

# Statistical mechanics for networks of real neurons

Leenoy Meshulam<sup>\*</sup>

*Department of Physics, Department of Neuroscience, and Carney Institute for Brain Science,  
Brown University,  
Providence, Rhode Island 02912, USA  
and Center for Computational Neuroscience,  
University of Washington,  
Seattle, Washington 98195, USA*

William Bialek<sup>†</sup>

*Joseph Henry Laboratories of Physics and Lewis-Sigler Institute for Integrative Genomics,  
Princeton University,  
Princeton, New Jersey 08544, USA*



(published 6 November 2025)

Perceptions and actions, thoughts and memories result from coordinated activity in hundreds or even thousands of neurons in the brain. It is an old dream of the physics community to provide a statistical mechanics description for these and other emergent phenomena of life. These aspirations appear in a new light because of developments in our ability to measure the electrical activity of the brain, sampling thousands of individual neurons simultaneously over hours or days. The progress that has been made in bringing theory and experiment together is reviewed, with a focus on maximum entropy methods and a phenomenological renormalization group. These approaches have uncovered new, quantitatively reproducible collective behaviors in networks of real neurons and provide examples of rich parameter-free predictions that agree in detail with experiments.

DOI: [10.1103/jcrn-3nrc](https://doi.org/10.1103/jcrn-3nrc)

## CONTENTS

I. Introduction	1
II. Some History	2
A. Prehistoric times	3
B. From perceptrons to deep networks	4
C. Symmetric networks	5
D. Perspectives	7
III. New Experimental Methods	7
A. Electrode arrays	8
B. Multiple electrodes in 3D	9
C. Imaging methods	9
D. Perspectives	11
IV. Maximum Entropy as a Path to Connect Theory and Experiment	12
A. Basics of maximum entropy	12
B. First connections to neurons	14
C. Larger networks of neurons	17
D. Doing more and doing less	23
V. A Unique Test	27
A. Many groups of $N = 100$ neurons	27
B. Maximum entropy models for subgroups	28
C. Success depends on the sampling density	29
D. Precision matters	33
VI. Criticality	35
A. Avalanches and dynamics	35
B. An effective thermodynamics	38
C. Bridges between dynamics and thermodynamics	40

D. Alternatives	41
VII. Renormalization Group for Neurons	45
A. Taking inspiration from the RG	45
B. By analogy with real-space methods	46
C. By analogy with momentum shell methods	49
D. RG as a path to understanding	50
VIII. Outlook	52
Acknowledgments	54
Appendix A: Sequences, Flocks, and More	54
1. Protein families	54
2. Collective behavior	56
3. Ecology and metabolism	59
4. Coda	60
Appendix B: Inference	60
References	61

## I. INTRODUCTION

Neural networks have been an inspiring source of physics problems for generations. The current revolution in artificial intelligence (LeCun, Bengio, and Hinton, 2015; Minaee *et al.*, 2024) has roots in classical work that appeared slightly more than 60 years ago in *Reviews of Modern Physics* (Rosenblatt, 1961; Block, 1962; Block, Knight, and Rosenblatt, 1962). The explicit effort to build network models grounded in statistical mechanics began in the 1970s (Cooper, 1973; Little, 1974; Little and Shaw, 1975, 1978) and received important stimuli in the early 1980s (Hopfield, 1982, 1984), with connections made to then-new ideas about spin glasses (Mézard, Parisi, and Virasoro, 1987; Amit, 1989). In the context of these models, one could use statistical mechanics to address not only the dynamics of a network but

\*leenoy@brown.edu

†wbialek@princeton.edu

also the way in which it learns from experience (Levin, Tishby, and Solla, 1990; Watkin, Rau, and Biehl, 1993). There is a path from this early work to current efforts of the theoretical physics community to understand the successes of machine learning (Carleo *et al.*, 2019; Mehta *et al.*, 2019; Roberts, 2021; Roberts and Yaida, 2022). In Sec. II we provide a brief guide to this rich history, emphasizing points that seem especially relevant for recent developments connecting theory and experiment.

In the long history of physicists' engagement with neural networks, it must be admitted that the search for tractable models often loosened the connection of theory to experiments on real brains. This problem became more urgent as methods became available to simultaneously monitor the electrical activity of tens, hundreds, or even thousands of neurons while animals engage in reasonably natural behaviors; see Sec. III. If we imagine a statistical mechanics for neural networks, these tools give us access to something like a Monte Carlo simulation of the microscopic degrees of freedom. This explosion of data calls out for new methods of analysis and creates new opportunities for interactions between theory and experiment.

Roughly 20 years ago it was suggested that maximum entropy methods could provide a very direct bridge from the new data on large numbers of neurons to explicit statistical physics models for these networks; see Sec. IV. In the simplest version of this approach, measurements on the mean activity and pairwise correlations among neurons result in an Ising spin glass model for patterns of activity in the network. All the couplings in the Ising model are determined by the measured correlations, and one can make parameter-free predictions for higher-order properties of the network. The surprise was that these predictions, at least in some cases, are extraordinarily successful; see Secs. IV.C and V.

The phenomenological success of the maximum entropy approach raised several questions. Should we expect this success to generalize, or was there something special about the first examples? Does success tell us something about the underlying network? If the models are so accurate, perhaps we should take them seriously as statistical physics problems: Where are real networks in the phase diagram of possible networks (see Sec. VI)? Can these models be given different interpretations, for example, in terms of a smaller number of “latent variables” that are encoded by the network?

The relatively simple statistical physics models constructed via maximum entropy are in some cases more successful than complex models motivated by biological details. Why should simple models work? In condensed matter physics we often describe macroscopic, emergent phenomena using models that are much simpler than the underlying microscopic mechanisms. This works not because we are lucky but because the renormalization group (RG) tells us that in many cases there is only a small number of relevant operators, so models simplify as we restrict our attention to longer length scales. Inspired by these ideas, there have been efforts to explicitly coarse grain the patterns of activity observed in large networks; see Sec. VII. The very first such efforts revealed surprisingly precise scaling behaviors, in some cases with exponents that are reproducible to the second decimal place. These initial results now have been confirmed in other systems.

As with maximum entropy methods, the success of coarse graining in uncovering interesting collective behaviors of real

neurons raises several questions. The observation of scaling suggests that the dynamics of these networks is controlled by some nontrivial fixed point of the RG. Can we identify this fixed point? Alternatively, can the observed scaling behaviors emerge in some other way? Away from the usual constraints of symmetry and locality, networks of neurons invite us to think about coarse graining more generally as lossy data compression. Does this connection between RG and information theory produce genuinely new physics, or just a new way of understanding old results? If there is an RG approach to the dynamics of neurons, is there a parallel RG analysis of the animal behaviors that these networks generate?

When physicists first wrote statistical mechanics models for neural networks, it was not clear whether these models should be taken as metaphors or whether they should be taken seriously as theories of real brains.<sup>1</sup> If forced to choose, most people would vote for metaphors since real brains surely are too complicated to be captured in the physicists' drive for simplification. While it emphatically is too soon to claim that we have a theory of the brain, progress that we review here makes clear that we can have the precise quantitative connections between theory and experiment that we have in the rest of physics. As experiments on the physics of living systems improve, we should ask more of our theories.

Finally, in case thinking about the brain is not sufficient motivation, networks of neurons provide a prototype of living systems with many degrees of freedom; see Appendix A. Even a single protein molecule typically is composed of more than 100 amino acids, and the structures and functions of these molecules emerge from interactions among these many more microscopic elements. At the next scale up, membrane patches and protein droplets self-organize in ways that most likely reflect phase separation. The identities and internal states of cells are determined by the expression levels of large numbers of genes that form an interacting regulatory network. In developing embryos, and tissues more generally, the movements of individual cells organize into macroscopic flows. In populations of bacteria, swarms of insects, schools of fish, and flocks of birds, we see collective movements and decision-making. In all these examples—and, of course, in networks of neurons—what we recognize as the functional behavior of living systems is a macroscopic behavior that emerges from interactions of many components on a smaller scale.

In the inanimate world, statistical mechanics provides a powerful and predictive framework within which to understand emergent phenomena. It is an old dream of the physics community that we could have a statistical mechanics of emergent phenomena in the living world as well. We encourage the reader to think of what we review here as progress toward realizing this dream.

## II. SOME HISTORY

Today, neural network models are known to many different communities: physicists and applied mathematicians,

---

<sup>1</sup>One can trace the metaphorical description of coordinated activity in the brain as being like collective effects in a magnet back even further, at least to the work of Cragg and Temperley (1954).

computer scientists and engineers, and neurobiologists and cognitive scientists. Neural networks are at the heart of an ongoing revolution in artificial intelligence and are making their way into many aspects of scientific data analysis, from cell biology to CERN. Here we provide a (perhaps idiosyncratic) reminder of how some of these ideas developed.

### A. Prehistoric times

The engagement of physicists with neurons and the brain has a long and fascinating history. Our modern understanding of electricity has its roots in the 1700s with observations on nerves and muscles. The understanding of optics and acoustics that emerged in the 1800s was continuous with the exploration of vision and hearing. This involved thinking not only about the optics of the eye or the mechanics of the inner ear but also about the inferences that our brains can derive from the data collected by these physical instruments.

The idea that the brain is made of discrete cells, connected by synapses, dates to the late 1800s (Ramón y Cajal, 1894). Electrical signals from individual nerve cells (neurons) were first recorded in the 1920s, starting with the cells in sense organs that provide the input to the brain (Adrian, 1928). Observing these small signals required instruments no less sensitive than those in contemporary physics laboratories. The crucial observation is that neurons communicate by generating discrete, identical pulses of voltage across their membranes; these pulses are called action potentials or spikes.<sup>2</sup>

By the 1950s there was a clear mathematical description of the dynamics underlying the generation and propagation of spikes (Hodgkin and Huxley, 1952). Perhaps surprisingly, the terms in these equations could be taken literally as representing real physical components—ion channel proteins that allow the flow of specific ions across the cell membrane and that open and close (“gate”) in response to the transmembrane voltage. The progress from macroscopic phenomenology to the dynamics of individual channels is a beautiful chapter in the interaction of physics and biology. Classic textbook accounts given by Aidley (1998) and Dayan and Abbott (2001) discussed phenomenological models for spiking activity, while a broader biological context was provided by Kandel *et al.* (2012). Rieke *et al.* (1997) described the way in which sequences of spikes represent information about the sensory world, and Bialek (2012) connected channels and spikes to other problems in the physics of biological systems.

Even before the mechanisms were clear, people began to think about how the quasidigital character of spiking could be

---

<sup>2</sup>Although beyond the scope of our discussion, spikes from a single neuron can sometimes have slightly different waveforms, and there are classical experiments showing that this has consequences for synaptic transmission. When spikes come in rapid succession, there is a systematic change in their shape, but this does not carry information beyond the arrival times of the spikes themselves. On longer timescales, neuromodulators and other effects can also change spike waveforms. These effects would be visible in direct electrical measurements (see Secs. III.A and III.B) but largely invisible in imaging experiments (see Sec. III.C). We are not aware of work that links changes in spike shape to the collective behavior in networks, except implicitly through changes in synaptic strength.

harnessed to perform computations (McCulloch and Pitts, 1943). This came after the foundational work of Turing (1937) on the theory of computation, but before any practical modern computers. The goal was to show that the basic facts known about neurons were sufficient to support computing essentially anything. On the one hand, this is a positive theoretical development: the brain could be a computer, in a deep sense. On the other hand, it is disappointing since, if the brain is a universal computer, there is not much more that one can say about the dynamics.

The way in which computation emerges from neurons in this early work clearly involves interactions among large numbers of cells in a network. Although single neurons can have remarkably precise dynamics in relation to sensory inputs and motor outputs (Rieke *et al.*, 1997; Nemenman *et al.*, 2008; Hires *et al.*, 2015; Srivastava *et al.*, 2017), there are many indications that our perceptions and actions, thoughts, and memories typically are connected to the activity in many hundreds, perhaps even hundreds of thousands, of neurons. Relevant activity in these large networks must be coordinated or collective.

The idea that collective neural activity in the brain might be described with statistical mechanics was greatly influenced by observations on the electroencephalogram (EEG) (Wiener, 1958). The EEG is a macroscopic measure of activity, traditionally done by simply placing electrodes on the scalp, and the existence of the EEG is *prima facie* evidence that the electrical activity of many, many neurons must be correlated. There was a notable demonstration by Adrian in which he sat quietly, eyes closed, with electrodes attached to his head. The signals, sent to an oscilloscope, showed the characteristic “alpha rhythm” that occurs in resting states, an oscillation at  $\sim 10$  Hz. When asked to add two numbers in his head, the rhythm disappeared, replaced by less easily described patterns of activity (Adrian and Matthews, 1934). This should dispel any lingering doubts that our mental lives are related to the electrical activity of our brains.

In the simplest models for neural dynamics, we describe the state of each neuron  $i$  at time  $t$  by a binary or Ising variable  $\sigma_i(t)$ ;  $\sigma_i(t) = +1$  means that the neuron is active, and  $\sigma_i(t) = 0$  means that the neuron is silent.<sup>3</sup> We imagine the dynamics proceeding in discrete time steps  $\Delta\tau$ . Each neuron sums inputs from other neurons, weighted by the strength  $J_{ij}$  of the synapse or connection from cell  $j \rightarrow i$ , and neurons switch into the active state if the total input is above a threshold,

$$\sigma_i(t + \Delta\tau) = \Theta \left[ \sum_j J_{ij} \sigma_j(t) - \theta_i \right]. \quad (1)$$

The nature of the dynamics is encoded in the matrix  $J_{ij}$  of synaptic strengths. If we think about arbitrary matrices, then

---

<sup>3</sup>For the moment *active* is a deliberately vague term. We could mean that the cell is in some relatively sustained state, perhaps steadily firing action potentials over a reasonable fraction of a second. Alternatively, we might be looking in very small time windows and asking about the presence or absence of single spikes. Resolving this vagueness is essential when we later connect theory with experiment.

the dynamics can be arbitrarily complex; progress depends on simplifying assumptions. It is useful to organize our discussion around two extreme simplifications. But keep in mind as we follow these threads that many of the developments occurred in parallel and that there was considerable crosstalk.

## B. From perceptrons to deep networks

One popular simplification is to assume that  $J_{ij}$  has a feedforward, layered structure. This is the perceptron architecture (Rosenblatt, 1961; Block, 1962; Block, Knight, and Rosenblatt, 1962), illustrated in Fig. 1(a), which is simpler to analyze precisely because there are no feedback loops. It is convenient to also label the neurons by the layer  $\ell$  in which they reside and to generalize from binary variables to continuous ones such that

$$x_i^{(\ell+1)} = g \left[ \sum_j W_{ij}^{(\ell+1)} x_j^{(\ell)} - \theta_i^{(\ell+1)} \right], \quad (2)$$

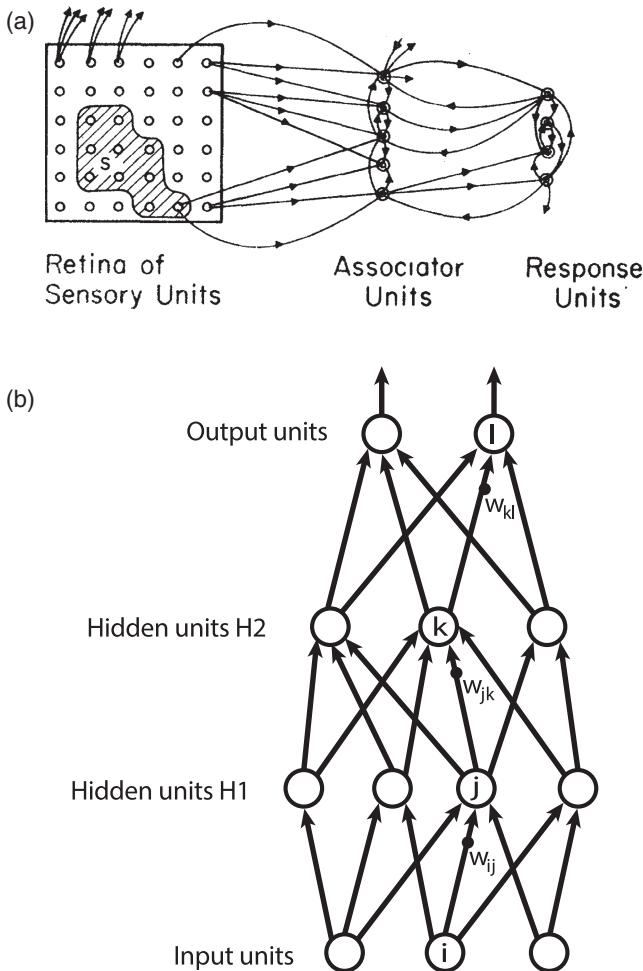


FIG. 1. Neural networks with a feedforward architecture, or perceptrons. (a) An early version of the idea. From Block, 1962. (b) A modern version with additional hidden layers. The first steps in the modern AI revolution involved similar networks, with many hidden layers, that achieved human-level performance on image classification and other tasks. From LeCun, Bengio, and Hinton, 2015.

where the propagation through layers replaces propagation through time and  $g[\cdot]$  is a monotonic nonlinear function. Thus, each neuron computes a single projection of its inputs from the previous layer and then outputs a nonlinear function of this projection.

In the limit in which  $g[\cdot]$  becomes a step function, we recover binary variables, and neuron  $i$  in layer  $\ell + 1$  can be thought of as dividing the space of its inputs in half, with a hyperplane perpendicular to the vector

$$\mathbf{V} = \{V_j\} = \left\{ W_{ij}^{(\ell+1)} \right\}. \quad (3)$$

Thus, the elementary computation is a binary classification of inputs,

$$\mathbf{x} \rightarrow y = \Theta(\mathbf{V} \cdot \mathbf{x} - \theta) \quad (4)$$

We could imagine having access to many examples of the input vector  $\mathbf{x}$  labeled by the correct classification  $y$  and thereby learning the optimal vector  $\mathbf{V}$ . This picture of learning to classify was present already ca. 1960, although it would take the full power of modern statistical physics to say that we really understand it. Crucially, if we think of  $\{x_i\}$  or  $\{\sigma_i\}$  as the microscopic variables in the system and  $J_{ij}$  as the interactions among these variables, then learning is statistical mechanics in the space of interactions (Gardner, 1988; Gardner and Derrida, 1988; Levin, Tishby, and Solla, 1990; Watkin, Rau, and Biehl, 1993).

Although many of the computations done by the brain can be framed as classification problems, such as attaching names or words to images, few can be solved by a single step of linear separation. Again, this was clear at the start, but development of these ideas took decades. Enthusiasm was dampened by an emphasis on what two layer networks could not do (Minsky and Papert, 1969), but eventually it became clear that multilayer perceptrons are much more powerful (LeCun, 1987; Lapedes and Farber, 1988), and theorems were proven to show that these systems can approximate any function (Hornik, Stinchcombe, and White, 1989). As with the simple perceptron, optimal weights  $W$  can be learned by fitting to many examples of input-output pairs. Importantly this does not require access to the “correct” answers at every layer. Instead, if we work with continuous variables, then the goodness of fit across many layers can be differentiated using the chain rule, and errors can be propagated back through the network to adjust the weights (Rumelhart, Hinton, and Williams, 1986).

Fast forward from the late 1980s to the mid-2010s. The few layers of early perceptrons became the many layers of “deep networks,” in the spirit of Fig. 1(b). Comparing Figs. 1(a) and 1(b) emphasizes the continuity of ideas across the decades. Advances in computing power and storage made it possible not only to simulate these models efficiently but also to solve the problem of finding optimal synaptic weights by comparing them to millions, or even billions, of examples. These explorations led to networks so large that the number of weights needed to specify the network vastly exceeded the number of examples. Contrary to well established intuitions, these “overparametrized” models worked by generalizing to new examples rather than overfitting to the training data.

Although we do not fully understand them, these developments have fueled a revolution in artificial intelligence (AI).

### C. Symmetric networks

Feedforward networks have the property that if  $J_{ij}$  is nonzero, then  $J_{ji} = 0$ . Hopfield (1982, 1984) considered the opposite simplification: if neuron  $i$  is connected to neuron  $j$ , then neuron  $j$  is connected to neuron  $i$  and the strength of the connection is the same, so  $J_{ij} = J_{ji}$ . In this case the dynamics in Eq. (1) have a Lyapunov function: at each time step the “energy”

$$E = -\frac{1}{2} \sum_{ij} \sigma_i J_{ij} \sigma_j + \sum_i \theta_i \sigma_i \quad (5)$$

either decreases or stays constant. The evolution of the network state stops at local minima of the energy  $E$ , and only at these local minima. We recognize this energy function as an Ising model with pairwise interactions among the spins (neurons). This explicit connection of neural dynamics to statistical physics triggered an avalanche of work, and textbook accounts of these ideas appeared quickly (Amit, 1989; Hertz, Krogh, and Palmer, 1991).

It is useful for visualizing the dynamics of symmetric networks that they can be realized by simple circuit components, using amplifiers with saturating outputs in place of neurons as in Fig. 2. As with perceptrons, one can generalize to soft spins, now in continuous time; one version of these dynamics is

$$\tau \frac{dx_i}{dt} = -x_i + \sum_j J_{ij} g(x_j). \quad (6)$$

These models have the same collective behaviors as Ising spins (Hopfield, 1984).

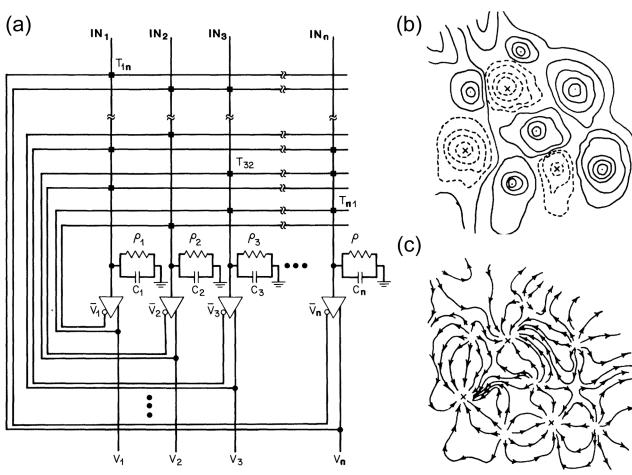


FIG. 2. Equivalent circuit and dynamics in a symmetric network. (a) A model neural circuit in electric components. (b) Schematic energy function for the circuit in (a). Solid contours are above a mean level and dashed contours are below, with X’s marking fixed points at the bottom of the energy valleys. (c) Corresponding dynamics shown as a flow field. From Hopfield and Tank, 1986.

A crucial point is that one can “program” symmetric networks to place local minima at desired states. Since the dynamics will flow spontaneously toward these minima and stop, we can think of this programming as storing memories in the network, which then can be recovered by initializing the state anywhere in the relevant basin of attraction. Taking the mapping of the Lyapunov function to an energy more seriously, this memory storage represents a sculpting of the energy landscape, which is a more general idea. As an example, we can think about the evolution of amino acid sequences in proteins sculpting the energy landscape for folding.

To illustrate the idea of memory storage, consider the case where the thresholds  $\theta_i = 0$ . Suppose that we can construct a matrix of synaptic weights such that

$$J_{ij} = J \xi_i \xi_j, \quad (7)$$

where the  $\xi_i = \pm 1$  are again a set of (now fixed) binary or Ising variables. Then the energy function becomes

$$E = -\frac{J}{2} \sum_{ij} \sigma_i \xi_i \xi_j \sigma_j = -\frac{J}{2} \left( \sum_i \sigma_i \xi_i \right)^2 = -\frac{J}{2} (\vec{\sigma} \cdot \vec{\xi})^2. \quad (8)$$

Because both  $\vec{\sigma}$  and  $\vec{\xi}$  are binary vectors, the energy is minimized ( $E$  is maximally negative) when these vectors are equal.<sup>4</sup> If want to be a bit fancier, we can transform  $\sigma_i \rightarrow \tilde{\sigma}_i = \sigma_i \xi_i$ , which is a discrete gauge transformation. We then realize that Eq. (8) is gauge equivalent to the mean-field ferromagnet.

Crucially, we can generalize this construction as

$$J_{ij} = J \left( \xi_i^{(1)} \xi_j^{(1)} + \xi_i^{(2)} \xi_j^{(2)} + \dots + \xi_i^{(K)} \xi_j^{(K)} \right). \quad (9)$$

If a network has  $N$  neurons and the number of these terms  $K \ll N$ , then the vectors  $\vec{\xi}^{(\mu)}$  are typically orthogonal and the energy function will have multiple minima at  $\vec{\sigma} = \vec{\xi}^{(\mu)}$ . We then have a model that stores  $K$  memories.

To make this more rigorous, we imagine that the states of the network not only are the minima of the energy function but also are drawn from a Boltzmann distribution at some inverse temperature  $\beta$ . It is plausible that this emerges from a noisy version of the dynamics in Eq. (1). Then we have

$$P(\vec{\sigma}) = \frac{1}{Z} \exp [-\beta E(\vec{\sigma})], \quad (10)$$

$$E(\vec{\sigma}) = -\frac{J_0}{N} \sum_{ij=1}^N \sum_{\mu=1}^K \sigma_i \xi_i^\mu \xi_j^\mu \sigma_j, \quad (11)$$

where we use the usual normalization of interactions by a factor  $N$  to insure a thermodynamic limit. Because the stored

<sup>4</sup>Because we set the thresholds to zero, the globally sign-flipped solution  $\vec{\sigma} = -\vec{\xi}$  is also allowed.

patterns are fixed, this is a statistical mechanics problem with quenched disorder, a special kind of mean-field spin glass. As a first try, we can take the stored patterns to be random vectors, which might make sense if we are describing a region of the brain where the mapping between the features of what we remember and the identities of neurons is abstract. We can measure the success of recalling memories by measuring the order parameters

$$m_\mu = \overline{\langle \vec{\xi}^\mu \cdot \vec{\sigma} \rangle}, \quad (12)$$

where  $\langle \dots \rangle$  denotes an average over the “thermal” fluctuations in the neural state  $\vec{\sigma}$  and  $\overline{\dots}$  denotes an average over the random choice of the patterns  $\vec{\xi}^\mu$ .

Shortly before the introduction of these models, there were dramatic developments in the statistical mechanics of disordered systems, including the solution of the fully mean-field Sherrington-Kirkpatrick spin glass model (Mézard, Parisi, and Virasoro, 1987). These tools could be applied to neural networks, resulting in a phase diagram mapping the order parameters  $\{m_\mu\}$  as a function of the fictitious temperature and the storage density  $\alpha = K/N$ , all in the thermodynamic limit  $N \rightarrow \infty$  (Amit, Gutfreund, and Sompolinsky, 1985, 1987). In the limit of zero temperature, below a critical  $\alpha_c = 0.138$  only one of the  $m_\mu$  will be nonzero, and it takes values close to 1; this survives to finite temperatures. Thus, there is an entire phase in which this model provides effective even if not perfect recall. By now, we think of neural network models not as an application of statistical mechanics but as a source of problems.

An important feature of the dynamics is that it is associative. Many initial states will relax to the same local minimum of the energy, which is equivalent to saying that the same memory can be recalled from many different cues. In particular, we can imagine that the many bits represented by the state  $\{\sigma_i\}$  can be grouped into features, for example, parts of the image of a face or the sound of the person’s voice. Under many conditions, if one set of features is given and the others randomized, the nearest local minimum will have all the features correctly aligned (Hopfield, 1982). The fact that our mind conjures an image in response to a sound or a fragrance once seemed mysterious; this provides a path to demystification, built on the idea that stored and recalled memories are collective states of the network.

The synaptic matrix in Eq. (9) has an important feature. Suppose that the network is currently in some state  $\vec{\sigma}$  and we want to add this state to the list of stored memories—i.e., we want the network to learn the current state. Following Eq. (9), we should change the synaptic weights

$$J_{ij} \rightarrow J_{ij} + J\sigma_i\sigma_j. \quad (13)$$

First, we note that the connection between neurons  $i$  and  $j$  changes in a way that depends only on these two neurons. This locality of the learning rule is in a way noteworthy since we might have thought that sculpting the energy landscape would require more global manipulations. Second, the change in

synaptic strength depends on the correlation between the presynaptic neuron  $j$  and the postsynaptic neuron  $i$ . If the cells are active together, the synapse should be strengthened. This simple rule is sometimes summarized by saying that neurons that “fire together wire together,” and there is considerable evidence that real synapses change in this way. Indeed, although this idea has its origins in classical discussions (James, 1904; Hebb, 1949), more direct measurements demonstrating that correlated activity leads to long-lasting increases of synaptic strength came only in the decade before Hopfield’s work (Bliss and Lømo, 1973).

In the first examples, the goal of computation was to recover a stored pattern from partial information (associative memory). Beyond memory, Hopfield and Tank (1985) soon showed that one could construct networks that solve classical optimization problems and that many biologically relevant problems could be cast in this form (Hopfield and Tank, 1986). Hopfield (2024) gave a modern perspective on these foundational developments.

A different generalization was the Boltzmann machine (Ackley, Hinton, and Sejnowski, 1985). Here the idea is that the states of individual neurons represent the states of individual variables in the world (for example, intensities in the pixels of an image), and the distribution over network states becomes a model for the joint distribution of these variables. The network is still described by the energy function in Eq. (5), but there is explicit randomness in the dynamics so that the distribution over states is guaranteed to be the Boltzmann distribution. Because the energy function is built only from interactions between pairs of neurons, one cannot generate arbitrary distributions. To match the complex structures in the distribution of real-world signals, one divides the neurons into visible and hidden units, with visible units representing variables in the outside world and hidden units serving an internal computational function. When we average over the states of hidden units, there are effective higher-order interactions among the visible units, and this provides the power needed to describe the distributions of real images, words, etc.

When the visible units are actually driven by incoming signals, the states of the hidden units can be interpreted as features or causes of the structure in these signals. This is especially true in the restricted Boltzmann machines where there are connections between visible and hidden units but not among neurons within these groups (Smolensky, 1986). Alternatively, if we let the network equilibrate without inputs, the hidden neurons drive the visible units to sample their distribution, in effect generating new images or sentences. This interplay of analysis and synthesis resonates decades later in generative AI and large language models. While it is a long and winding path, Hinton (2024) emphasized the role of Boltzmann machines as “historical enzymes” in catalyzing progress.

Parallel to these developments in statistical physics models of neural networks, the idea of simulated annealing (Kirkpatrick, Gelatt, and Vecchi, 1983) led people to take the mapping between “computational” problems of optimization and the “physical” problems of finding minimum energy states of many-body systems much more seriously. This led, for example, to connections between statistical

mechanics and computational complexity (Kirkpatrick and Selman, 1994; Monasson *et al.*, 1999). From an engineering point of view, models for neural networks connected immediately to the possibility of using modern chip design methods to build analog rather than digital circuits (Mead, 1989). Taken together, these simple symmetric models of neural networks formed a nexus among statistical physics, computer science, neurobiology, and engineering.

#### D. Perspectives

Our emphasis in this review is on networks of real neurons. But it would be foolish to ignore what is happening in the world of engineered artificial networks, which proceed at an alarming pace, realizing many of the old dreams for AI. Not long ago we would have emphasized the tremendous progress being made on problems such as image recognition and game playing, where deep networks achieved something that approximates human-level performance. Today, popular discussion is focused on generative AI, with networks that produce text and images that have a striking realism. Our theoretical understanding of why these things work remains quite weak. There are engineering questions about which practical problems can be solved with confidence by such systems and ethical questions about how humanity will interact with these machines. The successes of AI have even led some to suggest that physicists' notions of understanding might themselves be superseded. In contrast, many physicists are hopeful that ideas from statistical mechanics will help us to build a better understanding of modern AI (Carleo *et al.*, 2019; Mehta *et al.*, 2019; Roberts and Yaida, 2022).

In a different direction, many physicists have been interested in more explicitly dynamical models of neural networks (Vogels, Rajan, and Abbott, 2005), as in Eq. (6). Guided by the statistical physics of disordered systems, one can study networks in which the matrix of synaptic connections is drawn at random, perhaps from an ensemble that captures some established features of real connectivity patterns. These same ideas can be used for probabilistic models of binary neurons; notable developments include the development of a dynamical mean-field theory for these systems (van Vreeswijk and Sompolinsky, 1998).

Against the background of these theoretical developments, there has been a revolution in the experimental exploration of the brain, driven by techniques that combine methods from physics, chemistry, and biology. We believe that this provides an unprecedented opportunity to connect statistical physics ideas with quantitative measurements on network dynamics in real brains. We first turn to an overview of the experimental state of the art.

### III. NEW EXPERIMENTAL METHODS

Much of what we know about the brain has been learned by recording the electrical activity of one neuron at a time with metal microelectrodes. If we have a theoretical framework in which interesting things happen through collective activity in the network, however, it is difficult to see how we could make

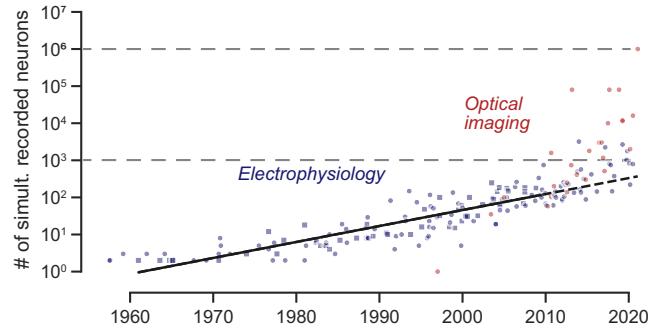


FIG. 3. Summary of the growth in scale of neural recordings Number of neurons recorded simultaneously with electrodes and electrode arrays (blue symbols); squares are from the earlier survey by Stevenson and Kording (2011). The number of neurons recorded simultaneously with optical imaging methods is indicated in red. Exponential growth for the electrode recordings is displayed as a black line, with a doubling time of  $7.4 \pm 0.4$  yr. Adapted from Urai *et al.*, 2022.

progress without experimental methods for recording from many neurons simultaneously.<sup>5</sup> Several groups were already recording from pairs of neurons in the 1960s, but systematic efforts to record from many neurons took until the 1980s.

For five decades we saw exponential growth in the number of neurons that can be monitored simultaneously with arrays of electrodes (see Fig. 3), with a doubling time of  $7.4 \pm 0.4$  yr (Stevenson and Kording, 2011). Impressively, progress followed essentially the same pace over the past decade, so  $\sim 10^3$  cells now are accessible almost routinely in many different brain areas and many different organisms; these developments are described in Secs. III.A and III.B. This century also brought a fundamentally new technique, with animals genetically engineered so that their neurons produce fluorescent proteins with fluorescence intensity modulated by electrical activity (see Sec. III.C); these methods are approaching  $\sim 10^6$  neurons (Demas *et al.*, 2021; Manley *et al.*, 2024). This progress creates new challenges for data analysis, but, more deeply, new opportunities for testing once-speculative theories. These developments also have a beauty of their own. Before we begin, note that as methods diversified, “recording from  $N$  neurons” came to mean different things, so a simple plot of  $N_{\max}$  versus time does not capture everything that is going on in these experiments, and these features of the experiments matter for theory.

Our emphasis here is on methods to record the electrical activity of many neurons simultaneously. We would be remiss if we did not point to the revolutionary developments in mapping the structure of these neuronal networks. It is a long-standing dream of the neuroscience community to visualize the wiring diagram of the brain, cataloging all the synaptic connections among neurons. While individual connections

<sup>5</sup>It is important that the “order parameters” in these theories are not simply the summed activity of all the neurons in the network, and hence do not correspond simply to something like the EEG. If we take the Hopfield model as an example, near its capacity the patterns of activity live in a space with dimensionality proportional to the size of the network itself (there are many order parameters), so there should not really be a simple path to dimensionality reduction.

have been studied for decades, the idea is to give a global view of all the connections, which has been termed the connectome by analogy with the transition from thinking about individual genes to thinking about an organism's entire genome.<sup>6</sup>

The first connectome was for the nervous system of the nematode *Caenorhabditis elegans*, which has just 302 neurons (White *et al.*, 1986). Progress has accelerated; recent milestones include the dense reconstruction of a 1 mm<sup>3</sup> volume of mouse visual cortex (MICrONS Consortium, 2025), identifying  $\sim 10^5$  neurons and  $\sim 5 \times 10^8$  synapses, and the full connectome of an adult fruit fly brain, with  $\sim 10^5$  neurons and  $\sim 5 \times 10^7$  synapses (Dorkenwald *et al.*, 2024). It is tempting to think one could use the connectome directly to build the matrix of synaptic connections,  $J_{ij}$  in the models of Sec. II, but it is not clear how to do this quantitatively. The most recent work in *C. elegans* even shows that there are functional connections between neurons that have no direct structural connection, mediated by molecules that diffuse through the extracellular space, and that these interactions can occur on timescales comparable to conventional synapses (Randi *et al.*, 2023). While it is encouraging to see connectomes being revealed, much work is needed before these data will have a compelling impact on our thinking about the patterns of neural activity.

### A. Electrode arrays

Rather than studying neurons in an intact brain, one can culture the cells in a dish, allowing them to connect into a network. Circa 1980, it was appreciated that the culture dish could be instrumented with an array of electrodes, giving access to the electrical activity of many (if not all) of the neurons in such artificial networks (Pine and Gilbert, 1982). Most of the brain is three dimensional, so this does not generalize, but the retina can be quite flat, at least locally. Placing a patch of a dissected retina onto an array of electrodes gives access to the “ganglion cells” that carry information from the eye to the brain and come together to form the optic nerve (Meister, Pine, and Baylor, 1994). Techniques progressed from recording a handful of cells simultaneously to arrays that can capture tens—and eventually hundreds—of cells, as shown in Figs. 4(a)–4(c) (Litke *et al.*, 2004; Segev *et al.*, 2004; Marre *et al.*, 2012). In some cases it is possible to achieve electrode densities high enough to record not only from large numbers of ganglion cells (more than 100) but also from a large fraction of the ganglion cells in a small patch of the retina, so we can access everything that the brain “sees” about a small patch of the visual world. These sorts of experiments have become routine in retinas from salamanders, mice, and primates whose visual systems are much like our own. There are efforts to scale up to recording from more than 1000 cells in this way (Tsai *et al.*, 2017).

<sup>6</sup>More generally we can think about “-omics” as the exhaustive study of systems—the proteome instead of individual proteins, the metabolome instead of individual metabolites, etc. Genome was originally a portmanteau of gene and chromosome, but perhaps the more general use of -omics is justified by the Greek σῶμα, meaning body; hence the connectome as the (whole) body of connections in the brain.

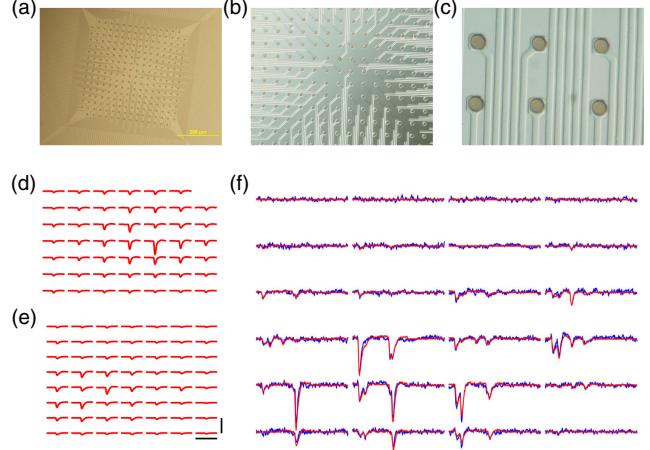


FIG. 4. Array of 252 electrodes for recording from the retina. (a)–(c) Views of the electrode array at increasing magnification. The distance between electrodes is 30  $\mu\text{m}$ . (d),(e) Examples of the stereotyped voltage traces [the templates  $T_{n\alpha}(\tau)$  in Eq. (14)] associated with two different cells. The scale bars are 6.5 ms and 200  $\mu\text{V}$ . (f) Raw voltage traces (blue lines) and reconstruction (red lines) made by superposing templates as in Eq. (14). The snippets are 20 ms in duration. From Marre *et al.*, 2012.

In electrode arrays each electrode picks up signals from multiple cells and each cell appears on multiple electrodes. Thus, there is a deconvolution problem referred to as spike sorting. This can be solved because the spikes generated by individual neurons are stereotyped. Concretely, this means that we can write the voltage  $v_n(t)$  on the  $n$ th electrode as a sum of terms contributed by action potentials from cell  $\alpha$  at times  $t_i^\alpha$ ,

$$v_n(t) = \sum_\alpha \sum_i T_{n\alpha}(t - t_i^\alpha) + \eta_n(t), \quad (14)$$

where the  $T_{n\alpha}(\tau)$  are “templates” that express how cells appear at electrodes and  $\eta_n(t)$  is residual noise; see Figs. 4(d)–4(f). In outline one can learn these templates by finding candidate spike events that stand well above the background noise, clustering them, using the cluster centers as matched filters to identify more candidate spikes, and iterating. There are many challenges in turning this outline into a working algorithm. For the multielectrode arrays used in recording from the retina, see the discussions by Prentice *et al.* (2011) and Marre *et al.* (2012). An important test of spike sorting is that spikes from a single neuron should never come closer in time than a refractory period of  $\sim 1$  ms.

Before searching for collective behaviors in the population of neurons, experiments with multielectrode arrays provide an efficient way of exploring the properties of many individual cells. Neurons throughout the brain can be divided into cell types, with different types exhibiting, for example, different responses to sensory inputs, different three-dimensional structures, and, more recently, different patterns of gene expression. The retina is a classic example, with classification based on structure dating back to the classic work of Ramón y Cajal (1893). Electrode arrays provide a direct view of how cells of a particular type tile the retina in a lattice and how the lattices of different cell types interdigitate (Field and

Chichilnisky, 2007; Roy *et al.*, 2021).<sup>7</sup> In addition to classification based on their responses to visual inputs, the templates  $T_{na}(\tau)$  derived from spike sorting can be thought of as electrical images of each cell, and these images also aid in the classification of neural cell types (Wu *et al.*, 2023).

## B. Multiple electrodes in 3D

A different approach is to insert multiple electrodes deep into brain tissue, which also has a long history. Where classical experiments brought a metal tip as close as possible to a single neuron, it was appreciated that multiple closely spaced tips, for example, those with wires twisted into a stereotrode or tetrode, could resolve multiple neurons from a small volume (McNaughton, O’Keefe, and Barnes, 1983; Wilson and McNaughton, 1993). The introduction of methods from semiconductor fabrication made it possible to build arrays of 100 silicon needles that could be inserted into the cortex (Jones, Campbell, and Normann, 1992).

Jumping ahead two decades, further miniaturization has led to integrated arrays of multiple electrodes along a single shaft coupled with preprocessing electronics, as illustrated in Fig. 5 (Jun *et al.*, 2017). The most recent such devices have more than 1000 sensors along a single probe capable of resolving hundreds of individual neurons (Steinmetz *et al.*, 2021). Although it is most common to deploy these arrays in studies on rodent brains, they can also be adapted to primates, where comparisons to the human brain are easier (Trautmann *et al.*, 2023). Alternative methods make use of polymer materials for flexible electrodes (Chung *et al.*, 2019). In particular, these permit long-term recordings that monitor the same neurons over weeks or months, for example, as the animal learns (Zhao *et al.*, 2023). Importantly, all these methods, as with classical single neuron recordings, provide access to the full stream of action potentials generated by each neuron down to millisecond precision.

Although our emphasis is on basic scientific questions, an important stimulus for continued development of these techniques is their potential for clinical applications. In particular, there is the program of constructing “brain-computer interfaces,” where electrode arrays monitor the activity of many neurons in the motor cortex and these signals are decoded to generate commands for a robot arm or cursor (Serruya *et al.*, 2002; Taylor, Tillery, and Schwartz, 2002; Carmena *et al.*, 2003; Musallam *et al.*, 2004). More recently, these techniques have emerged from the laboratory to experimental treatments of humans (Hochberg *et al.*, 2012; Willett *et al.*, 2021). This is a rapidly developing field in which not only experimental methods but also our theoretical understanding of neural dynamics and coding is contributing to practical medical goals.

<sup>7</sup>Although generally forming a lattice, the regions of the visual world to which individual cells respond (“receptive fields”) can be irregular. Experiments using the electrode arrays also show that the irregularities in the receptive fields of neighboring cells are coordinated such that they interlock and provide more uniform coverage of the visual world (Gauthier *et al.*, 2009). For a theoretical discussion, see Liu, Stevens, and Sharpee (2009).

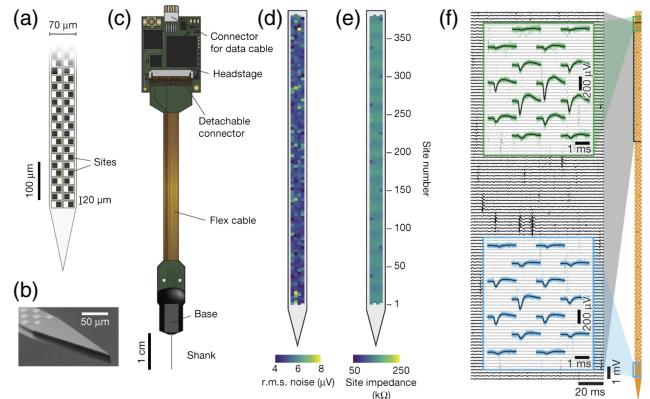


FIG. 5. “Neuropixel” probe, with 384 electrodes arrayed along a single shank. (a) Schematic of a probe tip showing a checkerboard layout of active electrode sites. (b) Scanning electron microscope image of a probe tip. (c) Probe packaging, including flexible cable and head stage electronics for data transmission. (d) Example of root-mean-square voltage noise levels in a bandwidth that captures the action potentials;  $\delta V_{\text{rms}} = 5.1 \pm 0.6 \mu\text{V}$ . (e) Typical site impedance in saline measured for each site with sinusoidal 1 nA injected currents at 1 kHz;  $Z = 149 \pm 6 \text{ k}\Omega$ . (f) A short segment of raw voltage recordings in the mouse brain. Insets: short snippets from multiple nearby electrodes that are identified as spikes from the same neuron, with 30 waveforms superposed to illustrate the stereotypy of these signals. The angle with which the shank penetrated the brain was chosen to sample many different areas. Upper electrodes are in the motor cortex, while lower electrodes are in the dorsal tenia tecta. From Jun *et al.*, 2017.

Versions of these tools have been commercialized, leading to an explosive increase in large-scale experiments across a wide range of brain regions in many different animals; a snapshot of this activity was provided by Steinmetz *et al.* (2018). As with the electrode arrays in Sec. III.A, signals from individual neurons appear at multiple electrodes and individual electrodes pick up multiple neurons, so there is a problem with spike sorting. With thousands of neurons, this problem is on a much larger scale than before, and there is a particular drive to have fully automated methods (Chung *et al.*, 2017). Progress continues, but the problem is not fully solved. We add that different analyses are sensitive to different systematic errors in the sorting process.

## C. Imaging methods

It is an old idea that we might be able to see the electrical activity of neurons, literally. The first implementation was with voltage-sensitive dyes that insert into the cell membrane and have optical properties (absorption or fluorescence) that shift in response to the large electric fields associated with the action potential (Cohen and Salzberg, 1978). The exploration of the brain (and living systems more generally) was revolutionized by the discovery that there are proteins that are intrinsically fluorescent without the need of cofactors (Johnson *et al.*, 1962; Shimomura, Johnson, and Saiga, 1962). These proteins were then tuned, by changing their amino acid sequences, to have different colors as well as fluorescence that

responds to environmental signals (Tsien, 2009). Decades after their initial discovery, genetic engineering allowed the insertion of these sequences into the genome (Prasher *et al.*, 1992; Chalfie *et al.*, 1994), placing them under the control of regulatory elements that are active in neurons or even in restricted classes of neurons.

Taking inspiration from voltage-sensitive dyes, the ideal would be to have a genetically encoded, fluorescent membrane protein that responds directly to the voltage across the membrane. There is continuing progress toward this goal (Jin *et al.*, 2012; Abdelfattah *et al.*, 2019; Villette *et al.*, 2019; Platisa *et al.*, 2023), but current indicator molecules are not quite sufficient for long-term recordings from large populations of neurons. What we do have are fluorescent proteins that respond to changes in intracellular calcium concentration, which provides a slightly indirect, low-pass-filtered trace of electrical activity; these now are widely used (Tian *et al.*, 2012; Chen *et al.*, 2013; Zhang *et al.*, 2023). To make the most of these signals requires sophisticated microscopy, such as scanning two-photon methods (Helmchen and Denk, 2005). With these tools we can observe reasonably large areas of the brain with single cell resolution, as in Figs. 6(a)–6(c). In addition, there are now engineered proteins that can be inserted into the membrane and act as light-gated channels, making it possible to inject controlled pulses of current into individual neurons to both excite and inhibit these cells through optical control (Rickgauer, Deisseroth, and Tank, 2014; Packer *et al.*, 2015).

In many regions of the brain, we do not see the full dynamical behavior of neuronal networks unless the animal is engaged in behavior. Evidently, having the “sample” moving and behaving is in tension with high-resolution microscopy. One solution is to miniaturize the microscope so that the animal can carry the instrument as it moves through its environment (Ziv *et al.*, 2013; Zong *et al.*, 2017). Alternatively, one can hold the animal’s head fixed under a stationary microscope but allow it to run on a rotating ball, using the movement of the ball to compute how the animal would have moved through the environment. This computed trajectory is then used to generate virtual reality (Harvey *et al.*, 2009; Dombeck *et al.*, 2010). An example of a virtual reality setup is shown in Fig. 16(b). It is possible to simulate not only the animal’s visual experience of running through the world but also its olfactory experience (Radvansky and Dombeck, 2018).

Imaging methods allow flexible trade-offs among spatial resolution, temporal resolution, the area over which one records, and the signal-to-noise ratio for each individual cell. As seen in Figs. 6(d) and 16(c), there is a regime in which the transient periods of neural activity stand out well above the background noise of the measurements from individual cells. If the aim is to simultaneously record from as many neurons as possible, one can reach every neuron in the brain of smaller animals, such as larval zebra fish, at the expense of visiting each neuron less frequently (Ahrens *et al.*, 2013). More generally, it is possible to combine methods, providing single cell recordings at high time resolution while monitoring a much larger area of the brain at lower resolution (Barson *et al.*, 2020).

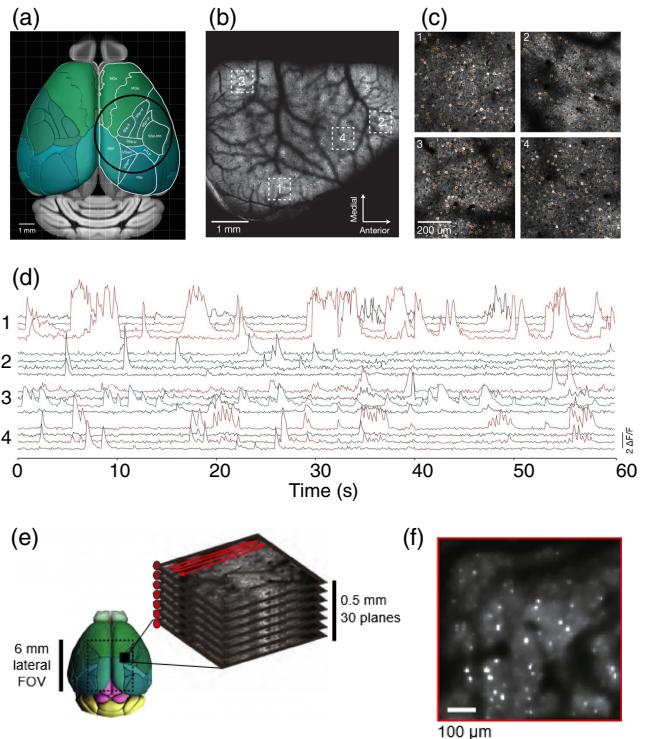


FIG. 6. Large-scale imaging of neural activity. (a) Schematic of the dorsal surface of the mouse cortex. The overlaid circle corresponds to the field of view when imaging with the “mesoscope” setup (5 mm radius). (b) Mesoscope image from a mouse brain expressing the fluorescent calcium indicator protein GCaMP6f. (c) Four fields of view [indicated in (b)], with regions of interest (orange) drawn manually around individual neurons. (d) Normalized fluorescence activity traces for 16 neurons extracted from the four regions in (b). Sampling rate 9.6 Hz. (e) An example field of view for volumetric imaging via light bead microscopy, which recently enabled monitoring  $\sim 10^6$  neurons simultaneously. The fast pulse structure of the laser source is used to produce “beads” (red dots) at different depths across 0.5 mm, and this is then scanned laterally, enabling volumetric recording. (f) The standard deviation of fluorescence in an imaging plane 183  $\mu\text{m}$  below the cortical surface in a mouse brain expressing GCaMP6f. (a)–(d) Adapted from Sofroniew *et al.*, 2016. (e),(f) From Manley *et al.*, 2024.

A special case is the small worm *Caenorhabditis elegans*, which has only 302 neurons in total. As in many invertebrates, these neurons have names and numbers and thus are identifiable across individuals. *C. elegans* was the first organism in which the pattern of synaptic connectivity was traced at electron microscope resolution (White *et al.*, 1986), and this connectome has been revisited with modern methods (Varshney *et al.*, 2011; Cook *et al.*, 2019). The worm is largely transparent, so optical methods can be used directly to monitor and drive neural activity without dissection, even in freely moving worms (Leifer *et al.*, 2011). Recordings from more than 100 neurons in this system reflect a macroscopic fraction of all the neurons, so we are approaching “whole brain” imaging with single cell resolution (Nguyen *et al.*, 2016). The neurons in *C. elegans* do not generate the discrete, stereotyped action potentials that are familiar in other

organisms, so the graded fluorescence signals in imaging experiments are a more direct correlate of slower, continuous electrical dynamics. Advances in experimental technique make it possible to identify neurons as their activity is monitored, placing them in the context of the known connectivity, and the combination of recording and stimulation has resulted in “pump-probe” measurements that map the functional connections between more than 10 000 pairs of cells (Randi *et al.*, 2023). These data provide the opportunity to formulate and test more global theoretical ideas about network dynamics.

If we want to visit each neuron often enough to make full use of the time resolution allowed by the calcium response of the fluorescent proteins, then there will be limits on the number of neurons that can be monitored. Scanning in two dimensions, one can now reach more than 1000 neurons, as in the example discussed in Secs. V and VII. Scanning in depth poses additional challenges (Weisenburger *et al.*, 2019; Zhang *et al.*, 2021), but new “light bead” methods make use of the short timescale of laser pulses to collect from multiple depths almost simultaneously, as shown in Figs. 6(e) and 6(f) (Demas *et al.*, 2021; Manley *et al.*, 2024). These methods are pushing toward monitoring  $10^6$  cells.

The raw data from an imaging experiment are a movie: fluorescence intensity versus time in each of  $\sim 10^6$  pixels. What we want are signals labeled by the cells that generate them, not by pixels. This involves two essential steps: discarding all changes in light intensity that result from sources other than electrical activity (primarily motion of the brain) and grouping together the pixels that belong to each cell. In many cases these steps need to be taken in three dimensions, combining signals from a “z stack” in which the microscope’s plane of focus has been stepped through the thickness of the brain region under study. These are challenging problems in data analysis, and a wide range of mathematical and algorithmic ideas have been brought to bear: local correlations (Smith and Häusser, 2010), dictionary learning (Pachitariu *et al.*, 2013), graph-cut-related algorithms (Kaifosh *et al.*, 2014), independent component analysis (Mukamel, Nimmerjahn, and Schnitzer, 2009), and non-negative matrix factorization (Maruyama *et al.*, 2014).

The fact that neurons generate discrete action potentials means that if we look in small time bins, the natural variables are binary, inviting a connection to Ising models. Calcium-sensitive indicators do not give us direct access to the time resolution that is needed for this binary description. There are several efforts to reconstruct the spikes that underlie the  $\sim 100$  ms calcium signals, but we suspect that these will be overtaken by advances in engineering directly voltage-sensitive proteins. An alternative that we use below is to discretize the calcium signals, admitting that the resulting binary variables necessarily refer to “active” and “inactive” states of the cell rather than to the presence or absence of action potentials; see Fig. 16(c).

#### D. Perspectives

Experimental methods for monitoring the electrical activity of neurons continue to evolve rapidly. We now look ahead and make some predictions about where the methods will be in

five or ten years. We caution that we are theorists surveying the state of experiments.

In recordings based on electrodes and electrode arrays, we can expect two major trends. The first is better coverage and higher sampling density. It is tempting to focus on the largest-scale experiments, as these are perhaps the most tantalizing opportunities to test the applicability of statistical physics ideas. In practice, however, more neurons often come at the expense of lower sampling density, which matters deeply for comparison with theory (see Sec. V.C), so one would like to be careful. We expect that the push for whole brain coverage soon will be complemented by a push for denser sampling: instead of choosing between high-density sampling in a small region, often in 2D, or sparse sampling of much larger areas in 3D, experiments will get much closer to recording every neuron in progressively larger volumes. The second trend is toward longer duration recordings, with a chronic presence of electrodes in the animal brain. Recent efforts have provided proof of concept for continuous recordings that last for weeks; we expect this to become more routine, reaching toward experiments that last for days or even months (Dhawale *et al.*, 2017; Chung *et al.*, 2019). The central challenge is to verify that we are monitoring the exact same set of cells throughout the entire recording. The great advantage is that the animal can be monitored in its home cage, in different environments, and at different times of day as it engages in a fuller range of behaviors. The longest timescale recordings will give a unique view of neural dynamics during learning.

On the optical front, the growth in the number of neurons that we can (literally) see simultaneously has recently accelerated dramatically, as seen clearly in Fig. 3. Faster, more selective scanning is in the works, which should allow more imaging techniques to reach the realm of  $\sim 10^6$  neurons, with improved signal-to-noise ratio. Currently, when imaging  $10^5$ – $10^6$  neurons, the loss of temporal resolution is significant, with a drop to acquisition rates below 10 Hz. As with electrodes where sampling density in space matters, here it is the sampling density in time that can be problematic. There are trade-offs among speed, the number of neurons, the signal-to-noise ratio in each neuron, and the total amount of optical power delivered to the brain, but these are specific to each imaging modality, and we can hope for progress.

Another intriguing direction is selective acquisition. Following methods used in astrophysics, if we can concentrate on the exact locations of the neurons, we can scan more quickly and use the same number of photons more efficiently. In addition, there is steady improvement in methods to express both indicators and light-gated channels in the same cells, often targeting specific classes of cells. This will bring to larger animals the kind of complete survey of functional connectivity that currently is possible only in *C. elegans* (Randi *et al.*, 2023) and also make it possible to probe causal connections between neural activity and motor output.

Finally, a significant breakthrough would be if voltage-sensitive fluorescent proteins become fully viable. The demands are severe: proteins must respond on a millisecond timescale, with large amplitude changes in fluorescence, and cells must be programmed to insert these proteins into the membrane. When this happens, it will become possible to

monitor thousands to millions of neurons with a resolution where we see every individual action potential, giving us the precision of electrodes and the survey capacity of optical imaging.

#### IV. MAXIMUM ENTROPY AS A PATH TO CONNECT THEORY AND EXPERIMENT

New experimental methods create new opportunities to test our theories. For neural networks monitoring the electrical activity of tens, hundreds, or thousands of neurons simultaneously should allow us to test statistical approaches to these systems in detail. Doing this requires taking much more seriously the connection between our models and real neurons—a connection that sometimes has been tenuous. Can we really take the spins  $\sigma_i$  in Eq. (5) to represent the presence or absence of an action potential in cell  $i$ ? We do indeed make this identification, and our goal is an accurate description of the probability distribution out of which the “microscopic” states of a large network are drawn. Note that, as in equilibrium statistical mechanics, this would be the beginning, not the end, of our understanding.

Maximum entropy models provide a path that starts with data and constructs models that have a direct connection to statistical physics. Our focus here is on networks of neurons, but it is important that the same concepts and methods are being used to study a much wider range of living systems. There are important lessons to be drawn from seeing all these problems as part of the same project; see Appendix A.

##### A. Basics of maximum entropy

Consider a network of neurons labeled by  $i = 1, 2, \dots, N$ , each with a state  $\sigma_i$ . In the simplest case where these states of individual neurons are binary—active or inactive or spiking or silent—then the network as a whole has access to  $\Omega = 2^N$  possible states

$$\sigma = \{\sigma_1, \sigma_2, \dots, \sigma_N\}. \quad (15)$$

These states mean something to the organism: they may represent sensory inputs, inferred features of the surrounding world, plans, motor commands, recalled memories, or internal thoughts. To make progress, we need to know which of these (very) many possible patterns of neural activity actually occur, and how often. More formally, we want to understand the probability distribution  $P(\sigma)$ . We might also be interested in sequences of states over time,  $P[\{\sigma(t_1), \sigma(t_2), \dots\}]$ , but for simplicity we focus first on states at a single moment in time.

The distribution  $P(\sigma)$  is a list of  $\Omega$  numbers that sum to 1. Even for modest sized networks, this is a long list,  $\Omega \sim 10^{30}$  for  $N = 100$ . We emphasize that there is no way for us to measure all these numbers in a realistic experiment. More deeply, large networks could not visit all of their possible states in the age of the Universe, let alone the lifetime of a single organism. This should not bother us, since one can make similar observations about the states of molecules in the air around us or the states of all the atoms in a small grain of sand. The fact that the number of possible states  $\Omega$  is (beyond)

astronomically large does not stop us from asking questions about the distribution from which these states are drawn.

The enormous value of  $\Omega$  does mean, however, that answering questions about the distribution from which the states are drawn requires the answer to be, in some sense, simpler than it could be. If  $P(\sigma)$  really were just a list of  $\Omega$  numbers with no underlying structure, we could never make a meaningful experimental prediction. Progress in the description of many-body systems depends on the discovery of some regularity or simplicity. Without such simplifying hypotheses, nothing can be inferred from any reasonable amount of data. The maximum entropy method is a way of being explicit about our simplifying hypotheses.

We can imagine mapping each microscopic state  $\sigma$  into some perhaps more macroscopic observable  $f(\sigma)$ , and from reasonable experiments we should be able to estimate the average of this observable  $\langle f(\sigma) \rangle_{\text{expt}}$ . If we think this observable is an important and meaningful quantity, it makes sense to insist that any theory we write for the distribution  $P(\sigma)$  should predict this expectation value correctly,

$$\langle f(\sigma) \rangle_P \equiv \sum_{\sigma} P(\sigma) f(\sigma) = \langle f(\sigma) \rangle_{\text{expt}}. \quad (16)$$

There might be several such meaningful observables, so we should have

$$\langle f_{\mu}(\sigma) \rangle_P \equiv \sum_{\sigma} P(\sigma) f_{\mu}(\sigma) = \langle f_{\mu}(\sigma) \rangle_{\text{expt}} \quad (17)$$

for  $\mu = 1, 2, \dots, K$ . These are strong constraints, but so long as the number of these observables  $K \ll \Omega$ , there are infinitely many distributions consistent with Eq. (17). How do we choose among them?

There are many ways of saying in words how we want to make our choice among the  $P(\sigma)$  that are consistent with the measured expectation values of the observables. We want to pick the simplest or the least structured model. We do not want to inject any information into our model beyond what is given to us by the measurements  $\{\langle f_{\mu}(\sigma) \rangle_{\text{expt}}\}$ . From a different point of view, we want to draw states out of the distribution  $P(\sigma)$  to generate samples that are as random as possible while still obeying the constraints in Eq. (17). It might seem that each choice of words generates a new discussion—what do we mean, mathematically, by “least structured,” or “as random as possible”?

Introductory courses in statistical mechanics make some remarks about entropy as a measure of our ignorance about the microscopic state of a system, but this connection is often left vague. In laying the foundations of information theory, Shannon (1948) made this connection precise. If we ask a question, we have the intuition that we gain information when we hear the answer. If we want to attach a number to this information gain, then the unique measure that is consistent with natural constraints is the entropy of the distribution out of which the answers are drawn. Thus, if we ask for the microscopic state of a system, the information that we gain upon hearing the answer is (on average) the entropy of the distribution over these microscopic states. Conversely, if the entropy is less than its maximum possible value, this reduction

in entropy measures how much we already know about the microscopic state even before we see it. As a result, for states to be as random as possible—to be sure that we do not inject extra information about these states—we need to find the distribution that has the maximum entropy.

Maximizing the entropy subject to constraints defines a variational problem, extremizing

$$\tilde{S} = -\sum_{\sigma} P(\sigma) \ln P(\sigma) - \sum_{\mu=1}^K \lambda_{\mu} \left[ \sum_{\sigma} P(\sigma) f_{\mu}(\sigma) - \langle f_{\mu}(\sigma) \rangle_{\text{expt}} \right] - \lambda_0 \left[ \sum_{\sigma} P(\sigma) - 1 \right], \quad (18)$$

where  $\lambda_{\mu}$  are Lagrange multipliers. We include an additional term ( $\propto \lambda_0$ ) in Eq. (18) to constrain the normalization so that the probability  $P(\sigma)$  for each state  $\sigma$  can be treated as an independent variable. Then, finding an extremum,

$$\frac{\delta \tilde{S}}{\delta P(\sigma)} = 0, \quad (19)$$

yields the probability distribution

$$P(\sigma) = \frac{1}{Z(\{\lambda_{\mu}\})} \exp[-E(\sigma)], \quad (20)$$

$$E(\sigma) = \sum_{\mu=1}^K \lambda_{\mu} f_{\mu}(\sigma). \quad (21)$$

Thus, the model we are looking for is equivalent to an equilibrium statistical mechanics problem in which the “energy” is a sum of terms, one for each of the observables whose expectation values we constrain. The Lagrange multipliers become coupling constants in the effective energy. To finish the construction, we need to adjust these couplings  $\{\lambda_{\mu}\}$  to satisfy Eq. (17). In general, this is a difficult problem; see Appendix B. If we have a set of expectation values to match and we want to add one more, this simply adds one more term to the form of the energy function, but in general implementing this extra constraint requires all the coupling constants to be adjusted.

To make the connections explicit, recall that we can define thermodynamic equilibrium as the state of maximum entropy given the constraint of fixed mean energy. This optimization problem is solved using the Boltzmann distribution. In this view the inverse temperature is a Lagrange multiplier that enforces the energy constraint, in contrast to the usual view of controlling the temperature and predicting the energy. The Boltzmann distribution generalizes if other expectation values are constrained (Landau and Lifshitz, 1977).

The maximum entropy argument gives us the form of the probability distribution, but we also need the coupling constants. We can think of this as being an “inverse statistical mechanics” problem since we are given expectation values or correlation functions and need to find the couplings, rather than the other way around. Different formulations of this

problem have a long history in the mathematical physics community (Keller and Zumino, 1959; Kunkin and Firsch, 1969; Chayes, Chayes, and Lieb, 1984). An early application to living systems involved reconstructing the forces that hold together the array of gap junction proteins that bridge the membranes of two cells in contact (Braun, Abney, and Owicky, 1984). As attention focused on networks of neurons, finding the relevant coupling constants came to be described as the “inverse Ising” problem.

In statistical physics there is, in some sense, a force driving systems toward equilibrium, as encapsulated in the  $H$  theorem. In many cases this force triumphs and what we see is a state with maximal entropy subject only to a very few constraints. In the networks of neurons that we study here, there is no  $H$  theorem, and the list of constraints is quite long compared to what we are used to in thermodynamics. This means that the probability distributions we write will be mathematically equivalent to some equilibrium statistical mechanics problem, but they do not describe an equilibrium state of the system that we are actually studying. This somewhat subtle relationship between maximum entropy as a description of thermal equilibrium and maximum entropy as a tool for inference was outlined long ago by Jaynes (1957, 1982).

If we do not have any constraints, then the maximum entropy distribution is uniform over all  $\Omega$  states. Each observable whose expectation value we constrain lowers the maximum allowed value of the entropy, and if we add enough constraints, we eventually reach the true entropy and hence the true distribution. It often makes sense to group the observables into one-body terms, two-body terms, three-body terms, etc.. Having constrained all the  $k$ -body observables for  $k \leq K$ , the maximum entropy model makes parameter-free predictions for correlations among groups of  $k > K$  variables. This provides a powerful path for testing the model and defines a natural generalization of connected correlations (Schneidman *et al.*, 2003).

The connection of maximum entropy models to the Boltzmann distribution gives us intuition and practical computational tools. It can also leave the impression that we are describing a system in equilibrium, which would be a disaster. In fact, the maximum entropy distribution describes thermal equilibrium only if the observable that we constrain is the energy in the mechanical sense. There is no obstacle to building maximum entropy models for the distribution of states in a nonequilibrium system.

In the model of Hopfield (1982), the energy function emerges from an analysis of the dynamics. Here the energy function arises in describing the distribution of network states at a single moment in time. It is important to note that infinitely many different dynamical models can lead to the same distribution. In particular, finding that the distribution over network states is a Boltzmann distribution with some energy landscape does not imply that the dynamics of the network is Brownian motion on this landscape,<sup>8</sup> and most of

---

<sup>8</sup>Or a discrete analog in the case of Ising variables.

the dynamical models that reproduce the distribution do not even obey detailed balance.

Recent work showed how to explore a large family of dynamical models consistent with the maximum entropy distribution and applied these ideas to collective animal behavior (Chen *et al.*, 2023). There also are generalizations of the maximum entropy method to describe distributions of trajectories, as we discuss in Sec. IV.D; maximum entropy models for trajectories are sometimes called the maximum caliber approach (Pressé *et al.*, 2013; Ghosh *et al.*, 2020). Finally, we note that, for better or for worse, the symmetries that are central to many problems in statistical physics in general are absent from the systems that we study here. Flocks and swarms are an exception, as discussed in Sec. A.2.

We emphasize that maximum entropy is unlike usual theories. We do not start with a theoretical principle—or even a model. Rather, we start with some features of the data and test the hypothesis that these features alone encode everything that we need to describe the system. Whenever we use this approach, we are referring back to the basic structure of the optimization problem defined in Eq. (18) and its formal solution in Eqs. (20) and (21). But there is no single maximum entropy model, and each time we need to be explicit: which are the observables  $f_\mu$  whose measured expectation values we want our model to reproduce? Can we find the corresponding Lagrange multipliers  $\lambda_\mu$ ? Do these parameters have a natural interpretation? Once we answer these questions, we can ask whether these relatively simple statistical physics descriptions make predictions that agree with the experiment. There is an unusually clean separation between learning the model (matching observed expectation values) and testing the model (predicting new expectation values). In this sense we can think of maximum entropy as predicting a set of parameter-free relations among different aspects of the data. Finally, we have to think carefully about what it means for models to work. We begin with early explorations at relatively small  $N$  (see Sec. IV.B), then turn to a wide variety of larger networks (see Sec. IV.C), and finally address how these analyses can catch up to the experimental frontier (see Sec. IV.D).

## B. First connections to neurons

Suppose that we observe three neurons and measure their mean activity as well as their pairwise correlations. Given these measurements, should we be surprised by how often the three neurons are active together? Maximum entropy provides a way of answering this question, generating a “null model” prediction assuming that the entire correlation structure is captured in the pairs, and this was appreciated  $\sim 2000$  (Martignon *et al.*, 2000). Over the next several years, a more ambitious idea emerged: could we build maximum entropy models for patterns of activity in larger populations of neurons? The first target of this analysis was a population of neurons in the salamander retina as it responds to naturalistic visual inputs (Schneidman *et al.*, 2006).

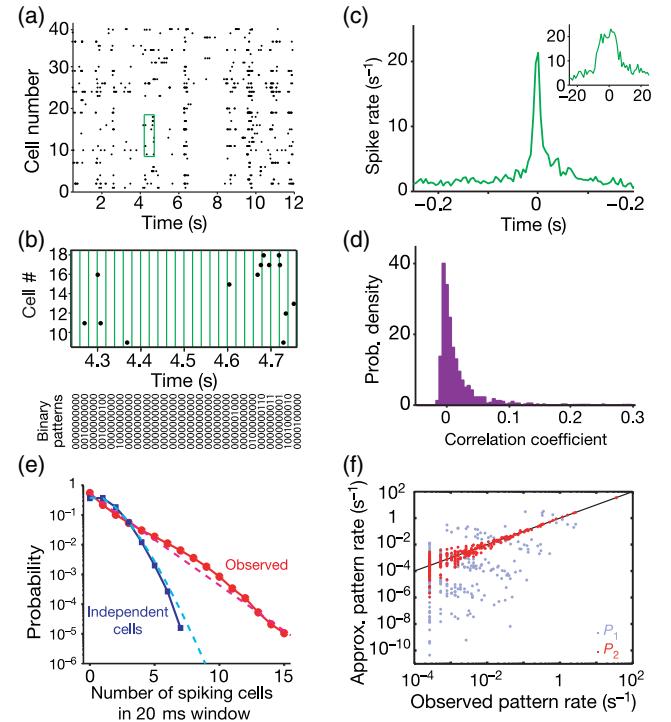


FIG. 7. Responses of the salamander retina to naturalistic movies. (a) Raster plot of the action potentials from  $N = 40$  neurons. Each dot represents a spike from one cell. (b) Expanded view of the green box in (a) showing the discretization of time into bins of width  $\Delta\tau = 20$  ms. The result (bottom panel) is that the state of the network is a binary word  $\{\sigma_i\}$ . (c) Correlations between two neurons. The results are shown as the probability per unit time of a spike in cell  $j$  (the spike rate) given that there is a spike in cell  $i$  at time  $t = 0$ . The plateau at long times should be the mean rate  $r_j = \langle \sigma_j \rangle / \Delta\tau$ . There is a peak with a width  $\sim 100$  ms, related to timescales in the visual input, and a peak with width  $\sim 20$  ms, as emphasized in the inset; this leads to the choice of bin size. (d) Distribution of (off-diagonal) correlation coefficients from Eq. (24) across the population of  $N = 40$  neurons. (e) Probability that  $K$  out of the  $N = 40$  neurons are active in the same time bin (red) compared with expectations if the activity of each neuron were independent of all the others (blue). Dashed lines are exponential (red) and Poisson (blue) and serve as a guide for the eye. (f) Predicted occurrence rates of different binary patterns vs the observed rates for the independent model  $P_1$  [Eqs. (29), (30); blue dots] and the pairwise maximum entropy model  $P_2$  [Eqs. (35), (33); red dots]. Adapted from Schneidman *et al.*, 2006.

In response to natural movies, the output neurons of the retina—the ganglion cells that carry visual signals from eye to brain and that as a group form the optic nerve—are sparsely activated, generating an average of just a few spikes per second each; see Figs. 7(a) and 7(b). Those initial experiments monitored populations of up to 40 neurons in a small patch of the retina, with recordings of up to 1 h. Pairs of neurons have temporal correlations with a relatively sharp peak or trough on a broad background that tracks longer timescales in the visual input; see Fig. 7(c). If we discretize time into bins of  $\Delta\tau = 20$  ms, then we capture most of the short time correlations but still have a low probability of seeing two

spikes in the same bin, so responses of neuron  $i$  become binary,<sup>9</sup>  $\sigma_i = \{0, 1\}$ .

If we define as usual the fluctuations around the mean,

$$\delta\sigma_i = \sigma_i - \langle\sigma_i\rangle, \quad (22)$$

then the datasets were large enough to get good estimates of the covariance

$$C_{ij} = \langle\delta\sigma_i\delta\sigma_j\rangle = \langle\sigma_i\sigma_j\rangle_c, \quad (23)$$

where  $\langle\cdots\rangle_c$  denotes the connected part. In many cases we have more intuition about the correlation matrix

$$c_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}}. \quad (24)$$

Importantly, these pairwise correlations are weak: almost all  $|c_{i\neq j}| < 0.1$ , and the bulk of these correlations are just a few percent; see Fig. 7(d). The recordings are long enough that these weak correlations are statistically significant, and almost none of the matrix elements are zero within errors. Correlations thus are weak and widespread, which seems to be common across many different regions of the brain.

If we look at only two neurons, the approximation that they are independent of one another is very good because the correlations are so weak. But if we look more globally, then the widespread correlations combine to have qualitative effects. As an example, we can ask for the probability that  $K$  out of  $N = 40$  neurons are active in the same time bin,  $P_N(K)$ , and we find that this has a much longer tail than would be expected if the cells were independent; see Fig. 7(e). Simultaneous activity of  $K = 10$  neurons already is  $\sim 10^3$  times more likely than in the independent model.

If we focus on  $N = 10$  neurons, then the experiments are long enough to sample all  $\Omega \sim 10^3$  states, and the probabilities of these different binary words depart dramatically from the predictions of an independent model; see Fig. 7(f). If we group the different binary words by the total number of active neurons, then the predictions of the independent model are actually anticorrelated with the real data. We emphasize that these failures occur despite the fact that pairwise correlations are weak and that they are visible at a relatively modest  $N = 10$ .

If we want to build a model for the patterns of activity in networks of neurons, it certainly makes sense to insist that we match the mean activity of each cell. At the risk of being pedantic, what this means explicitly is that we are looking for a probability distribution over network states  $P_1(\sigma)$  that has the maximum entropy while correctly predicting the expectation values,

$$m_i \equiv \langle\sigma_i\rangle_{\text{expt}} = \langle\sigma_i\rangle_{P_1}. \quad (25)$$

<sup>9</sup>The literature is mixed in sometimes choosing  $\sigma_i = \pm 1$  and sometimes  $\sigma_i = \{0, 1\}$ ; this choice is arbitrary. Here we use the  $\sigma_i = \{0, 1\}$  representation, which makes some things easier. In real neurons active and inactive states are emphatically not symmetric, so the elegance of the familiar  $\sigma_i = \pm 1$  is lost.

Referring back to Eq. (17), we find that the observables that we constrain become

$$\{f_\mu^{(1)}\} \rightarrow \{\sigma_i\}. \quad (26)$$

Note that  $i = 1, 2, \dots, N$ , where  $N$  is the number of neurons. To implement these constraints, we need one Lagrange multiplier for each neuron, and it is convenient to write this multiplier as an “effective field”  $h_i$  such that the general Eqs. (20) and (21) become

$$P_1(\sigma) = \frac{1}{Z_1} \exp[-E_1(\vec{\sigma})], \quad (27)$$

$$E_1(\vec{\sigma}) = \sum_\mu \lambda_\mu^{(1)} f_\mu^{(1)} \quad (28)$$

$$= \sum_{i=1}^N h_i \sigma_i. \quad (29)$$

We note that  $E_1$  is the energy function for independent spins in local fields, so the probability distribution over states factorizes,

$$P_1(\sigma) \propto \prod_{i=1}^N e^{-h_i \sigma_i}. \quad (30)$$

Thus, a maximum entropy model that matches only the mean activities of individual neurons is a model in which the activity of each cell is independent of all the others. We have seen that this model is in dramatic disagreement with the data.

A natural first step in trying to capture the nonindependence of neurons is to build a maximum entropy model that matches pairwise correlations. Thus, we are looking for a distribution  $P_2(\sigma)$  that has maximum entropy and matches the mean activities as in Eq. (25) and also the covariance of activity

$$C_{ij} \equiv \langle\delta\sigma_i\delta\sigma_j\rangle_{\text{expt}} = \langle\delta\sigma_i\delta\sigma_j\rangle_{P_2}. \quad (31)$$

In the language of Eq. (17), this means that we have a second set of relevant observables

$$\{f_\nu^{(2)}\} \rightarrow \{\sigma_i\sigma_j\}. \quad (32)$$

As before, we need one Lagrange multiplier for each constrained observable, and it is useful to think of the Lagrange multiplier that constrains  $\sigma_i\sigma_j$  as being a spin-spin coupling  $\lambda_{ij} = J_{ij}$ . When we recall that each extra constraint adds a term to the effective energy function, Eqs. (20) and (21) become

$$P_2(\sigma) = \frac{1}{Z_2(\{h_i; J_{ij}\})} e^{-E_2(\sigma)}, \quad (33)$$

$$E_2(\vec{\sigma}) = \sum_\mu \lambda_\mu^{(1)} f_\mu^{(1)} + \sum_\mu \lambda_\mu^{(2)} f_\mu^{(2)} \quad (34)$$

$$= \sum_{i=1}^N h_i \sigma_i + \frac{1}{2} \sum_{i\neq j} J_{ij} \sigma_i \sigma_j. \quad (35)$$

This is exactly an Ising model with pairwise interactions among the spins—not an analogy, but a mathematical equivalence.

Ising models for networks of neurons have a long history, as described in Sec. II.C. In their earliest appearance, these models emerged from a hypothetical, simplified model of the underlying dynamics. Here they emerge as the least structured models consistent with the measured properties of the network. As a result, we arrive not at some arbitrary Ising model, where we are free to choose the fields and couplings, but at a particular model that describes the actual network of neurons we are observing. To complete this construction, we have to adjust the fields and couplings to match the observed mean activities and correlations. Concretely, we have to solve Eqs. (25) and (31), which can be rewritten as

$$\langle \sigma_i \rangle_{\text{expt}} = \langle \sigma_i \rangle_{P_2} = \frac{\partial \ln Z_2(\{h_i; J_{ij}\})}{\partial h_i}, \quad (36)$$

$$\langle \sigma_i \sigma_j \rangle_{\text{expt}} = \langle \sigma_i \sigma_j \rangle_{P_2} = \frac{\partial \ln Z_2(\{h_i; J_{ij}\})}{\partial J_{ij}}. \quad (37)$$

With  $N = 10$  neurons this is challenging, but it can be done exactly since the partition function is a sum over just  $\Omega \sim 1000$  terms. Once we are done, the model is specified completely. Anything that we compute is a prediction, and there is no room to adjust parameters in search of better agreement with the data.

As noted, with  $N = 10$  neurons the experiments are long enough to get a reasonably full sampling of the probability distribution over  $\sigma$ . This provides the most detailed possible test of the model  $P_2$ , and in Fig. 7(f) we see that the agreement between theory and experiment is excellent, except for rare patterns where errors in the estimate of the probability are larger. Similar results are obtained for other groups of  $N = 10$  cells drawn out of the full population of  $N = 40$ . Quantitatively, we can measure the Jensen-Shannon divergence between the estimated distribution  $P_{\text{data}}(\sigma)$  and the model  $P_2(\sigma)$ . Across multiple choices of ten cells, this fluctuates by a factor of 2 around  $D_{\text{JS}} = 0.001$  bits, which means that it takes thousands of independent observations to distinguish the model from the data.

The architecture of the retina is such that many individual output neurons can be driven or inhibited by a single common neuron that is internal to the circuitry. This is one of many reasons that one might expect significant combinatorial regulation in the patterns of activity, and there were serious efforts to search for these effects (Schnitzer and Meister, 2003). The success of a pairwise model thus came as a considerable surprise.

The results in the salamander retina, with natural inputs, were quickly confirmed in the primate retina using simpler inputs (Shlens *et al.*, 2006). Those experiments covered a larger area and thus could focus on subpopulations of neurons belonging to a single class, which are arrayed in a relatively regular lattice. In this case not only did the pairwise model work very well, but the effective interactions  $J_{ij}$  were confined largely to nearest neighbors on this lattice.

Pairwise maximum entropy models also were reasonably successful in describing patterns of activity across  $N \leq 10$  neurons sampled from a cluster of cortical neurons kept alive

in a dish (Tang *et al.*, 2008). This work also pointed to the fact that dynamics did not correspond to Brownian motion on the energy surface.

These early successes with small numbers of neurons raised many questions. For example, the interaction matrix  $J_{ij}$  contained a mix of positive and negative terms, suggesting that frustration could lead to many local minima of the energy function or, equivalently, local maxima of the probability  $P(\sigma)$ , as in the Hopfield model; see Sec. II.C. Could these “attractors” have a function in representing the visual world? An important related consequence of the collective behavior in the Ising model is that if we know the state of all neurons in the network but one, then we have a parameter-free prediction for the probability that this last neuron will be active. Does this allow for error correction? To address these and other issues, one must go beyond  $N \sim 10$  cells, which was already possible experimentally. But at larger  $N$  one needs more powerful methods for solving the inverse problem that is at the heart of the maximum entropy construction, as described in Appendix B.

The equivalence to equilibrium models entices us to describe the couplings  $J_{ij}$  as interactions, but there is no reason to think that they correspond to genuine connections between cells. In particular,  $J_{ij}$  is symmetric because it is an effective interaction driving the equal-time correlations of activity in cells  $i$  and  $j$ , and these correlations are symmetric by definition. If we go beyond single time slices to describe trajectories of activity over time, then with multiple cells the effective interactions can become asymmetric and can break time-reversal invariance. Generalizations to stimulus-dependent maximum entropy model (see Sec. IV.D) introduce time variations, but the network is again statistically stationary in its response to stimuli. One could imagine extensions that capture learning or modulation of the network, perhaps coupling the time evolution of model parameters to changing input statistics through some normative principles or expressing learning rules by matching correlation structures over longer times. Going beyond stationary situations to building these extended models once seemed impossible, but the push for stable recordings of electrical activity over days and even weeks (see Sec. III) will create new opportunities.

The model defined by  $E_2$  is a spin glass, but it is a particular realization of a spin glass that is tuned to the measured behavior of neurons and neuron pairs. The cells appear to be on equal footing, but we do not assume homogeneity: different cells have different biases for activity or silence, and each cell has a unique pattern of interactions with all other cells. If cells can be grouped into types and interactions are type specific, then one might expect to see a block structure in the matrix  $J_{ij}$ , but this would be a result and not an assumption.

The idea of the Boltzmann machine is that purely pairwise interactions among neurons allow a subset of these cells (the visible units) to sample a much more complex distribution of states (Ackley, Hinton, and Sejnowski, 1985). Stated another way, the prediction is that populations of neurons that represent signals drawn from complex distributions should be part of larger networks that are described by simpler Ising models. Thus, if we could record from the entire network,

we would see a distribution over states that is well approximated by precisely the pairwise maximum entropy models that we are discussing here.<sup>10</sup>

Before leaving the early work, it is useful to step back and ask about the goals and hopes from that time. As previously reviewed, the use of statistical physics models for neural networks has a deep history. Saying that the brain is described by an Ising model captured both the optimism and (one must admit) the naivete of the physics community in approaching the phenomena of life. One could balance optimism and naivete by retreating to the position that these models are metaphors, illustrating what could happen rather than being theories of what actually happens. The success of maximum entropy models in the retina gave an example of how statistical physics ideas could provide a quantitative theory for networks of real neurons.

### C. Larger networks of neurons

The use of maximum entropy for networks of real neurons quickly triggered almost all of the following possible reactions: (a) It should never work, because systems are not in equilibrium, have combinational interactions, etc. (b) It could work, but only under uninteresting conditions. (c) It should always work since these models are highly expressive. (d) It works at small  $N$ , but this is a poor guide to what will happen at large  $N$ . (e) Sure, but why not use [favorite alternative], for which we have efficient algorithms?

Perhaps the most concrete response to these issues is simply to see what happens as we move to more examples, especially in larger networks. But we should do this with several questions in mind, some of which were explicit in the early literature (Roudi, Nirenberg, and Latham, 2009; Macke, Murray, and Latham, 2011). First, finding the maximum entropy model that matches the desired constraints—that is, solving Eq. (17)—becomes more difficult at larger  $N$ . Can we be sure that we are testing the maximum entropy idea and our choice of constraints rather than the efficacy of our algorithms for solving this problem?

Second, as  $N$  increases the maximum entropy construction becomes data hungry. This concern is often phrased as the usual problem of “overfitting,” when the number of parameters in our model is too large to be fully constrained by the data. But in the maximum entropy formulation, the problem is even more fundamental. The maximum entropy construction builds the least structured model consistent with a set of known expectation values. With a finite amount of data, if our list of expectation values is too long, then the claim that we “know” these features of the system simply is not true, and this problem arises even before we try to build the maximum entropy model.

Third, because correlations are spread widely in these networks, if one develops a perturbation theory around the

limit of independent neurons, then factors of  $N$  appear in the series, for example, for the entropy per neuron. Thus, success at modest  $N$  might mean that we are in a perturbative regime, which would be much less interesting. The question of whether success is perturbative is subtle, since at finite  $N$  all properties of the maximum entropy model are analytic functions of the correlations, and hence if we carry perturbation theory far enough, we will get the right answer (Sessak and Monasson, 2009).

Finally, in statistical mechanics we are used to the idea of a large  $N$  thermodynamic limit. Although this carries over to model networks (Amit, 1989), it is not obvious how to use this idea in thinking about networks of real neurons. Naive extrapolation of results from maximum entropy models of  $N = 10\text{--}20$  neurons in the retina indicated either that something special had to happen by  $N \sim 200$  or that the entropy would vanish. This was interesting because  $N \sim 200$  is the number cells that are “looking” at overlapping regions of the visual world (Schneidman *et al.*, 2006). A more sophisticated extrapolation imagines a large population of neurons in which mean activities and pairwise correlations are drawn at random from the same distribution as found in recordings from smaller numbers of neurons (Tkačík *et al.*, 2006, 2009). This sort of extrapolation is motivated in part by the observation that thermodynamic properties of the maximum entropy models learned for  $N = 20$  or 40 retinal neurons match the behavior of such random models at the same  $N$ . If we now extrapolate to  $N = 120$ , there are striking collective behaviors, and we ask whether these are seen in real data from  $N > 100$  cells.

Early experiments in the retina were already monitoring  $N = 40$  cells, and the development of numerical methods described in Appendix B quickly allowed these larger datasets to be analyzed (Tkačík *et al.*, 2006, 2009). With  $N = 40$  cells, one cannot check the predictions for probabilities of individual patterns  $P(\sigma)$ , but one can check the probability that  $K$  out of  $N$  cells are active in the same small time bin, as in Fig. 7(e), or the correlations among triplets of neurons. At  $N = 40$  we see the first hints that constraining pairwise correlations is not enough to capture the full structure of the network. There are disagreements between theory and experiment in the tails of the distribution  $P_N(K)$  and, more importantly, a few percent disagreement at  $K = 0$ . This may not seem like much, but since the network is completely silent in roughly half of the  $\Delta\tau = 20$  ms time bins, the data determine  $P_N(K = 0)$  precisely, and a 1% discrepancy is extremely significant.

A new generation of electrode arrays made it possible to record  $N = 100\text{--}200$  cells while densely sampling a small patch of the retina; see Sec. III.A. For example, these experiments could capture the signals from  $N_{\max} = 160$  ganglion cells in a  $(450 \mu\text{m})^2$  area of the salamander retina that contains a total of  $N \sim 200$  cells, and these recordings are stable for  $\sim 1.5$  h.

As explained in Appendix B, we can build maximum entropy models at larger  $N$  using a Monte Carlo simulation to estimate expectation values in the model, comparing them with the measured expectation values, and then adjusting the coupling constants to improve the agreement. This does not necessarily yield an exact solution to the constraint in Eq. (17),

<sup>10</sup>Immediately after the publication of Schneidman *et al.* (2006), Sejnowski forwarded an email exchange with Hinton in which they expressed their delight that real brains might be Boltzmann machines after all. This was of course a wildly optimistic extrapolation from the initial results. We thank Sejnowski for verifying this and giving permission to share the story.

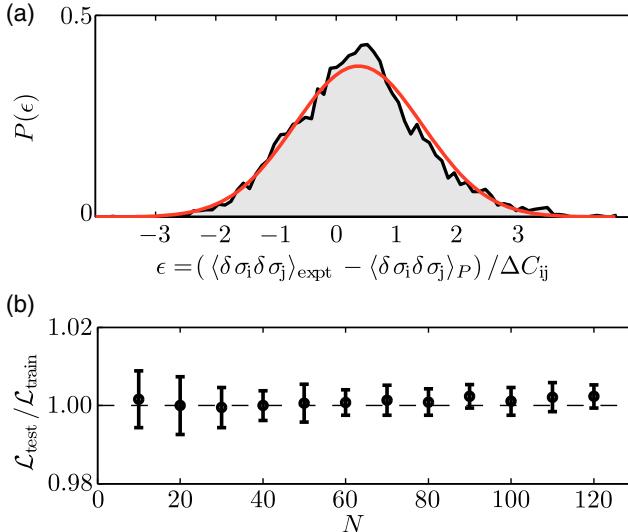


FIG. 8. Fitting, but not overfitting, with  $N \sim 100$  neurons. (a) Distribution of errors in the prediction of pairwise correlations, after adjusting the parameters  $\{h_i; J_{ij}\}$ , for  $N = 100$ . Prediction errors are in units of the measurement error  $\Delta C_{ij}$  for each element of the covariance matrix. The red line shows a Gaussian with zero mean and unit variance. (b) Log-likelihood [see Eq. (38)] of test data not used in constructing the maximum entropy model in units of the result for the training data. At  $N = 10$  it is not surprising that these agree, since the number of parameters  $\{h_i; J_{ij}\}$  is small. But we see that this agreement persists at approximately the 1% level out to  $N = 120$ , showing that even models for relatively large networks are not overfit. Adapted from Tkačik *et al.*, 2014.

but it seems acceptable since we are trying to match expectation values that are estimated from an experiment and they have errors. Figure 8(a) shows that with  $N = 100$  we can match the observed pairwise correlations within experimental error (Tkačik *et al.*, 2014). More precisely, the errors in predicting the elements of the covariance matrix  $C_{ij}$  [see Eq. (23)] are nearly Gaussian, with a variance equal to the variance of the measurement errors. This suggests, strongly, that one can successfully fit—but not overfit—a maximum entropy model to these data.

The test for fitting versus overfitting in Fig. 8(a) looks at each pair of cells individually, but part of the worry is that at large  $N$  we can have accurate estimates of individual elements  $C_{ij}$  while underdetermining the global properties of the matrix. We can take a familiar empirical approach, measuring the means  $\langle \sigma_i \rangle$  and covariances  $\langle \delta\sigma_i \delta\sigma_j \rangle_c$  in 90% of the data, using these to infer the parameters  $\{h_i; J_{ij}\}$  in a maximum entropy model, then testing the predictions of the model [see Eqs. (33) and (35)] on the remaining 10%. The fundamental measure of model quality is the log-likelihood of the data, which we can normalize per sample and per neuron as

$$\mathcal{L} = \frac{1}{N} \langle \log P(\sigma) \rangle_{\text{expt}}. \quad (38)$$

Figure 8(b) shows that  $\mathcal{L}$  is the same, to better than 1%, whether we evaluate it over the training data or over the test

data. This is true at  $N = 10$ , where there can be no question that we have enough samples, and it is true at  $N = 120$ .

Different networks of neurons in different organisms and different regions of the brain have different correlation structures. One should thus be wary of generalizations such as “an hour is enough data for 100 neurons.” But at least in the context of experiments on the retina, there is no question that maximum entropy models can be learned reliably from the available data, and that there is no overfitting. In other words, the models really are the solutions to the mathematical problem that we set out to solve (see Sec. IV.A): What is the minimal model consistent with a set of expectation values measured in an experiment? These models do not carry signatures of the algorithm that we used to find them, nor are they systematically perturbed by the finiteness of the data on which they are based. This answers the first two previously formulated questions.

Given that we can construct the maximum entropy models reliably, what do we learn? To begin, the small discrepancies in predicting the probability that  $K$  out of  $N$  neurons are active simultaneously,  $P_N(K)$ , become larger as  $N$  increases. The simplest solution to this problem is to add one more constraint, insisting that the maximum entropy model match the observed  $P_N(K)$  exactly. This adds only  $\sim N$  constraints to a problem in which we already have  $N(N+1)/2$ , so the resulting “ $K$ -pairwise” models are not significantly more complex.

Again, at the risk of being pedantic let us formulate the matching of the observed  $P_N(K)$  as constraining expectation values. If we introduce the Kronecker delta for integers  $n$  and  $m$ ,

$$\delta(n, m) = 1, \quad n = m, \quad (39)$$

$$\delta(n, m) = 0, \quad n \neq m, \quad (40)$$

then

$$P_N(K) = \left\langle \delta\left(K, \sum_i^N \sigma_i\right) \right\rangle. \quad (41)$$

Thus, to match  $P_N(K)$ , we want to enlarge our set of observables to include

$$\{f_\mu^{(\text{counts})}\} \rightarrow \left\{ \delta\left(K, \sum_i^N \sigma_i\right) \right\}. \quad (42)$$

As before, each new constraint adds a term to the effective energy,

$$E(\sigma) = \sum_\mu \lambda_\mu^{(\text{counts})} f_\mu^{(\text{counts})} = \sum_{K=0}^N \lambda_K \delta\left(K, \sum_i^N \sigma_i\right). \quad (43)$$

It is useful to think of this as an effective potential that acts on the summed activity,

$$\sum_{K=0}^N \lambda_K \delta\left(K, \sum_i^N \sigma_i\right) = V\left(\sum_i^N \sigma_i\right). \quad (44)$$

Putting the pieces together, the maximum entropy model that matches the mean activity of individual neurons, the

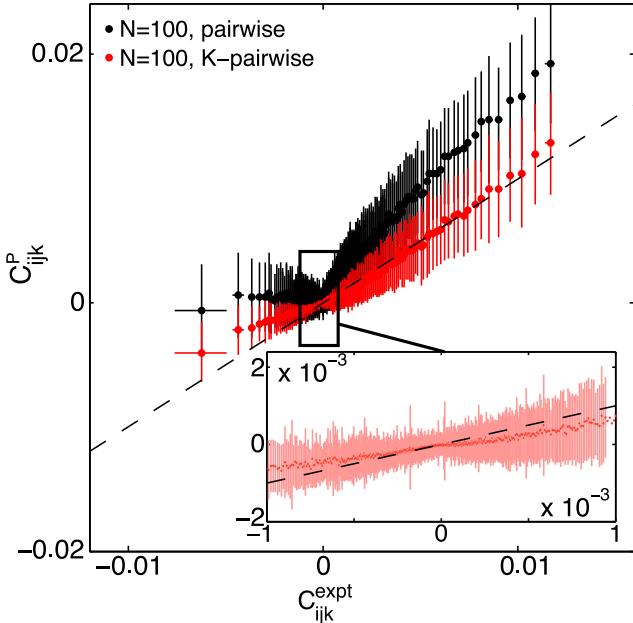


FIG. 9. Triplet correlations for  $N = 100$  cells in the retina. Measured  $C_{ijk}$  ( $x$  axis) vs value predicted by the model ( $y$  axis), shown for a single subgroup. The  $\sim 1.6 \times 10^5$  distinct triplets are grouped into 1000 equally populated bins. The error bars in  $x$  are the s.d. across the bin. The corresponding values for the predictions are grouped together, yielding the mean and the s.d. of the prediction ( $y$  axis). Inset: enlargement of the bulk of the predictions at small correlation for the  $K$ -pairwise model. The original reference used  $\sigma_i = \pm 1$ , so all  $C_{ijk}$  shown here are 8 times larger than they would be in the  $\sigma_i = \{0, 1\}$  representation. Adapted from Tkačik *et al.*, 2014.

correlations between pairs of neurons, and the probability that  $K$  out of  $N$  are active simultaneously takes the form

$$P_{2k}(\sigma) = \frac{1}{Z_{2k}} e^{-E_{2k}(\sigma)}, \quad (45)$$

$$E_{2k}(\sigma) = \sum_{i=1}^N h_i \sigma_i + \frac{1}{2} \sum_{i \neq j} J_{ij} \sigma_i \sigma_j + V \left( \sum_{i=1}^N \sigma_i \right). \quad (46)$$

We refer to this as the  $K$ -pairwise model (Tkačik *et al.*, 2014).

We can test this model immediately by estimating the correlations among triplets of neurons,

$$C_{ijk} = \langle (\sigma_i - \langle \sigma_i \rangle)(\sigma_j - \langle \sigma_j \rangle)(\sigma_k - \langle \sigma_k \rangle) \rangle. \quad (47)$$

Figure 9 shows the results with averages computed in both the pairwise and  $K$ -pairwise models, plotted versus the experimental values. The discrepancies are very small, although still roughly 3 times larger than the experimental errors in the estimates of the correlations themselves (Tkačik *et al.*, 2014). We see in Sec. V that one can sometimes get even better agreement. Note that the potential  $V$  that we add to match the constraint on  $P_N(K)$  does not carry any information about the identities of the individual neurons. It is thus interesting that including this term improves the predictions of all the triplet correlations, which do depend on neural identity.

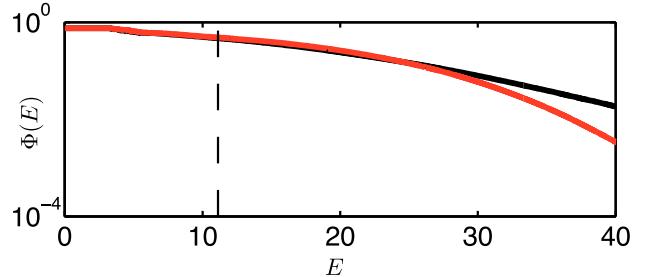


FIG. 10. The cumulative distribution of energies for  $N = 120$  neurons in the retina.  $\Phi(E)$  is defined in Eq. (48), and averages are over data (black line) or theory (red line). The dashed vertical line denotes an energy  $E_{2k}(\sigma)$  such that the particular state  $\sigma$  should occur once on average during the duration of the experiment. From Tkačik *et al.*, 2014.

With  $N = 100$  cells, we cannot check, as in Fig. 7(f), the probability of every state of the network. But the model assigns to every state an energy  $E_{2k}(\sigma)$ , and we can ask about the distribution of this energy over the states that we see in the experiment versus the expectation if states are drawn out of the model. To emphasize the extremes, we look at the high-energy tail,

$$\Phi(E) = \langle \Theta[E - E_{2k}(\sigma)] \rangle, \quad (48)$$

where  $\Theta(x)$  is the unit step function and the expectation value can be taken over the data or the theory. Figure 10 shows the comparison between theory and experiment. Note that the plot extends far past the point where individual states are predicted to occur once over the duration of the experiment, but we can make meaningful statements in this regime because there are exponentially many such states. Close agreement between theory and experiment extends out to  $E \sim 25$ , corresponding to states that are predicted to occur roughly once every 50 years.

This class of models predicts that neural activity will be collective. Thus, in a population of  $N$  cells, if we know the state of  $N - 1$ , we can make a prediction of the probability that the last cell will be active,<sup>11</sup>

$$P(\sigma_i = 1 | \{\sigma_{i \neq j}\}) = \frac{1}{1 + \exp[-h_i^{\text{eff}}(\{\sigma_{i \neq j}\})]}, \quad (49)$$

where we can think of the other neurons as applying an effective field to the one neuron that we focus on,

$$h_i^{\text{eff}}(\{\sigma_{i \neq j}\}) = E(\sigma_1, \sigma_2, \dots, \sigma_i = 1, \dots, \sigma_N) - E(\sigma_1, \sigma_2, \dots, \sigma_i = 0, \dots, \sigma_N). \quad (50)$$

For each neuron and each moment in time, we can calculate the effective field predicted by theory, with no free parameters, and we can group together all instances in which this field is in some narrow range and ask whether the probability of the cell

<sup>11</sup>The original presentation used  $\sigma_i = \pm 1$ , leading to a factor of 2 in the definition of the effective field; see Eq (25) of Tkačik *et al.* (2014).

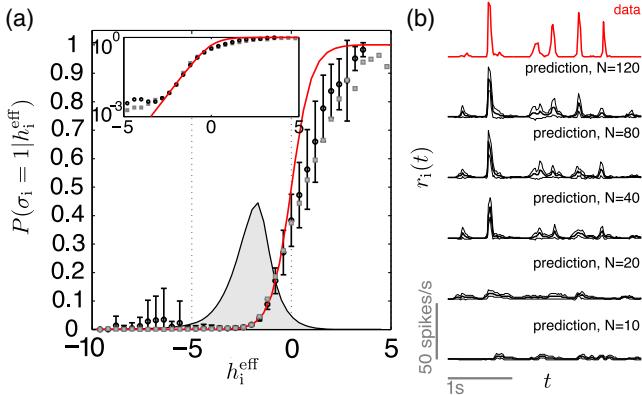


FIG. 11. Effective fields and the collective character of neural activity in the retina. (a) Probability that a single neuron is active given the state of the rest of the network, with  $N = 120$ . The points with error bars are the data, with the effective field computed from the model as in Eq. (50). The red line is the prediction from Eq. (49), and the gray points are the results from the purely pairwise rather than the  $K$ -pairwise model. The shaded gray region shows the distribution of fields across the experiment, emphasizing that the errors at large positive field are in the tail of the distribution. Inset: the same results on a logarithmic scale for probability. (b) Probability of a single neuron being active as a function of time in a repeated naturalistic movie, normalized as the probability per unit time of an action potential (spikes/s). Top (red) trace: experimental data. Lower (black) traces: predictions based on states of other neurons in an  $N$ -cell group based on Eqs. (49) and (50). The solid lines are the mean prediction across all repetitions of the movie, and the thin lines are the envelope  $\pm 1$  standard deviation. Adapted from Tkačik *et al.*, 2014.

being active agrees with Eq. (49). The results are shown in Fig. 11(a).

We see that the predictions of Eqs. (49) and (50) in the  $K$ -pairwise model agree well with experiment throughout the bulk of the distribution of effective fields, but also that discrepancies arise in the tails. These deviations are  $\sim 1.5$  times the error bars of the measurement but have some systematic structure, suggesting that we are capturing much but not all of the collective behavior under conditions where neurons are driven most strongly.

The results in Fig. 11(a) combine data across all times to estimate the probability of activity in one cell given the state of the rest of the network. It is interesting to unfold these results in time. In particular, the structure of the experiment was such that the retina saw the same movie many times, so we can condition on a particular moment in the movie, as shown for one neuron in Fig. 11(b). It is conventional to plot not the probability of being active in a small bin but the corresponding “rate” (Rieke *et al.*, 1997),

$$r_i(t) = \langle \sigma_i(t) \rangle / \Delta\tau, \quad (51)$$

where  $\sigma_i(t)$  denotes the state of neuron  $i$  at time  $t$  relative to (in this case) the visual inputs. We see in the top trace of Fig. 11(b) that single neurons are rarely active, with essentially zero probability of spiking between brief transients that generate on average one or a few spikes. This pattern is

common in response to naturalistic stimuli and is difficult to reproduce in models (Maheswaranathan *et al.*, 2023).

The maximum entropy models provide an extreme opposite point of view, making no reference to the visual inputs; instead activity is determined by the state of the rest of the network. We see that this approach correctly predicts sparse activity, with near zero rate between transients that are timed correctly relative to the input. Although here we see only one cell, the average neuron exhibits an  $r_i(t)$  that has  $\sim 80\%$  correlation with the theoretical predictions at  $N = 120$ . There is no sign of saturation, and it seems likely that we would make even more precise predictions from models based on all  $N \sim 200$  cells in this small patch of the retina. The possibility of predicting activity without reference to the visual input suggests that the “vocabulary” of the retina’s output is restricted and that, as with spelling rules, this should allow for error correction (Loback *et al.*, 2017).

Perhaps the most basic prediction from maximum entropy models is the entropy itself. There are several ways that we can estimate the entropy. To begin, in the  $K$ -pairwise model, we can see that the effective energy of the completely silent state from Eq. (46) is zero, which means that the probability of this state is simply the inverse of the partition function. Further, in this model the probability of complete silence matches what we observe experimentally. Thus, we can estimate the free energy of the model from the data, and then we can estimate the mean energy of the model from Monte Carlo, giving us an estimate of the entropy. An alternative is to generalize the model by introducing a fictitious temperature, as discussed in Sec. VI.B. Then at  $T = 0$  the entropy must be zero and at  $T \rightarrow \infty$  the entropy must be  $N \log 2$ , while the derivative of the entropy is related as always to the heat capacity. Thus, the entropy of our model for the real system at  $T = 1$  becomes<sup>12</sup>

$$S_N(T=1) = \int_0^1 dT \frac{C_v(T)}{T} = N \log 2 - \int_1^\infty dT \frac{C_v(T)}{T}, \quad (52)$$

where the heat capacity is related as usual to the variance of the energy,  $C_v = \langle (\delta E)^2 \rangle / T^2$ , which we can estimate from Monte Carlo simulations at each  $T$ . There is also a check requiring the two estimates in Eq. (52) to agree. All of these methods agree with one another at the percent level, with the results shown in Fig. 12(a).

The  $\sim 25\%$  reduction in entropy is significant, but more dramatic (and testable) is the prediction that the distribution over states is extremely inhomogeneous. Recall that if the distribution is uniform over some effective number of states  $\Omega_{\text{eff}}$ , then the entropy is  $S = \log \Omega_{\text{eff}}$  and the probability that two states chosen at random will be the same is  $P_c = 1/\Omega_{\text{eff}}$ . For nonuniform distributions we have  $S \geq -\log(P_c)$ . If neurons were independent, then with  $N$  cells we would have  $P_c \propto e^{-\alpha N}$ , and this is what we see in the data once they are shuffled to remove correlations; see Fig. 12(b). But the real data show a much more gradual decay with  $N$ ,

<sup>12</sup>We write  $C_v$  not because we are concerned about whether the volume is constant, but to avoid confusion with the covariance matrix  $C_{ij}$ .

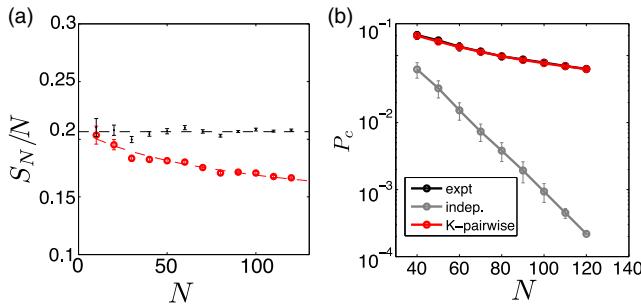


FIG. 12. Entropy and coincidences in the activity of the retinal network. (a) Entropy predicted in  $K$ -pairwise models (red line) and in the approximation that all neurons are independent (gray line). Models are constructed independently for many subgroups of size  $N$  chosen out of the total population  $N_{\max} = 160$ , and error bars include the variance across these groups. (b) Probability that two randomly chosen states of the network are the same, again for many subgroups of size  $N$ . The displayed results are for real data (black line), shuffled data (gray line), and the  $K$ -pairwise models (red lines). Adapted from Tkačik *et al.*, 2014.

and this is captured perfectly by the  $K$ -pairwise maximum entropy models.

At  $N = 120$  the logarithm of the coincidence probability (both measured and predicted) is an order of magnitude smaller than the entropy predicted by the model. Perhaps related is that the free energy per neuron—which, as discussed, can be obtained directly from the probability of the fully silent state—also decreases dramatically as  $N$  increases. At  $N = 120$  the free energy is just a few percent of either the entropy or the mean energy, reflecting near-perfect cancellation between these terms; one can also see this in a much simpler model that matches only  $P_N(K)$  and not the individual means or pairwise correlations (Tkačik *et al.*, 2013). These behaviors are captured by the  $K$ -pairwise model smoothly from  $N < 40$  through  $N > 100$ , indicating that what we learned at more modest  $N$  indeed does extrapolate up to a scale comparable to the entire population of cells in a patch of the retina. We will have to work harder to decide whether we can see the emergence of a true thermodynamic limit.

Finally, we address the question as to whether these results can be recovered as perturbations to a model of independent neurons. At lowest order in perturbation theory, there is a simple relationship between the observed correlations and the inferred interactions  $J_{ij}$  in the pairwise model (Sessak and Monasson, 2009), and we can check this relationship against the values of  $J_{ij}$  inferred from correctly matching the observed correlations. In the retina large deviations from lowest-order perturbation theory are already visible at  $N = 15$ , and correspondingly models built from the perturbative estimates of  $J_{ij}$  are orders of magnitude further away from the data than the full model (Tkačik *et al.*, 2014). Higher-order perturbative contributions to the entropy would be comparable to one another for  $N = 20$  retinal neurons even in a hypothetical network where all correlations were scaled down by a factor of 2 from the real data (Azhar and Bialek, 2010). We conclude that the success of maximum entropy models in describing networks of real neurons is not something we can understand in low-order perturbation theory. Interestingly, simulations of

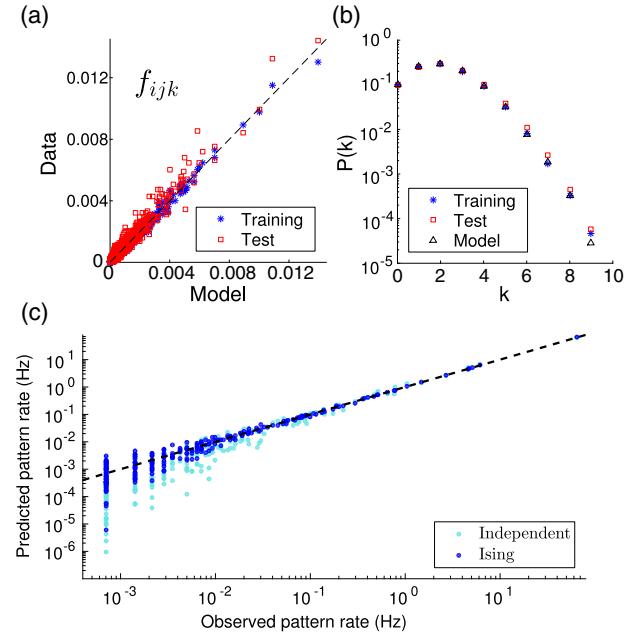


FIG. 13. Pairwise maximum entropy models describe collective behavior of  $N = 37$  neurons in the prefrontal cortex. (a) Observed vs predicted triplet correlations among all neurons. The training results (blue symbols) are predictions from the same segment of the experiment where the pairwise correlations were measured; the test results (red symbols) are in a different segment of the experiment. (b) Probability that  $K$  out of  $N$  neurons are active simultaneously, comparing predictions of the model with data in training and test segments. (c) Rate at which patterns of spiking and silence appear in a subset of ten neurons, with predicted vs observed rates compared in an independent model (cyan dots) and in the pairwise model (blue dots). Adapted from Tavoni *et al.*, 2017.

models with pure three- and four-spin interactions at  $N \sim 20$  show that pairwise maximum entropy models typically are good approximations to the real distribution both in the weak correlation limit and in the limit of strong, dense interactions (Merchan and Nemenman, 2016).

The retina is a special part of the brain, and one might worry that the success of maximum entropy models is somehow tied to these special features. It thus is important that the same methods work in capturing the collective behavior of neurons in distant parts of the brain. An example is in the prefrontal cortex, which is involved in a wide range of higher cognitive functions.

Experiments recording simultaneous activity from several tens of neurons in the prefrontal cortex were analyzed with maximum entropy methods; an example of the results is shown in Fig. 13 (Tavoni *et al.*, 2017). We see that these models pass the same tests as in the retina, correctly predicting triplet correlations, the probability of  $K$  out of  $N$  cells being active simultaneously, and the probabilities for particular patterns of activity in subgroups of  $N = 10$  cells. When extending this analysis across multiple experimental sessions, it was possible to detect changes in the coupling matrix  $J_{ij}$  as the animal learned to engage in different tasks. These changes were concentrated in subsets of cells that also were

preferentially reactivated during sleep between sessions. One should be careful about giving too mechanistic an interpretation of the Ising models that emerge from these analyses, but it is exciting to see the structure of the models connect to independently measurable functional dynamics in the network. This is true even in the farthest reaches of the cortex, the regions of the brain that we use for thinking, planning, and deciding.

The Ising model also gives us a way of exploring how the network would respond to hypothetical perturbations (Tavoni, Cocco, and Monasson, 2016). If we increase the magnetic field uniformly across all the cells in the population of prefrontal neurons, the predicted changes in activity are far from uniform. For some cells the response and the derivative of the response (susceptibility) are on a scale expected if neurons respond independently to applied fields, but there are groups of cells that coactivate much more, with susceptibilities peaking at intermediate fields. It is tempting to think that these groups of cells have some functional significance, and this is supported by the fact that in the real data (with no fictitious fields) the groups of cells identified in this way remain coactivated over relatively long periods of time.

At the opposite extreme of organismal complexity, the worm *C. elegans* is an attractive target for these analyses because one can record not only from a large number of cells but also from a large fraction of the entire brain at single cell resolution; see Sec. III.C. A major challenge is that these neurons do not generate discrete action potentials or bursts, so the signals are not naturally binary. A first step was to discretize the continuous fluorescence signals into three levels and construct a Potts-like model that matched the population of each state and the probabilities that pairs of neurons are in the same state (Chen *et al.*, 2019). Although these early datasets were limited, this simple model succeeded in predicting off-diagonal elements of the correlation matrix that were unconstrained, the probability that  $K$  of  $N$  neurons are in the same state, and the relative probabilities of different states in relation to the effective fields generated by the rest of the network. The fact that the same statistical physics approaches work in worms and in the mammalian cortex is encouraging, though we should see more compelling tests with the next generation of experiments.

A much different approach is to study networks of neurons that have been removed from the animal and kept alive in a dish. There is a long history of work on these “cultured networks,” and as previously noted (see Sec. III.A), some of the earliest experiments recording from many neurons were done with networks that had been grown onto an array of electrodes (Pine and Gilbert, 1982). Considerable interest was generated by the observation that patterns of activity in cultured networks of cortical neurons consist of “avalanches” that exhibit at least some degree of scale invariance; see Sec. VI.A. Recent work returns to these data and shows that detailed patterns of spiking and silence are well described by pairwise maximum entropy models, reproducing triplet correlations and the probability that  $K$  out of  $N = 60$  neurons are active simultaneously (Sampaio Filho *et al.*, 2024).

As a final example, we consider populations of  $N \sim 100$  neurons in the mouse hippocampus (Meshulam *et al.*, 2017). The hippocampus plays a central role in navigation and

episodic memory and is perhaps best known for its population of “place cells,” neurons that are active only when the animal moves to a particular position in its environment. First discovered in rodents (O’Keefe and Dostrovsky, 1971), it is thought that the entire population of these cells together provides the animal with a cognitive map (O’Keefe and Nadel, 1978). More recent work shows how this structure extends to three dimensions, and across hundreds of meters in bats (Tsoar *et al.*, 2011; Yartsev and Ulanovsky, 2013).

As the animal explores its environment or runs along a virtual track,<sup>13</sup> the mean activity of individual neurons is quite small, as in the previous examples. Most pairs of neuron have negative correlations, as expected if activity is tied to the position—if each cell is active in a different place, then on average one cell being active means that other cells must be silent, generating anticorrelations. Indeed, it is tempting to make a model of the hippocampus in which some positional signal is computed by the brain, with inputs from many regions, and each cell in the hippocampus is either active or silent, depending on the value of this positional signal. This model is specified by the “place fields” of each cell, the probability that a cell is active as a function of position, and these can be estimated directly from the data. Given the place fields, all other properties of the network are determined with no adjustable parameters.

The place field of cell  $i$  is defined by the average activity conditional on the position  $x$  along a track,

$$\langle \sigma_i \rangle_x = F_i(x). \quad (53)$$

If activity in each cell depends independently on position, then the pairwise correlations are driven by the fact that all cells experience the same  $x$ , drawn from some distribution  $P(x)$  across the experiment. The quantitative prediction is that

$$\begin{aligned} C_{ij} &\equiv \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle \\ &= \int dx P(x) F_i(x) F_j(x) \\ &\quad - \left[ \int dx P(x) F_i(x) \right] \left[ \int dx P(x) F_j(x) \right]. \end{aligned} \quad (54)$$

The covariance matrix elements  $C_{ij}$  have a pattern that is qualitatively similar to the real data but quantitatively far off. In particular, the eigenvalue spectrum of the matrix predicted in this way falls rapidly, while the real spectrum has a slow, nearly power-law decay (Meshulam *et al.*, 2017). This is a first hint that the neurons in the hippocampal network share information, and hence exhibit collective behavior, beyond simply place.

<sup>13</sup>The hippocampus is a prime example of a neural network that learns. But the experiments analyzed here involve the mouse spending ~30 min in a virtual environment with which it is already familiar. It thus seems likely that there is little if any learning about the environment across the dataset, and indeed the data pass various tests for stationarity. Perhaps we should say that the animal *moves through* rather than *explores* its environment.

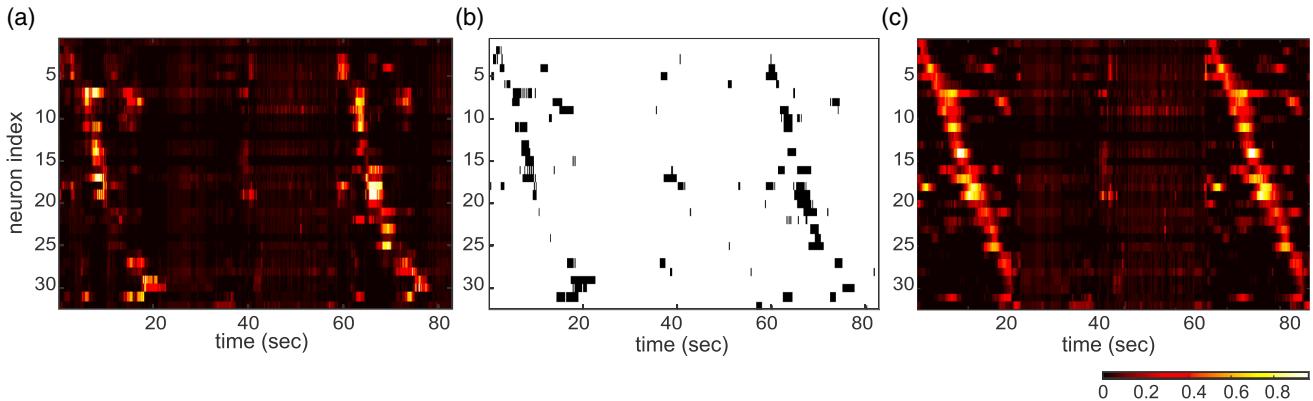


FIG. 14. Collective behavior in the mouse hippocampus. (a) Predicted probability of activity for single neurons, computed from the effective field in the pairwise maximum entropy model. The focus is on 32 place cells that should be active in sequence as the mouse runs along a virtual track. During the first run, cells 21–25 are predicted to “miss” their place fields, but all cells are predicted to be active in the second run. (b) Real data of place cell activity during two runs down the linear track, in the same time window as (a) and (c). Note the missed events for cells 21–25 in the first run. (c) Predicted probability from the independent place cell model. There is no indication of when fields should be missed. From [Meshulam et al., 2017](#).

A new generation of experiments monitoring more than 1000 neurons in the hippocampus provides unique opportunities for theory, as discussed in Secs. V and VII. Here we want to emphasize the way in which collective dynamics emerge from maximum entropy models of  $N \sim 100$  cells. Equations (49) and (50) and Fig. 11 remind us that models for the joint distribution of activity in a neural population also predict the probability for one neuron to be active given the state of the rest of the network. We can go through the same exercise for a population of cells in the hippocampus: construct the pairwise maximum entropy model and, for each neuron at each moment, compute the probability that it will be active given the state of all the other neurons; results are displayed in Fig. 14(a).

We see in Fig. 14(a) that, roughly speaking, cells are predicted to be active in sequence. This makes sense since these are place cells and the mouse is running at nearly constant speed along a virtual track, so cells with place fields arrayed along the track should be activated one after the other. Interestingly, the calculation leading to this prediction makes no reference to the virtual position of the mouse, or even to the idea of place fields, but only to the dependence of activity in one cell on the rest of the network. In this window of time, the mouse actually makes two trips along the track and, perhaps surprisingly, the predictions for the two trips are different. On the first trip, it is predicted that several of the cells will “miss” their place fields, while all cells should be active in sequence on the second trip. This is exactly what we see in the data; see Fig. 14(b). If neurons were driven only by the animal’s position this would not happen; see Fig. 14(c). Thus, what might have seemed like unpredictable variation really reflects the collective behavior of the network and is captured very well by the Ising model, with no additional parameters. We return to Ising models for the hippocampus in Sec. V.

#### D. Doing more and doing less

Is there any sense in which maximum entropy models are “better” than alternative models? The pairwise maximum

entropy models are singled out because they have the minimal structure needed to match the mean activity and two-point correlations in the network. But how different are they from other models that would also match these data? We could imagine, for example, that once we specify the full matrix of correlations, then the set of allowed models is tightly clustered in its predictions about higher-order structure in the patterns of activity, in which case saying that these models work does not say much about the underlying physics.

One can build a statistical mechanics on the space of probability distributions  $P(\sigma)$ , defining a *version space* as all the models that match a given set of pairwise correlations within some tolerance  $\epsilon$ . We can construct a Boltzmann weight over this space in which the entropy of the underlying distribution plays the role of the negative energy,

$$Q[P(\sigma)] \propto \delta \left[ 1 - \sum_{\sigma} P(\sigma) \right] U_{\epsilon}[P(\sigma); \{m_i, C_{ij}\}] \times \exp \left[ -\beta \sum_{\sigma} P(\sigma) \ln P(\sigma) \right], \quad (55)$$

where the first term in the product enforces normalization, the second term selects distributions that match expectation values within  $\epsilon$ , and the last term is the Boltzmann weight (Obuchi, Cocco, and Monasson, 2015). Note that this is the maximum entropy distribution of distributions (!) consistent with a particular mean value of the entropy and the measured expectation values. As  $\beta \rightarrow \infty$ ,  $Q$  condenses around the maximum entropy distribution, while as  $\beta \rightarrow 0$ , all distributions consistent with the expectation values are given equal weight.

If the matrix  $C_{ij}$  is chosen at random then one can use methods from the statistical mechanics of disordered systems to develop an analytic theory that compares the similarity of the true distribution to those drawn at random from the ensembles of distributions at varying  $\beta$ . In this random setting, the maximum entropy models are not special, and in a rough sense all models that match the low-order correlations are equally good approximations (Obuchi, Cocco, and

Monasson, 2015). Networks of real neurons have some particular correlations  $C_{ij}$  that are not typical in the ensemble of random  $C_{ij}$ . Indeed, we see in Sec. VI that these correlations are far from generic. Ferrari, Obuchi, and Mora (2017) explicitly tested the predictions of maximum entropy models versus the average model that matches  $C_{ij}$  for a population of retinal neurons. They found that the maximum entropy model gives a better description, and this advantage grows with  $N$ . They concluded that “for large networks it is better to pick the most random [maximum entropy] model than to pick a model at random.”

One way that we could do more in describing the patterns of neural activity is to address their time dependence explicitly. In particular, for the retina we know that the network is being driven by visual inputs. We can repeat the movie many times and ask about the mean activity of each cell at a given moment in the movie,  $\langle \sigma_i(t) \rangle$ . In addition, as before, we can measure the correlations between neurons at the same moment in time,  $\langle \sigma_i(t) \sigma_j(t) \rangle$ . Thus, we want to find a model for the distribution over sequences, or *trajectories*, of network states  $P_{\text{traj}}[\sigma(t)]$  that has maximal entropy and matches the time-dependent mean activity

$$m_i(t) \equiv \langle \sigma_i(t) \rangle_{\text{expt}} = \langle \sigma_i(t) \rangle_{P_{\text{traj}}}, \quad (56)$$

as well as the time averaged equal-time correlations

$$C_{ij} \equiv \frac{1}{\tilde{T}} \sum_t \langle \delta \sigma_i(t) \delta \sigma_j(t) \rangle_{\text{expt}} \quad (57)$$

$$= \frac{1}{\tilde{T}} \sum_t \langle \delta \sigma_i(t) \delta \sigma_j(t) \rangle_{P_{\text{traj}}}, \quad (58)$$

where  $\tilde{T}$  is the duration of our observations in units of the time bin width  $\Delta\tau$ .

This is an instance of the general structure presented in Sec. IV.A, where the first set of observables is of the form

$$\{f_\mu\} \rightarrow \{f_{i,t}\} = \{\sigma_i(t)\}. \quad (59)$$

To constrain the expectation value of each of these terms, we need a separate Lagrange multiplier, and as before we think of these as local fields that now depend on time,  $\lambda_{i,t} = h_i(t)$ . In addition, we have observables of the form

$$\{f_\mu\} \rightarrow \{f_{ij}\} = \left\{ \sum_t \sigma_i(t) \sigma_j(t) \right\}, \quad (60)$$

and for each of these we again have a separate Lagrange multiplier that we think of as a spin-spin coupling,  $\lambda_{ij} = J_{ij}$ . The general equations (20) and (21) now take the form

$$P_{\text{traj}}[\sigma(t)] = \frac{1}{Z_{\text{traj}}} \exp \{-E_{\text{traj}}[\sigma(t)]\}, \quad (61)$$

$$E_{\text{traj}}[\sigma(t)] = \sum_t \sum_i h_i(t) \sigma_i(t) + \frac{1}{2} \sum_t \sum_{ij} J_{ij} \sigma_i(t) \sigma_j(t). \quad (62)$$

In this class of models, correlations arise both because different neurons may be subject to correlated time-dependent fields and because of effects intrinsic to the network. If all of the correlations were driven by visual inputs, then matching the correlations would lead to  $J_{ij} = 0$ , but this never happens with real data.

One can go further and assume some form for the relation between the time-dependent field  $h_i(t)$  and the visual inputs. The simplest possibility is that the field is a spatially and temporally filtered version of the light intensity pattern shown to the retina (Granot-Atedgi *et al.*, 2013). Explicitly, if the light intensity varies in space and time as  $I(\vec{x}, t)$ , each neuron  $i$  is described by a linear kernel or receptive field  $K_i(\vec{x}, \tau)$  such that

$$h_i(t) = h_i^0 + \int d^2x \int d\tau K_i(\vec{x}, \tau) I(\vec{x}, t - \tau). \quad (63)$$

Such stimulus-dependent maximum entropy models can be seen as generalizations of the widely used “linear-nonlinear” models for single neurons (Dayan and Abbott, 2001). These also are the maximum entropy models consistent with the correlation between single neurons and the movie itself,  $\langle \sigma_i(t) I(\vec{x}, t - \tau) \rangle$ , which can be estimated without having to repeat the movie.

An alternative is to determine the time-dependent fields from experiments with a repeated movie and then fit a separate model to the dependence of the field on the input movie (Ferrari *et al.*, 2018). This two-step procedure has the advantage that incompleteness of the model for the stimulus dependence of the field does not influence the estimates of the interactions  $J_{ij}$ . Indeed, the fact that we can predict the activity of single neurons in the retina from other neurons, without reference to the visual input (see Fig. 11), means that the problem of disentangling stimulus dependence of the fields from true interactions is nontrivial.

An interesting question is how the decomposition into field and interactions connects to the distribution of sensory inputs. We know that single neurons adapt their apparent input-output relationships to the input statistics, perhaps in ways that maximize the magnitude or efficiency of the sensory information that is conveyed by the resulting sequence of action potentials,<sup>14</sup> and there are generalizations of this idea to populations of neurons (Tkačík *et al.*, 2010). In the language of the stimulus-dependent maximum entropy models, this suggests that the mapping from sensory inputs to the fields  $h_i(t)$  will change when we change the distribution of inputs. But what happens to the interactions  $J_{ij}$ ? Recent work on the retina suggests that the interaction matrix may be largely invariant across different ensembles of input movies (Hoshal *et al.*, 2023). Despite the fact that the interactions themselves do not vary with the input ensemble, their presence enhances the reliability with which brief segments of the neural response can be used to make choices among a set of possible ensembles.

Our discussion began with models that match the mean activity of each neuron and their pairwise correlations, with

---

<sup>14</sup>For a recent review, see Bialek (2024).

these correlations measured at equal times, resulting in Eqs. (33) and (35). The stimulus-dependent models capture time-dependent mean activity, where time is measured relative to the sensory inputs, but still match only the equal-time correlations. A natural alternative is to capture time-dependent correlations but simplify by matching only the global mean activity. Concretely, this means that we want to find a maximum entropy model for sequences or trajectories of states,

$$P_{\text{traj}2}[\sigma(t)] = \frac{1}{Z_{\text{traj}2}} \exp \{-E_{\text{traj}2}[\sigma(t)]\}, \quad (64)$$

where the subscript reminds us that this is a second model for trajectories. We want to match experimental observations of the mean activity, averaging over its time dependence

$$\bar{m}_i \equiv \frac{1}{T} \sum_t \langle \sigma_i(t) \rangle_{\text{expt}} = \langle \sigma_i(t) \rangle_{P_{\text{traj}2}}. \quad (65)$$

In addition, we want to match the pairwise correlations across time,

$$\begin{aligned} C_{ij}(\tau) &\equiv \frac{1}{T} \sum_t \langle \delta\sigma_i(t) \delta\sigma_j(t + \tau) \rangle_{\text{expt}} \\ &= \langle \delta\sigma_i(t) \delta\sigma_j(t + \tau) \rangle_{P_{\text{traj}2}}, \end{aligned} \quad (66)$$

where as before  $\delta\sigma_i(t) = \sigma_i(t) - \bar{m}_i$ . We assume that the system is statistically stationary, so correlations depend only on time differences.

This is another instance of the general structure presented in Sec. IV.A, where one set of observables is of the form

$$\{f_\mu^{(\text{means})}\} \rightarrow \left\{ \sum_t \sigma_i(t) \right\}, \quad (67)$$

and a second set of observables is of the form

$$\{f_\mu^{(\text{pairs},\tau)}\} \rightarrow \left\{ \sum_t \sigma_i(t) \sigma_j(t + \tau) \right\}. \quad (68)$$

As before, we identify an effective field  $h_i = \lambda_i$  as the Lagrange multiplier constraining the means of individual neurons, and the Lagrange multipliers that constrain the covariance of pairs of neurons separated by a time  $\tau$  can now be identified as time-dependent couplings  $J_{ij}(\tau)$ . The general equation (21) thus becomes

$$E_{\text{traj}2}[\sigma(t)] = \sum_t \sum_i h_i \sigma_i(t) + \frac{1}{2} \sum_{t\tau} \sum_{ij} \sigma_i(t) J_{ij}(\tau) \sigma_j(t + \tau). \quad (69)$$

While Eq. (69) is a natural counterpoint to the model of Eq. (62), it has been less widely explored.

Perhaps the most important feature of this class of models is that it gives us a chance to explore the breakdown of time-reversal invariance in neural activity. Note that in Eq. (69) we can swap indices on neurons and time together such that

$i, t \leftrightarrow j, t'$ , and this leaves the energy  $E_{\text{traj}2}$  unchanged. This means that  $J_{ij}(\tau) = J_{ji}(-\tau)$ . But time-reversal invariance would require  $J_{ij}(\tau) = J_{ij}(-\tau)$ , which might not be the result of solving the matching conditions in Eq. (66). This emphasizes the conceptual point that we can have maximum entropy models for systems whose dynamics violate detailed balance. It also opens a path to investigating more concretely how patterns of neural activity represent the arrow of time (Lynn et al., 2022a, 2022b).

In some cases pairwise maximum entropy models are not enough to capture the full structure of the network. A simple idea is that we need more constraints. We see how this worked in the retina, where fixing the distribution  $P_N(K)$  allowed for much closer agreement with the experiment; see Sec. IV.C. An alternative is that we do not need more terms, just different terms. Are there different paths to simplification, or at least to understanding why simplification is possible?

A key feature of pairwise models is that they involve matching  $\sim N^2$  features of the data, many fewer than the  $\sim 2^N$  parameters that would be required to describe an arbitrary probability distribution. But as the experimentally accessible  $N$  increases, eventually even  $\sim N^2$  becomes too large, and we cannot reliably determine the entire matrix of correlations. If we want to keep to the maximum entropy strategy, we have to find a way of working with fewer constraints, ideally  $\sim N$ .

An early idea was that, instead of matching the full matrix of correlations, we could match the distribution of these correlations across all pairs, which can be estimated more reliably (Castellana and Bialek, 2014). This problem actually is the inverse of the usual spin glass: rather than choosing interactions  $J_{ij}$  from a distribution and computing correlations, we choose the correlations from a distribution and infer the interactions. If we fix only the first two moments of the distribution of correlations, then the interactions develop a block structure reminiscent of the hierarchy of correlations in replica symmetry breaking (Mézard, Parisi, and Virasoro, 1987). We can find a phase diagram in the space of these moments that crucially depends on how they scale with  $N$ . As experiments progress from  $N \sim 100$  to  $N \sim 10^4$  and even  $N \sim 10^6$  (see Sec. III.C), it seems likely that this approach of constraining distributions will become more useful.

Another approach is to go back to the basic maximum entropy formulation but choose only a limited set of quantities whose expectation values we should constrain. These quantities need not be simple objects such as  $\sigma_i$  or  $\sigma_i \sigma_j$ . As an example, we could imagine a neuron somewhere else in the brain that takes inputs from the network we are studying, sums these inputs, and compares them with a threshold, as in the model of McCulloch and Pitts (1943) described in Sec. II.A. Concretely, we can consider the activity of these hypothetical neurons,

$$y_\mu = \Theta \left( \sum_{i=1}^N W_{\mu i} \sigma_i - \theta_\mu \right), \quad (70)$$

and ask for the maximum entropy distribution that matches the expectation values of  $\{y_\mu\}$  that we compute from the data.

Following the aforementioned arguments, this distribution has the form

$$P_{\text{out}}(\sigma) = \frac{1}{Z_{\text{out}}(\{g_\mu\})} \exp \left[ - \sum_{\mu=1}^K g_\mu y_\mu \right], \quad (71)$$

where the subscript reminds us that this model focuses on a possible output of the network rather than directly on the network state itself. We can set  $\theta_\mu = 1$  by choosing units for the weights. Note that the number of parameters  $\{g_\mu; W_{\mu i}\}$  is then  $(N+1)K$ , which is much less than  $N^2$  if the number of output neurons  $K \ll N$ . If we could view the weights and thresholds  $\{W_{\mu i}\}$  as given, then we would have only  $K$  parameters, set by the expectation values  $\{y_\mu\}$ , and we could even let  $K \gg N$  without concerns about undersampling in experiments of reasonable length. Surprisingly, one can make progress by choosing weights at random (Maoz *et al.*, 2020).

Figure 15 shows the behavior of models with random weights or projections  $\{W_{\mu i}\}$  as applied to a population of  $N = 178$  neurons in the visual cortex of a macaque monkey as the animal was shown relatively simple images. Experiments were long enough that one could construct, reliably, the pairwise and  $K$ -pairwise maximum entropy models, with more than 15 000 parameters. The performance of these models can be measured, as usual, by estimating the mean log-likelihood of the data in the model, and we see that the models that match correlations generated the data with  $\sim 100$  times higher probability than a model of independent neurons. More than half of this gain is achieved in the random

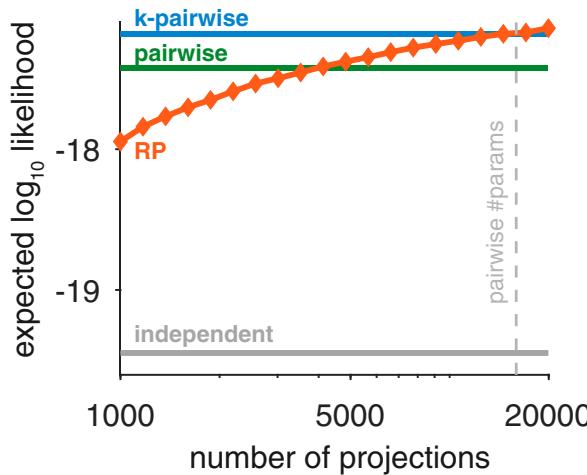


FIG. 15. The surprising success of random projections in describing a population of  $N = 178$  neurons in the visual cortex. RPs are the models defined in Eqs. (70) and (71); independent ( $P_1$ ), pairwise ( $P_2$ ), and  $K$ -pairwise models ( $P_{2k}$ ) are as described in Secs. IV.B and IV.C. In each case models are learned from random halves of the data and likelihoods are computed from the held out data; averages are shown over these random splits. For the RP models, there is also an average over many random choices for projections  $\{W_{\mu i}\}$  in Eq. (70); variations are small. The dashed line marks the point where the complexity of the RP models matches that of the pairwise models. From Maoz *et al.*, 2020.

projection models with only  $K = 1000$  projections, and the full performance is recovered if we allow  $K \sim 15\,000$  projections. This is not in itself an improvement on the  $K$ -pairwise models, but here the projections have been chosen at random.

By iteratively pruning the variables  $y_\mu$  that make weak contributions to the performance of the model and replacing them with new random projections, one arrives at models with the same performance but 10 times fewer parameters. One can make random choices from distributions in which different numbers of  $W_{\mu i}$  are allowed to be nonzero, and it is suggestive that performance is best when the “in degree” of the connections  $\{\sigma_i\} \rightarrow y_\mu$  is small, less than 10. These results indicate that relatively simple maximum entropy models that match a set of strongly nonlinear functions of the network state can be highly effective, although more work will be needed to understand their scaling with  $N$ . It is especially attractive that these functions can be interpreted as the activity of downstream neurons.

The strategy of choosing random projections and then editing these choices is surprisingly successful. This leaves the question, however, as to whether there is a best choice of constraints given a set of possibilities. Intuitively, we measure the quality of a model by the probability that it generates the data. More formally, a model for the distribution  $P(\sigma)$  defines a code in which each state is  $\sigma$  assigned a code word with length

$$\ell(\sigma) = -\log_2 P(\sigma) \text{ bits}, \quad (72)$$

and hence the average amount of space needed to describe the data is (Shannon, 1948; Cover and Thomas, 1991)

$$\langle \ell \rangle = -\langle \log_2 P(\sigma) \rangle_{\text{expt}} \text{ bits}. \quad (73)$$

But maximum entropy distributions are special because, when one substitutes in from Eqs. (20) and (21),

$$-\langle \ln P(\sigma) \rangle_{\text{expt}} = \ln Z + \sum_{\mu=1}^K \lambda_\mu \langle f_\mu(\sigma) \rangle_{\text{expt}} \quad (74)$$

$$= \ln Z + \sum_{\mu=1}^K \lambda_\mu \langle f_\mu(\sigma) \rangle_P \quad (75)$$

$$= -\langle \ln P(\sigma) \rangle_P = S[P]. \quad (76)$$

Thus—for maximum entropy models—the space into which we compress the real data is equal to the entropy of the distribution that we construct. This means that we will achieve the greatest compression by choosing constraints that minimize the entropy of the corresponding maximum entropy model, so we arrive at a “minimax entropy” principle (Zhu, Wu, and Mumford, 1997).

The minimax entropy principle is compelling, but intractable in general. If we want to constrain a subset of the pairwise correlations, then the problem simplifies significantly if we insist that the pairs we constrain form a tree with no loops. On a tree the forward statistical mechanics problem is exactly solvable, the entropy reduction can be written as a sum over the pairwise mutual informations, and there is a greedy

algorithm to find the pairs that maximize the sum. The result is that we can construct the optimal treelike model with minimal computational effort (Lynn, Yu, Pang, Bialek, and Palmer, 2025). The surprise is that in at least some cases, the optimal tree model captures some, though not all, features in the collective behavior of more than 1000 neurons (Lynn, Yu, Pang, Palmer, and Bialek, 2025). While such restricted models may not achieve the full accuracy that we hope for, they may provide a literal backbone for constructing more precise models. It remains to be seen if there are other limits in which the minimax entropy principle becomes tractable.

With many alternative approaches to simplification, we can ask for more efficient methods of exploration. This may be a place where modern machine learning can be helpful—using artificial networks not as models of real networks but rather to solve a wider range of inverse problems. We can take inspiration from recent work on fluid mechanics and weather prediction (Kochkov *et al.*, 2024), including inverse and control problems (Alhashim, Hausknecht, and Brenner, 2025). These results suggest that we may be able to learn the mapping from correlations to effective interactions, not only in the full pairwise model but also in its simplified and extended versions.

Finally, when should we expect simplified models to be possible at all? Again, the distribution  $P(\sigma)$  is a list of  $2^N$  positive numbers, constrained only adding up to 1. In principle, these numbers could be arbitrary, as in the random energy model (Derrida, 1981). But a single variable  $\sigma_i$  can share only a limited amount of information with the rest of the network  $\{\sigma_{j \neq i}\}$ . Since variables are binary, knowing the exact state of the rest of the network can provide at most 1 bit of information about  $\sigma_i$ , although “knowing the exact state of the rest of the network” involves specifying  $O(N - 1)$  bits. We can simplify our models if this knowledge can be compressed without losing information (Bialek, Palmer, and Schwab, 2020).

In an Ising model where variables live on a regular lattice and interact with their neighbors, the influence of the entire network on a single variable can be summarized by knowing the state of the neighboring variables, or  $\sim z$  bits, where  $z$  is the coordination number of the lattice. As long as interactions are short range, this number of relevant bits stays fixed even as  $N \rightarrow \infty$ . In models with long-range interactions, including mean-field models, the averaging over many variables reduces the variance of the effective field acting on a single spin, and this again allows for compression in our description of the interactions. Compressibility in the influence of the entire network on one variable is related to the subextensive behavior of mutual information between halves of the system, and this is violated in the random energy model and in cryptographic systems (Ngampruetikorn and Schwab, 2023). The information shared among neurons is compressible, even in cases where simple pairwise models fail (Ramirez *et al.*, 2024). While compressibility seems to be a requirement for simplification, it is not clear how to use compression to construct explicit simplified models.

## V. A UNIQUE TEST

As we started to see successful maximum entropy analyses of  $\sim 100$  neurons (see Sec. IV.C), the experimental frontier

moved to recording more than 1000 neurons from a single region of the brain. Among other things (see Sec. VII), these larger populations offer the chance to construct many different groups of  $N = 100$  cells from the same region of the brain, and to ask how the success or failure of models varies across these groups of neurons. We can do more and choose groups of cells such that the distribution of mean activities and pairwise correlations are essentially the same—different populations that look the same in the low-order statistics that are the inputs to the maximum entropy construction. Is the success of maximum entropy somehow guaranteed by the form of these low-order data? We later see that this is not the case and that success therefore points to the underlying structure in the network. More deeply, we see that when these models succeed, the match between theory and experiment is surprisingly precise, which may provide a more general lesson about the opportunities for theory in the physics of complex biological systems.

### A. Many groups of $N=100$ neurons

The initial application of this approach was in the mouse hippocampus, specifically the CA1 region, where cells are largely in a single plane (Meshulam *et al.*, 2021). A typical field of view is shown in Fig. 16(a). Scanning two-photon microscopy covers 500  $\mu\text{m}$  at a frame rate of 30 Hz to monitor the calcium-modulated fluorescence of more than 1000 cells as the animal runs repeatedly along a 4 m virtual track, as detailed in Fig. 16(b). These experiments have been done on multiple animals, in each case collecting  $\sim 30$  min of data. Roughly half of the cells that one sees in these recordings are place cells that are consistently active only when the mouse is in a small region of the track.

The raw data in these experiments are movies, as discussed in Sec. III.C. There is a conventional pipeline to associate groups of pixels with individual cells such that we have time series of fluorescence in response to electrical activity as in Fig. 16(c). We go one step further and reduce the signal from each cell  $i$  to a binary variable  $\sigma_i$ . The simplest approach is to set a threshold and, because baselines are stable and noise levels are low, this is unambiguous. We can do a little better, however, since if we see a pulse of fluorescence that falls from its peak and then recovers, we can use our understanding of the dynamics of calcium unbinding from the indicator molecule to identify a flicker between on and off states. Two examples of the binarization process can be seen in Figs. 16(d) and 16(e). A detailed description was given by Meshulam *et al.* (2017). As explained in Sec. IV.B, we can compute from these binary variables the mean activity<sup>15</sup>

$$m_i = \langle \sigma_i \rangle, \quad (77)$$

the covariance matrix

$$C_{ij} = \langle (\sigma_i - m_i)(\sigma_j - m_j) \rangle, \quad (78)$$

<sup>15</sup>For clarity, we repeat in this section some of the definitions provided previously.

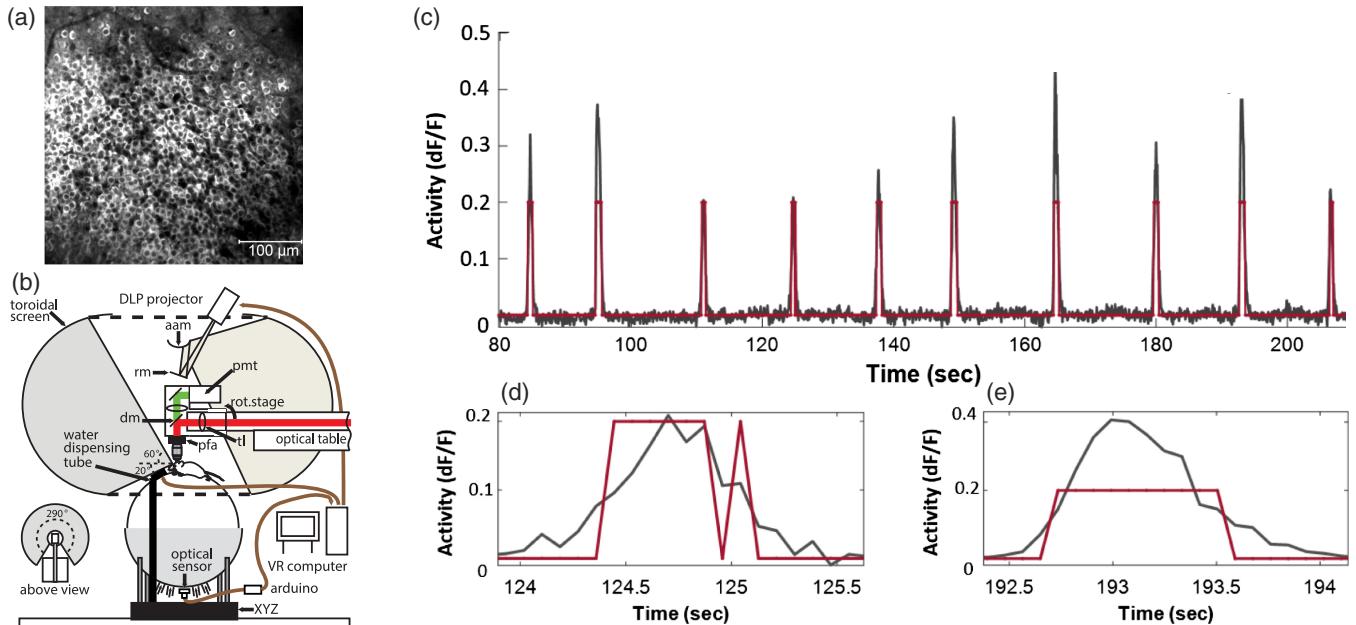


FIG. 16. Imaging electrical activity in the mouse hippocampus. (a)  $500 \times 500 \mu\text{m}^2$  region in the hippocampal area CA1. The image is constructed in  $1/30$  s frames using a scanning two-photon microscope. Here the fluorescence intensity is integrated over time such that each identifiable neuron is bright. (b) Virtual reality setup where the mouse's head is fixed while it runs on a ball. The rotation of the ball is used to compute the effective trajectory through space, driving a movie appropriate for these movements. (c) Continuous fluorescence signal from a single neuron (black line), emphasizing the high signal-to-noise ratio and the ease of defining a binary on/off-on-off version of the cell's activity (red line). (d) On a finer timescale, we understand enough about the dynamics of the indicator molecule to identify slow decay of a fluorescence transient as an on-off flickering of the underlying activity. (e) A simpler case where the cell is "on" when the fluorescence signal is above threshold. (a),(b) From Meshulam *et al.*, 2019. (c)–(e) From Meshulam *et al.*, 2017. Also courtesy of C. D. Brody, J. L. Gauthier, and D. W. Tank.

and the correlation matrix

$$\tilde{C}_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}}. \quad (79)$$

If we point randomly to one cell in the experiment and draw a circle of radius  $r = 0.07$  mm, then we have a dense sampling of  $N = 100$  cells, as shown in Fig. 17(a). If we increase the size of the circle until we enclose roughly twice as many cells, we could choose randomly and create a new population of  $N = 100$  cells. But this network would be noticeably different. In particular, the distribution of correlations between pairs of neurons,  $\tilde{C}_{ij}$ , would be different because there is a tendency for cells that are closer together to be more strongly correlated.

Rather than choosing completely at random, we can swap cells from the initial dense sampling with cells in the larger area, and for each swap we check the distribution of  $\tilde{C}_{ij}$  in the new population. Formally, in the  $k$ th circle, we have some distribution  $P_k(\tilde{C})$ , and in the  $(k+1)$ st circle, after each swap we have a new distribution  $P_{k+1}(\tilde{C})$ . We test the similarity of these distributions by estimating their Kullback-Leibler divergence,

$$D_{\text{KL}}[P_k(\tilde{C})||P_{k+1}(\tilde{C})] = \int d\tilde{C} P_k(\tilde{C}) \log_2 \left[ \frac{P_k(\tilde{C})}{P_{k+1}(\tilde{C})} \right]. \quad (80)$$

As we make successive swaps, we ensure that  $D_{\text{KL}}$  remains small until we have swapped half of the cells, at which point we enlarge the circle yet again. The result, as shown in Fig. 17, is a set of five subgroups of  $N = 100$  cells chosen from increasingly large areas and hence lower sampling densities, but with distributions of correlations that are almost all indistinguishable.<sup>16</sup>

From the formal perspective, we hold the distribution of correlations fixed as we look at different subgroups so that we are solving essentially similar problems. From the functional perspective, holding this distribution fixed also insures that a nearly constant fraction of the cells in each subgroup are place cells.

## B. Maximum entropy models for subgroups

Starting the construction of Fig. 17 from ten randomly chosen cells in each of three animals, we have 150 distinct examples of  $N = 100$  cell subgroups, all with very similar low-order statistics. For each of these subgroups, we can construct the maximum entropy model that matches the mean activity of each cell and the matrix of pairwise

<sup>16</sup>In Figs. 17–23, we compare populations of  $N = 100$  cells sampled at varying densities. In each case the results at highest (leftmost) and lowest (rightmost) densities were given by Meshulam *et al.* (2021).

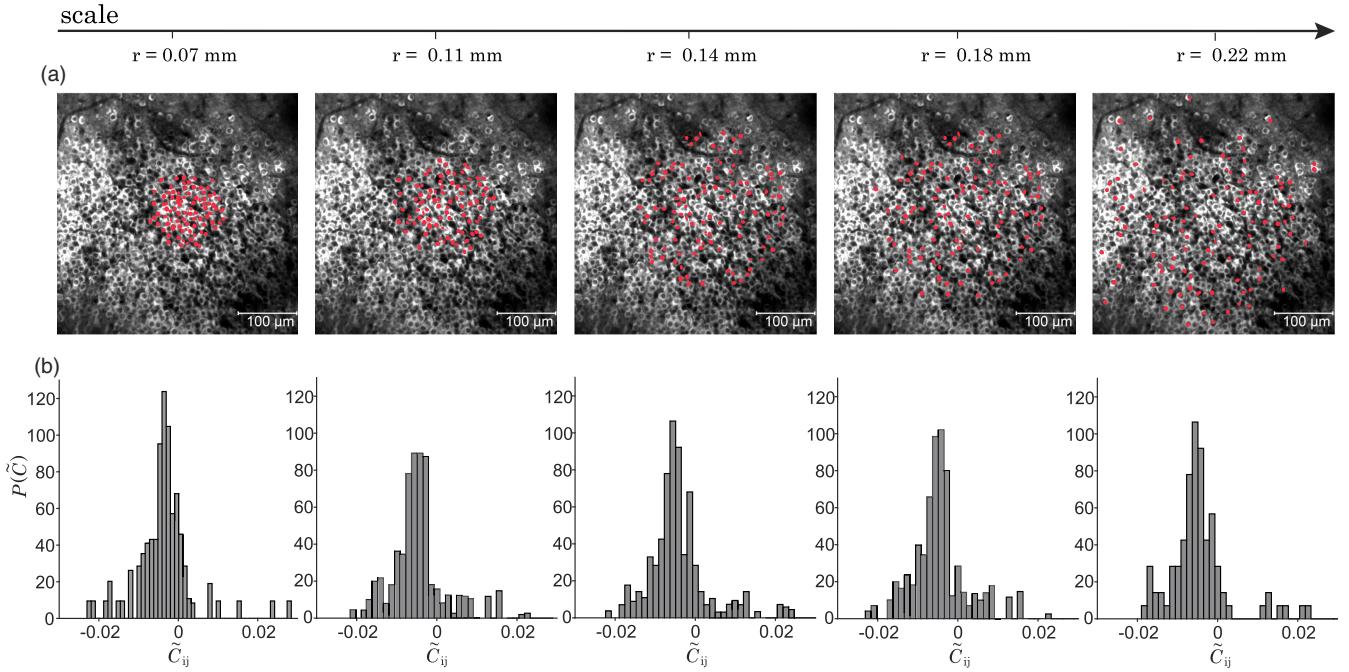


FIG. 17. Subgroups of  $N = 100$  cells with different sampling densities. (a) Image of the CA1 region in a mouse hippocampus showing fluorescence signals from more than 1000 neurons, as in Fig. 16(a). The red dots indicate the cells chosen, as described in the text, from five circles of increasing radius (top row). (b) Distribution of correlation coefficients from Eq. (79) across each population.

correlations. As a reminder, from Eqs. (33) and (35), the result is a model of the form

$$P_2(\sigma) = \frac{1}{Z_2(\{h_i; J_{ij}\})} e^{-E_2(\sigma)}, \quad (81)$$

$$E_2(\sigma) = \sum_{i=1}^N h_i \sigma_i + \frac{1}{2} \sum_{i \neq j} J_{ij} \sigma_i \sigma_j. \quad (82)$$

Note that this is the original form of the model discussed in Secs. IV.B and IV.C without the additional constraint added in Eq. (45) to give a better description of the retina. Methods for choosing the parameters  $\{h_i; J_{ij}\}$  to match the experimentally measured expectation values  $\{m_i; C_{ij}\}$  are summarized in Appendix B.

Drawing from the previous discussion, we can subject the predictions of these models to multiple tests.

- The probability that  $K$  out of  $N$  neurons are active simultaneously,  $P_N(K)$  from Eq. (41).
- The distribution of the effective energy,  $E = E_2(\sigma)$  from Eq. (82).
- The correlations among triplets of neurons,  $C_{ijk}$  from Eq. (47).
- The fine-grained structure of triplet correlations, comparing the model's prediction errors with the experimental errors in estimating these correlations.
- The probability that a single neuron  $i$  is active given the state of the rest of the network, as summarized by the effective field,  $h_i^{\text{eff}}(\{\sigma_{j \neq i}\})$  from Eqs. (49) and (50).
- The distribution of effective fields given the state of a single neuron.

We emphasize that each of these tests does not look at just a single number. Thus,  $N = 100$  cells have  $\sim 1.6 \times 10^5$  distinct triplet correlations, and the distribution of energies has a perhaps surprisingly rich structure.

We also note once more that there is no room for fitting in any of the tests. All of the parameters of our description are determined by matching the means and pairwise correlations, so everything else is a parameter-free prediction. The maximum entropy construction is the hypothesis that all signatures of collective activity in the network can be found in the low-order statistics and thus can be viewed as providing a set of predicted relations between these aspects of the data and the higher-order statistics.

### C. Success depends on the sampling density

Following the agenda outlined in Sec. V.B, we test the predictions of maximum entropy models against six distinct features of the data. We do this in populations of  $N = 100$  cells drawn from regions of different sizes (see Fig. 17), so we can see how the quality of predictions depends on sampling density. We see that there is a systematic decay in the quality of predictions as the density goes down and that some features of the data are easier to get right than others. Here we focus on describing the results from one example shown in Figs. 18–23. Additional examples from more animals are shown in Fig. 24. We summarize the results across all examples and provide perspective in Sec. V.D.

*Distribution of summed activity.* Starting with the first applications of maximum entropy ideas to neurons, it has been appreciated that an important signature of collective behavior is the probability  $P_N(K)$  that  $K$  out of the  $N$  neurons in the network are active in the same small time bins.

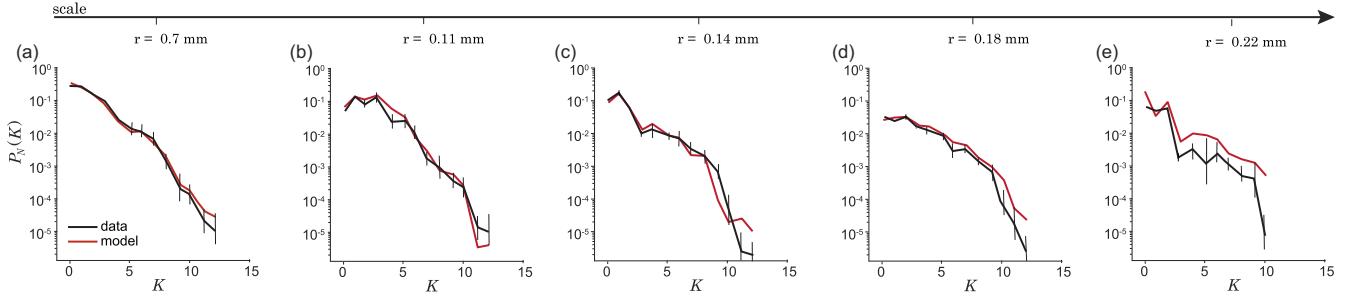


FIG. 18. Distribution of summed activity in  $N = 100$  hippocampal neurons sampled at different densities. (a) The probability  $P_N(K)$  that  $K$  out of the  $N = 100$  neurons in the population are simultaneously active for cells chosen from the smallest selection radius,  $r = 0.07$  mm, at the far left in Fig. 17(a). The model predictions (red line) are compared with the data (black line); the error bars are standard deviations across random halves of the experiment. (b)–(e) As in (a) but for populations chosen from larger areas, with  $r = 0.11, 0.14, 0.18$ , and  $0.22$  mm (top row), moving from the second column toward the right in Figs. 17.

Figure 18 shows the  $P_N(K)$  for the five groups of  $N = 100$  cells shown in Fig. 17. We see that the most spatially contiguous group has the best quantitative agreement with the data, even down to small probabilities, e.g.,  $P_N(K = 12) \sim 10^{-5}$ ; see Fig. 18(a). The observed  $P_N(K)$  changes as we sample cells less densely from larger areas, but the corresponding maximum entropy models predict these changes reasonably well out to a sampling radius  $r = 0.18$  mm; see Figs. 18(b)–18(d). Finally, when we sample from the largest area, predictions fail completely, with disagreements larger than experimental errors already present at  $K = 3$ ; see Fig. 18(e).

*Distribution of effective energy, or surprise.* Maximum entropy models predict the probability of every pattern of activity and silence in the network or, equivalently, how surprised we should be by each of these microscopic states. The negative logarithm of this probability defines an effective energy, and we can compare the distribution of this energy across the states that occur in the data with the distribution predicted by the model, as in Fig. 19. Overall, model predictions are in excellent quantitative agreement with experimental observations for the subgroups selected from the two smallest radii; see Figs. 19(a) and 19(b). We start seeing disagreements at  $r = 0.14$  mm, but even then only in the tails of rare events  $E > 24$ ; see Fig. 19(c). For the largest

two radii, disagreements are more notable; see Figs. 19(d) and 19(e). For networks built by sparse sampling from the largest radii, significant prediction errors are already visible at  $E \sim 7$ ; Fig. 19(e). We see that the experimental distribution  $P(E)$  has fine-scale features that might be mistaken for noise but are not; they are reproduced using the maximum entropy model for the most densely sampled networks.

*Trends in triplet correlations.* The maximum entropy models that we consider here match the two-neuron correlations in the network, so a natural test is to ask about three-neuron or triplet correlations, as in Eq. (47),

$$C_{ijk} \equiv \langle (\sigma_i - m_i)(\sigma_j - m_j)(\sigma_k - m_k) \rangle. \quad (83)$$

For  $N = 100$  cells, there are  $\sim 1.6 \times 10^5$  distinct ways to choose a triplet. In Fig. 20 we group the observed triplet correlations into bins and show the mean and standard deviation of predicted correlations in each bin; perfect predictions would fall on a line of unit slope. For the three subgroups selected from the most compact regions [see Figs. 20(a)–20(c)], and hence with the most dense sampling, predictions are close to the line across the full dynamic range of the data. For  $r = 0.18$  mm the model begins to underestimate the larger, less common correlations,

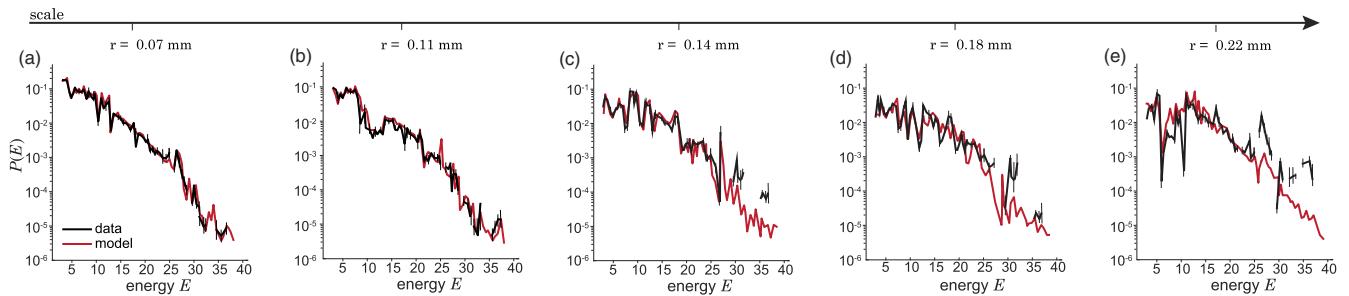


FIG. 19. Distribution of effective energy, or surprise, in  $N = 100$  hippocampal neurons sampled at different densities. (a) The distribution  $P(E)$  of effective energies or log probabilities that the model assigns to every possible state in the network, for cells in the smallest selection radius,  $r = 0.07$  mm, on the far left in Fig. 17. The distribution over states predicted by the model (red line) is compared with the distribution over states as they occur in the experiment (black line); both are computed with a bin size  $\Delta E = 0.75$ . The error bars are standard deviations across random halves of the experiment. (b)–(e) As in (a) but for populations chosen from larger areas, with  $r = 0.11, 0.14, 0.18$ , and  $0.22$  mm (top row), moving from the second column to the right in Fig. 17.

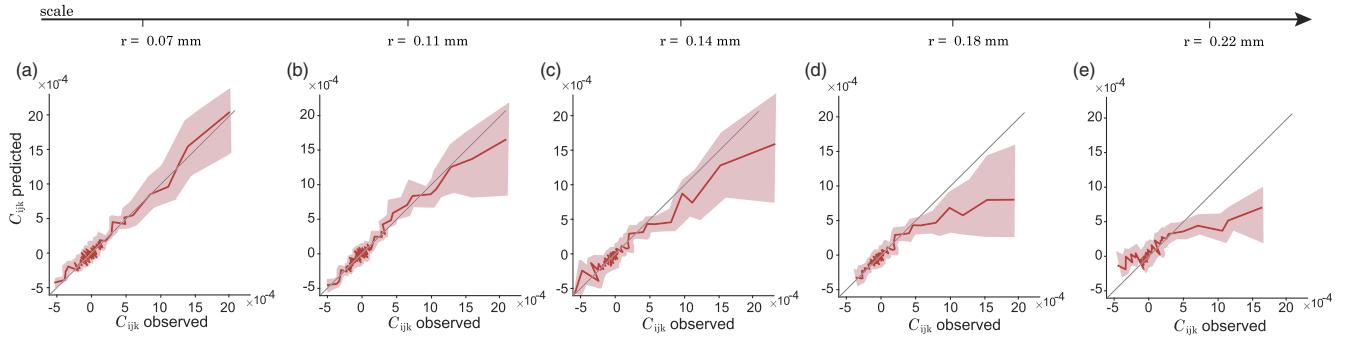


FIG. 20. Trends in triplet correlations in  $N = 100$  hippocampal neurons sampled at different densities. (a) Predicted vs observed triplet correlations  $C_{ijk}$  for cells in the smallest selection radius,  $r = 0.07$  mm, shown at the far left in Fig. 17. On the  $x$  axis, the  $\sim 1.6 \times 10^5$  distinct triplet correlations are grouped together into 100 adaptive bins. The corresponding values computed from the model, which are binned in the same way, are shown on the  $y$  axis. The shaded area indicates the standard deviation of predictions within each bin. (b)–(e) As in (a) but for populations chosen from larger areas, with  $r = 0.11, 0.14, 0.18$ , and  $0.22$  mm (top row), moving from the second column to the right in Fig. 17.

$|C_{ijk}| = \gtrsim 3 \times 10^{-4}$ ; see Fig. 20(d). Finally, with neurons chosen sparsely from the largest area, there is limited success with the smallest  $|C_{ijk}|$  and systematic underestimates of the absolute correlations over most of the dynamic range; see Fig. 20(e).

*Triplet correlations in detail.* Figure 20 tests the ability of the maximum entropy models to capture the trends in triplet correlations, but does not tell us whether the individual elements of the correlation tensor  $C_{ijk}$  are correct in detail. To get at this, we compare the errors in the model’s predictions with the errors in measurement. Once again, we collect the observed correlations into small bins, and within each bin we compute the root-mean-square error in the model predictions and estimate the root-mean-square errors in measurement of the correlation itself from the data. We focus, in particular, on the bulk of the triplets with  $|C_{ijk}| < 4 \times 10^{-4}$ . Figure 21 compares these prediction and measurement errors across groups of  $N = 100$  cells drawn from increasingly large areas. We see that the two measures of error are essentially identical in the smallest, most densely sampled network. Without overfitting, it is difficult to perform any better than this; see Fig. 21(a). As we sample cells from larger regions, the two error measures gradually separate [see Figs. 21(b)–21(d)] until

the prediction errors are consistently larger than the experimental errors across the full range of correlations that we probe here [see Fig. 21(e)].

*Collective behavior and effective fields.* One of the characteristics of a population whose behavior is collective is that the activity in the network as a whole can be strongly predictive of individual member’s activity. Using the equivalence between our maximum entropy model and an Ising model with competing interactions, this predictive power is summarized by an effective field  $h_i^{\text{eff}}$  acting on each neuron, as in Eq. (50). In the model used here [Eq. (82)], this becomes

$$\begin{aligned} h_i^{\text{eff}} &= E(\sigma_1, \dots, \sigma_i = 0, \dots, \sigma_N) - E(\sigma_1, \dots, \sigma_i = 1, \dots, \sigma_N) \\ &= h_i + \sum_{j \neq i} J_{ij} \sigma_j. \end{aligned} \quad (84)$$

The effective field predicts the probability for any single neuron to be active at a single moment in time, given the active or silent state of all the other neurons in the population at the same time point; from Eq. (49),

$$P(\sigma_i = 1 | h_i^{\text{eff}}) = \frac{1}{1 + \exp(-h_i^{\text{eff}})}. \quad (85)$$

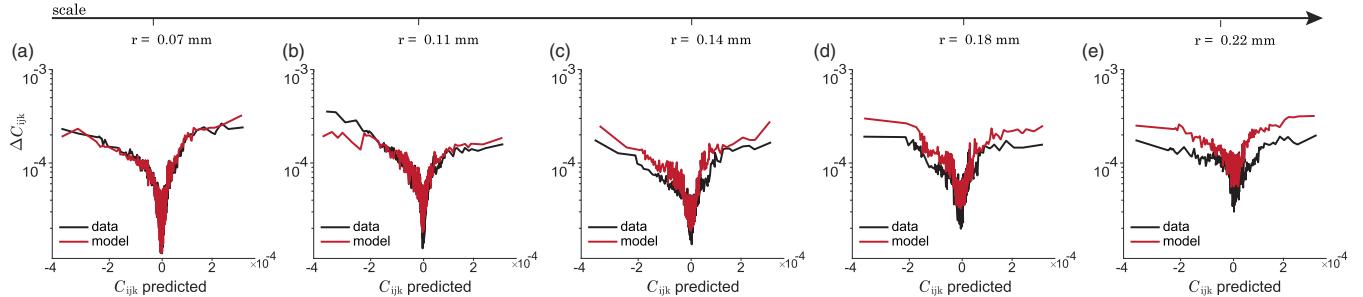


FIG. 21. Detailed triplet correlations for  $N = 100$  hippocampal neurons sampled at different densities. (a) Comparison of the maximum entropy model prediction errors (red) for individual triplet correlations,  $C_{ijk}$  in Eq. (83), with the measurement errors (black line) from the data itself. For cells in the smallest selection radius,  $r = 0.07$  mm, as shown at the far left in Fig. 17. Values on the  $x$  axis are grouped together into 500 adaptive bins. (b)–(e) As in (a) but for populations chosen from larger areas, with  $r = 0.11, 0.14, 0.18$ , and  $0.22$  mm (top row), moving from the second column to the right in Fig. 17.

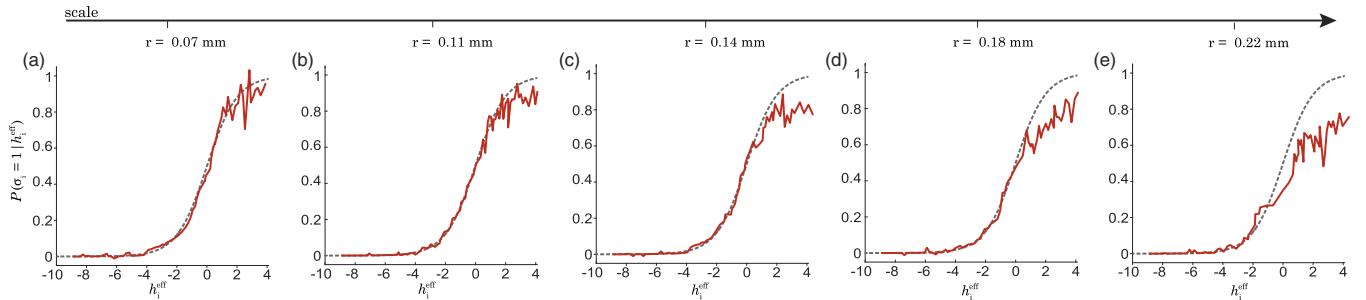


FIG. 22. Collective behavior and effective fields in  $N = 100$  hippocampal neurons sampled at different densities. (a) Probability of individual neurons to be active given the state of the network, which is summarized by the effective field from Eq. (49). For cells in the smallest selection radius,  $r = 0.07$  mm, as shown at the far left in Fig. 17. Data in red, prediction of Eq. (50) in the dashed black line. (b)–(e) As in (a) but for populations chosen from larger areas, with  $r = 0.11, 0.14, 0.18$ , and  $0.22$  mm (top row), moving from the second column to the right in Fig. 17.

In Fig. 22 we examine the quality of these predictions as a function of sampling density, as with previous tests. For each cell  $i$  and for every moment in time, we can compute the effective field from the state of all the other neurons, and we can then estimate the probability that the cell is active given that the field falls into a small bin. We see that the agreement between theory and experiment is good for a network built from the most dense sampling [see Fig. 22(a)], even at the extremes of the effective field. The quality of predictions falls gradually as we sample with lower density from larger areas; see Figs. 22(b)–22(e). In particular, with dense sampling there are moments when the effective field is large enough that we predict a cell to be active with near certainty, and these predictions are correct. This strong (if rare) prediction fails as we look at lower-density populations, and this error spreads to lower and lower probabilities until the models even fail at negative fields.

*Inferring the effective field.* If the effective field acting on a neuron is large and positive, our models predict that the neuron should be active; quantitatively, the model predicts the probability of activity as in Fig. 22. Can we turn this around and use the activity or silence of one cell to predict the state of the rest of the network, as summarized by the effective field? These questions are related through Bayes's rule,

$$P(h_i^{\text{eff}} | \sigma_i = 1) = \frac{1}{P(\sigma_i = 1)} P(\sigma_i = 1 | h_i^{\text{eff}}) P(h_i^{\text{eff}}), \quad (86)$$

and similarly for  $\sigma_i = 0$ . Because the distribution of effective fields has a nontrivial form, it is not easy to guess how these distributions will look. In particular, we want to see that active neurons point to a state of the network that generates large positive fields, and conversely for inactive neurons, such that  $P(h_i^{\text{eff}} | \sigma_i = 1)$  and  $P(h_i^{\text{eff}} | \sigma_i = 0)$  are distinguishable. We test this distinguishability in Fig. 23, expressing the effective field as the predicted probability of activity through Eq. (85). We see that when we build networks via dense sampling from a small region, the two distributions are almost nonoverlapping [see Fig. 23(a)] such that the activity or silence of a single cell is maximally informative about the state of the network as a whole. Overlap is visible as soon as we sample from  $r = 0.11$  mm [see Fig. 23(b)] and continues to grow [see Figs. 23(c) and 23(d)] until, at the sparsest sampling from the largest radius, the two distributions overlap almost completely [see Fig. 23(e)]. Note that as the quality of predictions falls off with increasing sampling radius, the distribution of fields conditional on an active neuron (purple) moves toward the distribution conditional on a silent neuron (yellow), rather than both changing toward each other. This is consistent with the errors in Fig. 22 starting at large positive field and spreading toward lower values. Taken together, these results show that as the radius increases, the model predicts more false negatives, i.e., misses predicting the activity of a neuron that was active. More precisely, the model fails to connect

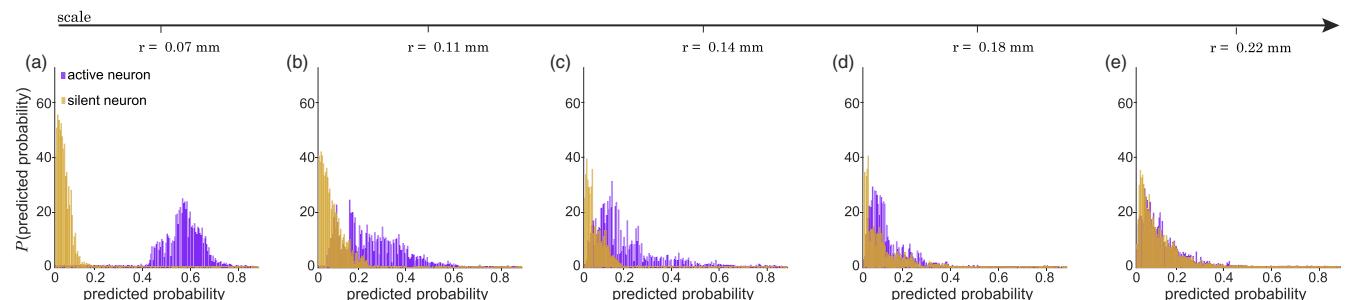


FIG. 23. Inferred effective fields in  $N = 100$  hippocampal neurons sampled at different densities. (a) Distribution of effective fields  $h_{\text{eff}}$  given that a neuron is active (purple) or silent (gold), with the effective field measured using the predicted probability of activity from Eq. (85). The cells shown are in the smallest selection radius,  $r = 0.07$  mm, as shown at the far left in Fig. 17. (b)–(e) As in (a) but for populations chosen from larger areas, with  $r = 0.11, 0.14, 0.18$ , and  $0.22$  mm (top row), moving from the second column to the right in Fig. 17.

active neurons with the associated states of the network. It also suggests that false negatives are more difficult to avoid than false positives.

#### D. Precision matters

The first thing that we notice is that theory and experiment really can agree exceedingly well. We see this especially in the (a) panels of Figs. 18–23 that refer to sampling  $N = 100$  neurons from the smallest radius, and in the left columns of the examples from two additional mice in Fig. 24. It is particularly striking that maximum entropy models can

reproduce bumps and wiggles in the energy distribution that one might have dismissed as noise, though this would be inconsistent with the measured error bars, and that the bulk of the  $\sim 10^5$  individual triplet correlations are reproduced within experimental error. We saw similar results in the first such analysis of the hippocampus (Meshulam *et al.*, 2017), but it is reassuring to see that this detailed quantitative success is reproducible across different populations of neurons in independent experiments on multiple animals. While we expect this sort of reproducibility in physics, we should not take it for granted in the complex context of a functioning brain.

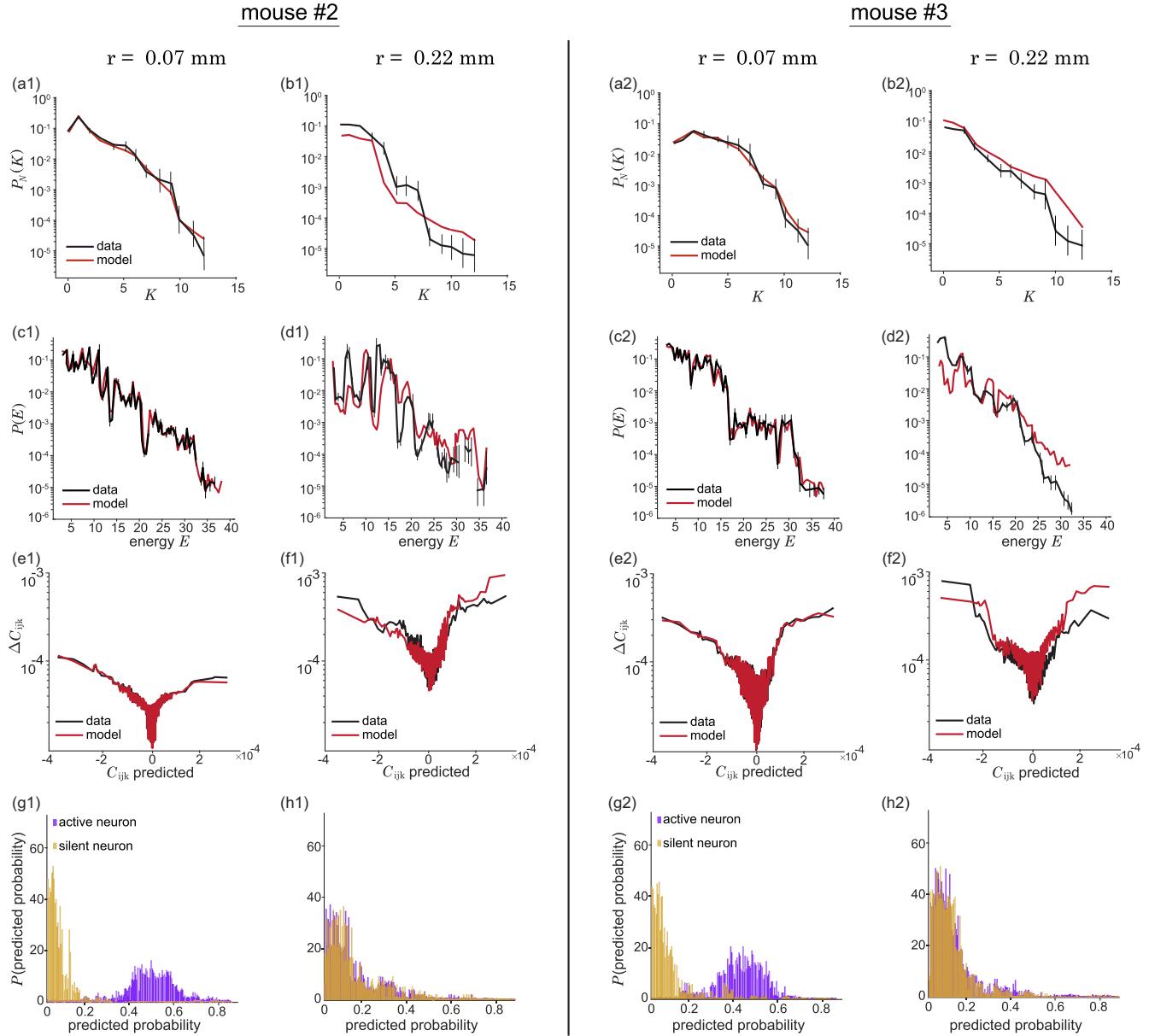


FIG. 24. Model predictions examples for two more mice. Predictions shown for the subgroups constructed from the two extreme radii, with the same starting point. (a)(1)–(h)(1) Predictions from experiments in mouse 2. (a)(2)–(h)(2) Predictions from mouse 3. (a),(c),(e), (g) Predictions for the subgroup sampled from the smallest radius (neighboring cells). (b),(d),(f),(h) Predictions for the subgroup sampled from the largest radius (the entire field of view). (a),(b) Distributions of summed activity as in Fig. 18 for mouse 1. (c),(d) Distributions of effective energy as in Fig. 19. (e),(f) Detailed triplet correlations as in Fig. 21. (g),(h) Inferences of the effective field from the activity of a single neuron as in Fig. 23.

It is equally important to note that success is not automatic. If we look only at  $N = 100$  cells from the smallest radius, where we have essentially complete sampling of a local network, everything works and it is difficult to assess the significance of this result. It could be that the models are so expressive that they can explain anything. It could also be that, while we see the multiple tests of the model as being different, the model or the real network ties these different quantities together so strongly that all succeed or fail together. Looking at networks built by sampling at lower density from larger regions, we see that neither of these ideas is correct. Different networks can be more or less well described by pairwise maximum entropy models even though they have similar low-order statistics,<sup>17</sup> and different features of the data can be captured or not by these models, suggesting that there is a hierarchy of difficulty to the different experimental tests. More deeply, success or failure of the model must be telling us something about the underlying network beyond what we see in the pairwise correlations.

It is useful to look at the performance of the model at a middle level of sampling density ( $r = 0.14$  mm), corresponding to the (c) panels in Figs. 18–23. We see that the predictions for the probability of  $K$  neurons being active simultaneously [see Fig. 18(c)], for the distribution of effective energies [see Fig. 19(c)], and for the trends in triplet correlations [see Fig. 20(c)] are not bad. If this were the best we had seen, we might think it was a success. But when we look at the triplets in detail [see Fig. 21(c)], the activity of neurons as a function of the effective field [see Fig. 22(c)], and the ability to infer the effective field from the activity of the individual neurons [see Fig. 23(c)], the agreement between theory and experiment is noticeably worse. This trend continues as we build networks by sampling the same number of cells from larger areas.

While the quality of predictions generally goes down at lower sampling densities, these failures happen in a well-defined order. Good predictions of  $P_N(K)$  survive the longest (see Fig. 18), followed by the distribution of effective energies (see Fig. 19) and trends in triplet correlations (see Fig. 19), which are roughly equivalent in performance. The three more challenging properties to predict also follow an internal hierarchy, with the inference of the effective fields being the most difficult, as significant disagreements arise even at  $r = 0.11$  mm; see Fig. 23.

We now identify two particularly interesting examples of model success and failure. To begin, we can have models that capture the trends in triplet correlations [see Fig. 20(c)], while the prediction errors for individual triplet correlations are outside of experimental error [see Fig. 21(c)]. The failure to predict individual triplet correlations is a hint that something more serious is going wrong, and again this gets worse at lower sampling density. Another observation is that a model can give a decent description of how the activity of a single cell depends on the network state through the effective field [see Fig. 22(c)], while it is almost impossible to distinguish the distributions of fields consistent with that cell being active or

silent [see Fig. 23(c)]. This is not so much a disagreement with the data as a breakdown in the interpretability of the model. We want to be able to say that, because behavior is collective, an active cell is responding to a positive field imposed by the rest of the network, but this proves to be the most fragile of predictions.

The ability of these statistical physics models to reproduce all  $N^3/3! \sim 10^5$  triplet correlations within the errors of the measurements gives a sense for the power of this theoretical approach. Not long ago, it seemed sensible to consider alternative models that capture qualitative features of the triplet correlations (Macke, Opper, and Bethge, 2011), but now we see that it is possible for pairwise models to predict all the triplet correlations within errors. Importantly, this success depends on the sampling density.

Even if a large network of neurons is described exactly by a pairwise model, the distribution over states in a subnetwork will be described only approximately by such models. The approximation gets better if the interactions in the entire network are largely within the subnetwork. The success of pairwise models when applied to dense sampling from restricted areas suggests, strongly, that interactions and inputs are spatially local. Although still somewhat controversial, this locality is consistent with more direct measurements taken over many years (Wiener, Paul, and Eichenbaum, 1989; Hampson, Simeral, and Deadwyler, 1999; Rickgauer, Deisseroth, and Tank, 2014).

The hippocampus is especially interesting because of the well-studied place cells that play a role in navigation, as discussed in Sec. IV.C. We have seen how a model in which cells respond independently to the animal's position fails to capture the variability of responses in repeated movements through the same region, while the maximum entropy models predict this behavior as a response of individual neurons to the state of the entire network without reference to position; see Fig. 14. In looking more generally at alternative models (see Sec. VI.D), we see that the independent place cell model also fails to account for the triplet correlations (see Fig. 31). This failure is dramatic, but only because we have seen the detailed quantitative success of the maximum entropy approach. The conclusion is that cells in the hippocampus share information about more than just place, and this information is captured by the couplings in the Ising model. This information can be quantified in bits (Meshulam *et al.*, 2017), and this picture is consistent with models in which place selectivity itself is an emergent property of the network (Treves, Migliolo, and Parisi, 1992).

The success of pairwise maximum entropy models may be more surprising because the different examples of  $N = 100$  neurons really are different, even though the distributions of low-order statistics are similar. We can see this directly from the data by comparing examples of  $P_N(K)$  from three different animals; Figs. 18(a), 24(a)(1), and 24(a)(2). We can make the same point looking at the data through the lens of the models by comparing examples of  $P(E)$ ; see Figs. 19(a), 24(c)(1), and 24(c)(2). If we look at the models themselves, they all are spin-glass-like, with fluctuations in the couplings  $J_{ij}$  from link to link that are comparable to the mean coupling, as shown in Fig. 25. All of the models are in a regime of reasonably strong

<sup>17</sup>In particular, this lays to rest the early speculation that success or failure of these models could be predicted from the mean activity of the neurons alone (Roudi, Nirenberg, and Latham, 2009).

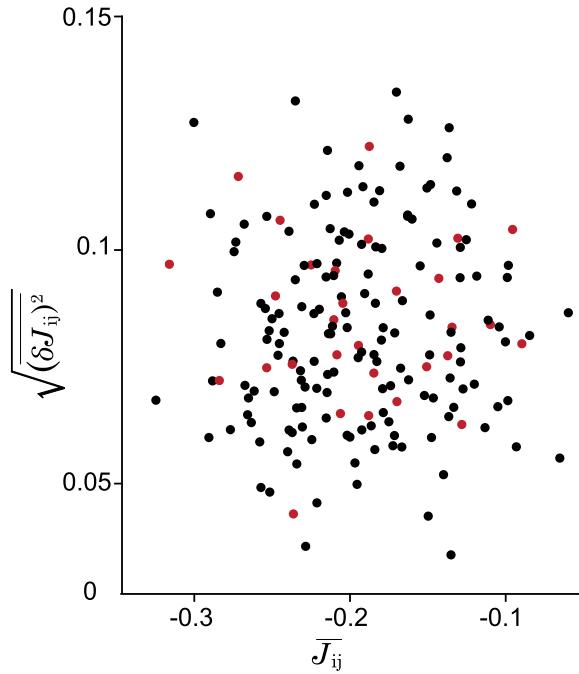


FIG. 25. Mean vs standard deviation of the coupling constants in maximum entropy models constructed for each of the 150 populations of all sampling radii in three animals. The populations associated with the smallest radii are shown in red (ten from each animal).

coupling, with  $N\bar{J} \sim N\overline{(\delta J)^2} \sim 1$ , but all the models are different in the precise form of the matrix<sup>18</sup>  $J_{ij}$ ; these differences from subgroup to subgroup are larger than expected if the matrix elements were drawn independently from a fixed underlying distribution. Importantly, these differences are present even in the populations drawn from the smallest radii, where the models are most successful. On the one hand, these observations indicate that relatively simple statistical physics models of real living systems are succeeding in capturing how particular systems behave in detail. On the other hand, this leaves open the question of whether there is something more universal in this behavior, to which we return in Sec. VII.A.

Finally, there is a more general lesson to be drawn from this larger-scale survey. As the quality of measurements on biological systems improves, we should aspire to the kind of detailed, quantitative theory-experiment comparison that we expect in other areas of physics.

## VI. CRITICALITY

Correlations between two neurons in a network typically are weak but widespread. This is reminiscent of what happens in mean-field models. As an example, for a mean-field

ferromagnet, all the pairwise correlations are equal and  $C \sim 1/N$  (Sethna, 2021; Kivelson, Jiang, and Chang, 2024). If we take this analogy seriously, then correlations in a network with  $N \sim 100$  cells should be  $C \sim 0.01$ , which is in fact smaller than what we see. More seriously, while we can observe a varying number of neurons, the actual size of the network is fixed by the patterns of connectivity, and the “real” values of  $N$  are even larger. The familiar statistical physics models thus make it difficult to understand how the correlations, averaged over all pairs of cells, can reach  $N\bar{C} \gg 1$ . There are two broad possibilities: such large correlations either could be driven by fluctuating external fields or could emerge from tuning of the system close to a critical point.

Living systems are not random combinations of their components, and it is a challenge to define what is special. If the number of interacting components is large, we might expect that behaviors can be organized into a phase diagram. Critical points in the phase diagram are special in many ways: collective coordinates can be infinitely sensitive to variations in external parameters; correlations can extend throughout the system, far beyond the range of direct interactions; and fluctuations and responses can occur over a wide range of timescales, with the longest timescale growing with the size of the system. For these reasons and more, many groups have suggested that biological systems might be tuned, or self-tuned, to a critical point (Bak, 1996; Mora and Bialek, 2011; Muñoz, 2018).

### A. Avalanches and dynamics

The idea of criticality in networks of neurons was given considerable stimulus by the emergence of models for self-organized criticality (Bak, Tang, and Wiesenfeld, 1987; Tang *et al.*, 1987), in which, as the name suggests, dynamical systems can “tune themselves” to criticality rather than requiring precise adjustment of some underlying parameters. The simplest models of self-organized criticality describe a stylized sandpile, with sand dropping randomly onto the surface, and criticality is the statement that the avalanches that collapse the high peaks in the pile occur in all sizes, with a power-law distribution. The first suggestion that criticality might be relevant to the brain was the observation of “neural avalanches” in the activity of neural networks in a dish with an array of electrodes on its bottom surface (Beggs and Plenz, 2003). Activity in these systems consists of long periods of quiet punctuated by bursts, and these bursts are avalanchelike in the sense that the random occurrence of activity in one or a few cells triggers activity in other cells, thereby spreading through the network. Power laws are seen not only in the amplitude of the avalanches but also in their duration and in the mean amplitude as a function of duration. Average trajectories of avalanches with different duration can be rescaled to a universal form (Friedman *et al.*, 2012).

In the earliest experiments, activity was defined by the signal at a single electrode exceeding some threshold, and scaling often was confined to a narrow range. More recent experiments resolved the spikes from single neurons in intact brains (Fontenelle *et al.*, 2019) and observed scaling over three decades (Fontenelle *et al.*, 2024). An example, from neurons in the motor cortex of a behaving mouse, is shown in Fig. 26.

<sup>18</sup>As usual, we write  $\overline{\cdots}$  to denote an average over “disorder” in the parameters of a model, as distinguished from  $\langle \cdots \rangle$ , which denotes an average over variables drawn from the model at fixed parameters. Here the disorder is the variation of couplings across all  $N(N-1)/2 \sim 5000$  distinct pairs in each population of  $N = 100$  neurons.

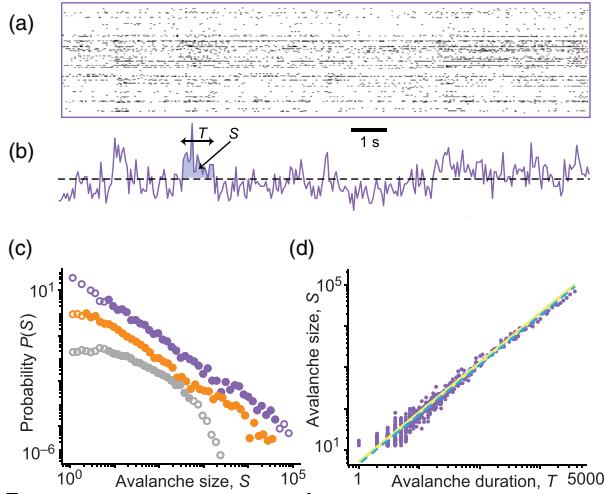


FIG. 26. Avalanches in a population of  $N = 208$  neurons in the motor cortex of a mouse as it runs along a track. (a) Spike rasters from all  $N$  cells, where a dot represents the occurrence of a spike. (b) Total number of spikes in  $\Delta\tau = 50$  ms bins. The dashed line marks a threshold to define an avalanche. The area  $S$  defines the avalanche size and  $T$  indicates the duration. (c) Distribution of avalanche sizes  $P(S)$ , where the threshold has been set to the eighth percentile in the distribution of spike count. Shown are the full data (purple); a projection onto the largest five principal components of the activity (orange), which preserves much though not all of the scaling behavior; a projection onto the remaining  $N - 5$  components (gray). The distributions are shifted vertically for clarity. (d) Avalanche size  $S$  vs duration  $T$  measured in units of  $\Delta\tau$ . Each point is a single avalanche event. Lines are power laws from different model predictions. From [Fontenele et al., 2024](#).

A central feature of criticality is critical slowing down. In a low-dimensional dynamical system, we expect to see one slow mode appear as the system parameters approach a bifurcation, but in a system with many degrees of freedom, we can see a macroscopic density of modes with decay rates approaching zero. In many cases this is understandable as a result of dynamic scaling, as discussed in Sec. VII.A. [Solovey et al. \(2015\)](#) took a more phenomenological approach, analyzing electrocorticographic (ECoG) recordings in primates.

ECoG is done by placing an array of electrodes on the surface of the brain. This is similar to an EEG, which uses an electrode array on the surface of the skull. ECoG cannot resolve individual neurons but offers higher spatial resolution than EEGs. It often is used in neurosurgery to map brain areas in humans. The dynamics of the voltage signals  $\{V_\mu(t)\}$  are nonlinear, but one can make progress in a locally linear approximation. Concretely, the linear approximation is

$$V_\mu(t) = \sum_\nu A_{\mu\nu} V_\nu(t - \Delta\tau) + \epsilon_\mu(t), \quad (87)$$

where  $\mu = 1, 2, \dots, 128$  is an index on the electrodes, the time resolution  $\Delta\tau = 1$  ms, and  $\epsilon_\mu$  is a noise term that we try to minimize by adjusting the dynamical matrix  $A_{\mu\nu}$ . Because we expect linearity to work only locally, the dynamical matrix is

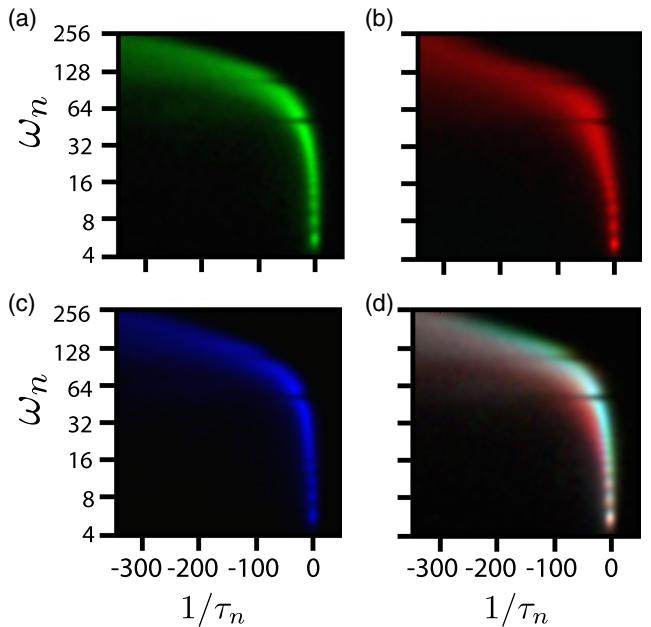


FIG. 27. Spectra of ECoG fluctuations in primate cortex. Frequencies  $\omega_n$  and decay rates  $1/\tau_n$  are extracted from the eigenvalues of the local dynamics matrix, as in Eq. (89), and the density of these points is mapped across a long recording. (a) Results in an awake animal (green). (B) Results under anesthesia (red). (C) Results after recovery from anesthesia (blue). (D) Superposition of the three spectra. The similarity of results before and after anesthesia is shown by the dominance of cyan rather than separate blue and green regions. The faster decays under anesthesia are shown by the leftward displacement of the red density. Adapted from [Solovey et al., 2015](#).

fit to short (500 ms) segments of the data. In each segment we can find the spectrum of the dynamical matrix,

$$\sum_\nu A_{\mu\nu} \phi_\nu^n = \Lambda_n \phi_\mu^n, \quad (88)$$

$$\Lambda_n = e^{-(i\omega_n + 1/\tau_n)\Delta\tau}, \quad (89)$$

which defines a collection of modes with frequencies  $\omega_n$  and relaxation times  $\tau_n$ . Combining data across many segments, we find a density in the  $(\omega, 1/\tau)$  plane, as shown in Fig. 27. We see that there is a substantial concentration of modes with large values of  $\tau$ , almost touching the stability line  $1/\tau = 0$ ; see Fig. 27(a). Perhaps even more notably, the density shifts away from the stability line, toward shorter relaxation times, as the animal is anesthetized [see Fig. 27(b)], and the slow modes then reappear as the animal wakes up [see Fig. 27(c)]. Not only do we see signs of critical slowing down, but they are associated with consciousness as opposed to sleep.

Analyses of ECoG recordings and avalanches share the need for making choices. Avalanches need to be defined, at least by a threshold, time is discretized, and care sometimes needs to be taken in marking the ends of these events. The dynamics of ECoG signals surely are nonlinear, and the locally linear approximation uncovers interesting structure, but one might worry that eigenvalue spectra have a clear meaning only in this approximation. Given that there are

choices to be made, one view is that there is a correct version of these choices, and the other view is that (within reason) these choices should not affect our conclusions. In the case of avalanches, power-law size distributions are visible across many choices of parameters, but from the start it has been noticed that scaling exponents vary (Beggs and Plenz, 2003). Given this sensitivity to choices in the analysis, it is not clear whether we should expect universality of exponents across different networks of neurons. As we emphasize in Sec. VII.A in the examples, we understand that scaling is very precise and, in a sense, becomes clearer the more closely we look. In addition, criticality is more than power laws. Recent work has drawn attention to universality in the temporal form of avalanches (Friedman *et al.*, 2012) and has pushed for more precise tests of scaling over larger dynamic range (Fontenele *et al.*, 2024), but achieving this has been challenging. We hope that improved experimental methods will make it possible to address these issues more fully.

A much simpler notion of dynamical criticality arises in thinking about how the brain integrates signals over time. As an example, we (and other animals) move our eyes to compensate for the rotation of our head. This requires that our eye muscles apply a force related to the rotational displacement; otherwise, the eyes would relax back to their resting position. But we measure the rotation of our head using our vestibular system, and this is an inertial sensor. Viscosity of the fluid inside the semicircular canal converts the acceleration signal into a velocity signal, but it is left to our brains to do one more integral, converting velocity into displacement. This “oculomotor integrator” has been studied for decades in many different organisms. A large class of models for the underlying circuits can be approximated as dynamical systems that have a line attractor, with position along the line corresponding to the position of the eye (Seung, 1996). The linearized dynamics of such a system has a true zero mode, so the network is poised at a bifurcation or critical point between stable and unstable behavior. In practice, the integration is leaky, but on timescales orders of magnitude longer than the relaxation times of individual neurons in the network (Aksay *et al.*, 2001), so these systems must be close to critical. How this relates to the underlying connections among neurons is an active topic of investigation in model organisms ranging from zebra fish to primates and exemplifying current efforts to map synaptic connectivity completely (Joshua and Lisberger, 2015; Vishwanathan *et al.*, 2024).

Finally, we discuss dynamical criticality in relation to learning. Many of these considerations are common to a broader class of dynamical systems, so we consider continuous variables  $\mathbf{x} = \{x_i(t)\}$  that obey general equations of motion,

$$\frac{dx_i}{dt} = F_i(\mathbf{x}, t; \boldsymbol{\theta}), \quad (90)$$

where  $\boldsymbol{\theta} = \{\theta_\alpha\}$  are the adjustable parameters that we imagine can be learned by assessing the performance of the network, for example, following the gradient of some cost function. We assume that this cost can be measured locally in time and that the total cost  $\mathbf{C}$  is an integral over time such that

$$\mathbf{C} = \int dt \mathcal{C}[\mathbf{x}(t), t]. \quad (91)$$

As an example, some of the variables  $\mathbf{x}$  could be motor outputs, with  $\mathcal{C}$  measuring the distance between these outputs and some desired trajectory. The cost depends implicitly on the parameters  $\boldsymbol{\theta}$  through the equations of motion, which makes it difficult to compute how the cost changes when we change parameters.

One strategy to make the dependence on parameters more explicit is to attach extra terms to the local cost  $\mathcal{C}$  that act as Lagrange multipliers to enforce the equations of motion. Rather than finding the minimum of  $\mathbf{C}$  with respect to the parameters, we minimize the action

$$\mathcal{S} = \int dt \mathcal{L}[\mathbf{x}(t), \lambda(t), t; \boldsymbol{\theta}], \quad (92)$$

$$\mathcal{L} = \mathcal{C}[\mathbf{x}(t), t] - \sum_i \lambda_i(t) \left[ \frac{dx_i}{dt} - F_i(\mathbf{x}, t; \boldsymbol{\theta}) \right]. \quad (93)$$

Note that along a trajectory that obeys the equations of motion, we have

$$\frac{d\mathbf{C}}{d\theta_\alpha} = \frac{d\mathcal{S}}{d\theta_\alpha}. \quad (94)$$

The use of Lagrange multipliers as auxiliary dynamical variables goes back to Pontryagin<sup>19</sup> and has found wide applications in control theory. It reappears in the use of field-theoretic methods for classical stochastic dynamics (Martin, Siggia, and Rose, 1973), and its relevance to connecting learning and dynamics in networks of neurons was emphasized by Krishnamurthy, Can, and Schwab (2022).

Taking the derivatives in Eq. (94), we find that

$$\begin{aligned} \frac{d\mathbf{C}}{d\theta_\alpha} &= \sum_i \int dt \left[ \frac{\partial \mathcal{C}}{\partial x_i} \frac{dx_i}{d\theta_\alpha} - \lambda_i(t) \frac{d}{dt} \frac{dx_i}{d\theta_\alpha} \right] \\ &\quad + \sum_i \int dt \lambda_i(t) \left[ \frac{\partial F_i}{\partial \theta_\alpha} + \frac{\partial F_i}{\partial x_j} \frac{dx_j}{d\theta_\alpha} \right]. \end{aligned} \quad (95)$$

Further, if we are careful about the boundary conditions, the extremum with respect to  $\mathbf{x}(t)$  can be written as an equation for the dynamics of the Lagrange multipliers,

$$\frac{\delta \mathcal{S}}{\delta x_i(t)} = 0 \quad (96)$$

$$\Rightarrow \frac{d\lambda_i(t)}{dt} = - \sum_j \frac{\partial F_j}{\partial x_i} \lambda_j(t) - \frac{\partial \mathcal{C}}{\partial x_i}. \quad (97)$$

Substituting this into Eq. (95) and again integrating by parts, we find that all the terms with  $dx_i/d\theta_\alpha$  cancel, leaving

$$\frac{d\mathbf{C}}{d\theta_\alpha} = \sum_i \int dt \lambda_i(t) \frac{\partial F_i}{\partial \theta_\alpha}. \quad (98)$$

---

<sup>19</sup>For an accessible source, see Pontryagin (1987).

We see from Eq. (97) that the auxiliary variables  $\lambda$  (locally) grow or shrink exponentially, and this is determined by the eigenvalues of the matrix  $\partial F_j / \partial x_i$  evaluated along the trajectory. Importantly, this is the transpose of the dynamical matrix that determines, through the equations of motion, whether two nearby trajectories  $x(t)$  and  $x(t) + \delta x(t)$  separate or converge over time. Thus, if the network dynamics is fully stable, with negative Lyapunov exponents, then  $\lambda$  will decay exponentially, and through Eq. (98) the gradient of the cost with respect to parameters also will be exponentially small, making it difficult to learn. Conversely, if the network dynamics are strongly chaotic, with positive Lyapunov exponents, then  $\lambda$  will grow exponentially, and so will the gradient, again making it difficult to learn. The only way to insure that the gradient of the cost function has  $\mathcal{O}(1)$  contributions from all along the trajectory is for the network dynamics to be characterized by Lyapunov exponents near zero—the regime of dynamical criticality. Recurrent networks near criticality may also be more effective because they have access to a wider range of timescales. These observations are broadly in agreement with empirical results (Bertschinger and Natschläger, 2004; Pascanu, Mikolov, and Bengio, 2013; Vorontsov *et al.*, 2017).

## B. An effective thermodynamics

Now that we can construct accurate models for the statistical mechanics of real neural networks, it becomes natural to ask whether there is a thermodynamics that emerges as  $N \rightarrow \infty$ . While heat and temperature do not have any meaning in these systems, thermodynamics is about the interplay of energy and entropy (Sethna, 2021; Kivelson, Jiang, and Chang, 2024), and these have clear significance for networks of neurons. We have written the probability distribution over patterns of activity as

$$P(\sigma) = \frac{1}{Z} e^{-E(\sigma)}, \quad (99)$$

so the effective energy  $E(\sigma)$  is just the negative log probability of a state. The negative logarithm of the probability in turn has an information-theoretic meaning as the length of the ideal code word for describing each pattern of activity, or more simply as the natural measure of how surprised we should be when we observe that pattern (Shannon, 1948; Cover and Thomas, 1991; Mézard and Montanari, 2009).

As a reminder, when we compute the partition function in Eq. (99), we have

$$Z = \sum_{\sigma} \exp [-E(\sigma)] \quad (100)$$

$$= \int dE \rho(E) e^{-E}, \quad (101)$$

where the density of states

$$\rho(E) = \sum_{\sigma} \delta[E - E(\sigma)] \quad (102)$$

becomes smooth at large  $N$  such that

$$\rho(E) \approx \frac{1}{\Delta E} e^{S(E)}, \quad (103)$$

where  $S(E)$  is the microcanonical entropy and  $\Delta E$  is a scale to get the units right. We expect, as usual, both energy and entropy to be proportional to the number of degrees of freedom  $N$  such that

$$E = N\epsilon, \quad (104)$$

$$\lim_{N \rightarrow \infty} \frac{S(E)}{N} = s(\epsilon = E/N), \quad (105)$$

and hence

$$Z \rightarrow \frac{N}{\Delta E} \int dE \exp [-Nf(\epsilon)], \quad (106)$$

where the free energy density  $f(\epsilon) = \epsilon - s(\epsilon)$ . At large  $N$  the dominant states are those with energy per degree of freedom  $\epsilon_*$  such that  $\partial s(\epsilon)/\partial \epsilon = 1$  and

$$Z \approx \frac{N}{\Delta E} e^{-Nf(\epsilon_*)} \int dE \exp [Ns''(\epsilon_*)(\epsilon - \epsilon_*)^2 + \dots]. \quad (107)$$

Thus, the “stiffness” that holds the log probability of states near its typical value is the negative second derivative of the entropy, and the resulting variance in the energy density is the specific heat  $c = 1/[-s''(\epsilon_*)]$ .

Equation (107) makes it clear that something special happens if  $s''(\epsilon_*) \rightarrow 0$  such that the (linear) stiffness holding the energy near its typical value vanishes. Formally, the variance of the energy, and hence the specific heat, diverges as  $N \rightarrow \infty$ . This is a critical point.

In statistical mechanics we have the equivalence of ensembles, telling us that what we compute at a fixed temperature is essentially the same as what we compute with fixed energy if the number of degrees of freedom is large (Sethna, 2021). In information theory the corresponding idea is “typicality,” i.e., that almost all the states that we actually see have the same log probability (Cover and Thomas, 1991; Mézard and Montanari, 2009). When the specific heat diverges, the fluctuations in log probability become very large, so the approach to typicality at large  $N$  becomes anomalously slow.

The fact that the microcanonical entropy is an increasing function of the energy means that states that are individually less likely are more numerous. For neurons there is a useful intuition based on the fact that spikes are less likely than silences. Thus, particular states in which more neurons are active are less probable than those in which fewer neurons are active. But there are more ways of arranging  $K$  spikes among  $N$  neurons when  $K$  is larger (until  $K = N/2$ , which essentially never happens). This trade-off between the frequency and multiplicity of states is exactly the trade-off between energy and entropy.

The typical states that we observe have an energy such that  $dS/dE = 1$ , which means that the trade-off between the frequency and multiplicity is balanced. Usually, this balancing

is local, but at a critical point it extends over a wider range of energies or frequencies.

In a finite population of neurons, we can never see a true divergence in the specific heat. What we can do is to ask whether the specific heat or variance in log probability is large when compared with hypothetical networks that have similar but slightly different properties. One way to construct such networks is to introduce a fictitious temperature,

$$P(\sigma) = \frac{1}{Z} e^{-E(\sigma)} \rightarrow \frac{1}{Z(T)} e^{-E(\sigma)/T}. \quad (108)$$

Varying  $T$  gives us one slice through the space of possible networks: at large  $T$  we find models where neurons are more active and less correlated than in the real network, while at small  $T$  we find models where neurons are less active and more correlated.

The initial exploration of thermodynamics for  $N = 40$  cells in the retinal network showed that the specific heat

$$c(T) = \frac{\langle (\delta E)^2 \rangle}{NT^2} \quad (109)$$

was large, and further that there is a peak in  $c(T)$  close to the model of the real network at  $T = 1$  (Tkačik *et al.*, 2006, 2009). This means that real networks have an unusually large dynamic range for the surprise carried by individual patterns of activity and that this is a property not shared by plausible but slightly different networks. We also can construct hypothetical networks in which individual elements of the correlation matrix are chosen at random from the observed distribution of matrix elements, and maximum entropy models for these randomized networks show almost identical thermodynamic behavior. But we can build random networks in this way at larger  $N$ , with the prediction that the peak of the specific heat should be even larger and closer to  $T = 1$  for  $N \sim 100$ . This prediction was confirmed in an analysis of next-generation experiments with  $N = 100\text{--}160$  (Tkačik *et al.*, 2015).

One may reasonably object by stating that temperature is an artificial construct. Perhaps more reasonable is to divide the effective energy function into one piece that controls the activity of individual neurons and one that controls their interaction, then ask what happens as we change the strength of interactions while keeping the mean activity of each neuron fixed. As an example, we can generalize the  $K$ -pairwise model of Eq. (45) to write

$$E_{2\alpha}(\sigma; \alpha) = E_{\text{ind}}(\sigma) + \alpha E_{\text{int}}(\sigma), \quad (110)$$

$$E_{\text{ind}}(\sigma) = \sum_{i=1}^N h_i(\alpha) \sigma_i, \quad (111)$$

$$E_{\text{int}}(\sigma) = \frac{1}{2} \sum_{i \neq j} J_{ij} \sigma_i \sigma_j + V \left( \sum_{i=1}^N \sigma_i \right). \quad (112)$$

Note that to fix the mean activity of each neuron, we must adjust the local field  $h_i$  as we change the interaction strength  $\alpha$ . If we start with the parameters that describe a population of

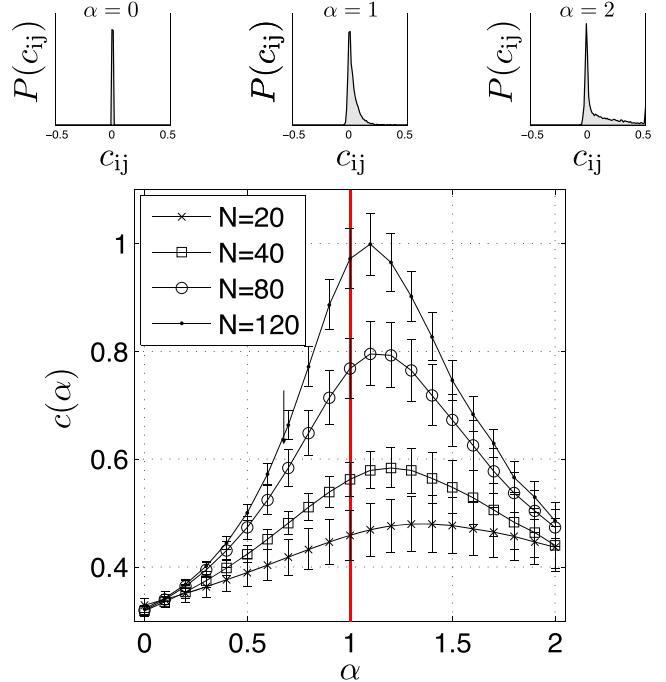


FIG. 28. Specific heat vs strength of interactions. We construct a series of maximum entropy models for networks such that all neurons have the same mean activity as in the real network but varying interactions and hence correlations; see Eqs. (111) and (112). Main panel: specific heat  $c(\alpha)$  vs interaction strength  $\alpha$  for different populations of  $N$  cells chosen out of an experiment on  $N = 160$  cells in the retina. The error bars are standard deviations across ten networks at each  $N$  and  $\alpha$ . Upper panels: distribution of correlation coefficients for all pairs of neurons at three values of  $\alpha$ ;  $\alpha = 1$  is the real network. From Tkačik *et al.*, 2015.

$N = 120$  neurons in the retina, we obtain the results given in Fig. 28 (Tkačik *et al.*, 2015).

As we change  $\alpha$ , we generate networks that in many ways are plausible. The extreme  $\alpha = 0$  describes neurons that turn on and off independently, which is extreme. But even  $\alpha = 2$  describes a network in which pairwise correlation is still reasonable, with a sharper peak at  $c_{ij} = 0$  and a longer tail. The strength of correlations varies monotonically with  $\alpha$ , but the specific heat does not—there is a peak within  $\sim 10\%$  of  $\alpha = 1$ . This peak is higher and closer to  $\alpha = 1$  at larger  $N$ . If we compare the  $K$ -pairwise model to the pure pairwise model, the peak is higher, sharper, and closer to the real system in the more accurate model. These effects also are clearer when the retina is responding to more naturalistic stimuli, even though the pattern of correlations is not simply inherited from the visual inputs; see Sec. VI.D.

We have emphasized that typical states in a Boltzmann-like distribution are those in which the trade-off between the frequency and multiplicity of states is balanced locally; at a critical point, this balance extends over a broader range of probabilities. A striking feature of the maximum entropy models learned from the retina is that the frequency-multiplicity balance extends almost perfectly over a finite range of probabilities such that the entropy is a nearly linear function of energy. This can be seen over a limited dynamic

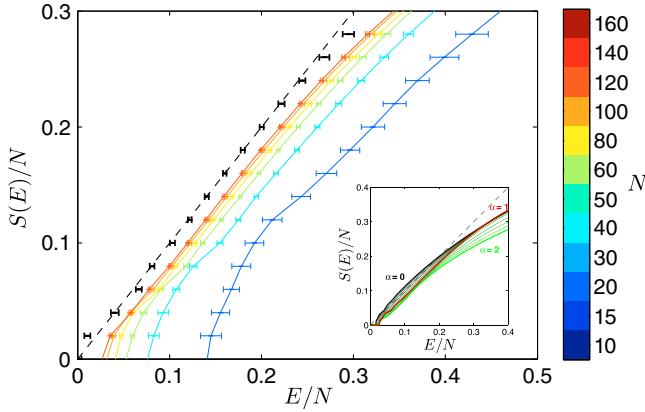


FIG. 29. Entropy vs energy in maximum entropy models for neural activity in the retina. Main panel: results for models at varying  $N$ , with black points based on extrapolation  $N \rightarrow \infty$ . The error bars are standard deviations across multiple networks at each  $N$ , and the dashed line is  $S = E$ . Inset: results at  $N = 120$  with varying  $\alpha$ , as in Fig. 28, showing that the near linearity of entropy vs energy breaks down as we move away from  $\alpha = 1$ . From Tkačik *et al.*, 2015.

range by directly counting states in the raw data, but the models make this prediction across 10 orders of magnitude in probability; see Fig. 29. Importantly, these models make correct predictions over this full range, as seen in Fig. 10. The linearity of entropy versus energy is equivalent to Zipf’s law for the rank ordered probabilities of individual states (Mora and Bialek, 2011) and breaks if we move away from  $\alpha = 1$ . The near linearity of entropy versus energy is seen also in much simpler maximum entropy models that capture the probability that  $K$  out of  $N$  neurons are active but discard information about the identity of the cells (Tkačik *et al.*, 2013).

The results in this section point strongly to the idea that real networks of neurons are poised at nongeneric values of their underlying parameters, generating phenomenology that we associate with critical behavior in simpler systems. We can construct models that are close to the real system but different, quantitatively, and aspects of this behavior fall away. In the (admittedly coarser) observations on the human brain, it even seems that one can drive the system away from critical behavior through anesthesia.

### C. Bridges between dynamics and thermodynamics

Notions of criticality in dynamics and thermodynamics seem very different. But one can also build maximum entropy models for temporal sequences of states, for example, matching pairwise temporal correlations. As noted, this is sometimes called maximum caliber (Pressé *et al.*, 2013; Ghosh *et al.*, 2020). Among other things, this dispels the idea that maximum entropy describes only equilibrium systems. For networks of neurons, we could be interested in either an autonomous description of the dynamics or a description that is locked to external signals, for example, the visual inputs to the retina. There also have been dynamical maximum entropy models for flocks of birds (Cavagna *et al.*, 2014).

If we try to match pairwise correlations not just at equal time but also at unequal times, we are asking quite a lot of the data and arrive at a complicated model. As a first try, one can build models for the summed activity of the network, that is, for the number of neurons  $K_t$  that are active in a small window of size  $\Delta\tau$  surrounding the time  $t$  (Mora, Deny, and Marre, 2015). As noted, applied to single time points this model focuses attention on the surprising trade-off between the probability and numerosity of network states with different numbers of spikes (Tkačik *et al.*, 2013).

Concretely, we ask for the maximum entropy model that matches the distribution of the number of active neurons at one moment in time  $P_N(K)$  from Eq. (41) and the joint distribution at pairs of times  $P(K_t, K_{t+\tau})$  for  $\tau = 1, 2, \dots, v$ . The resulting model is of the form

$$P_{\text{dyn}}(\{K_t\}) = \frac{1}{Z_{\text{dyn}}} \exp[-E_{\text{dyn}}(\{K_t\})], \quad (113)$$

$$E_{\text{dyn}}(\{K_t\}) = \sum_{t=1}^{\tilde{T}} V_N(K_t) - \sum_{t=1}^{\tilde{T}} \sum_{\tau=1}^v J_\tau(K_t, K_{t+\tau}), \quad (114)$$

where  $\tilde{T}$  is the large window of our observations, not to be confused with the temperature  $T = 1/\beta$ . We have to adjust the potential  $V_N(K)$  to match  $P_N(K)$ , and we adjust the interactions  $J_\tau(K, K')$  to match the joint probabilities  $P(K_t, K_{t+\tau})$  that  $K$  and  $K'$  neurons are active in bins separated by  $\tau$ . Because this is a one-dimensional model with local interactions, it can be solved exactly using transfer matrix methods, thereby avoiding a Monte Carlo simulation.

This approach was applied to experiments on a population of  $N = 185$  neurons in the rat retina responding to videos of randomly moving bars. The binary variables  $\sigma_i(t)$  mark the spiking versus silence of neuron  $i$  in a bin of width  $\Delta\tau = 10$  ms surrounding the time  $t$ ,  $K_t = \sum_i \sigma_i(t)$  (Marre *et al.*, 2012). As previously discussed, since we are matching correlations between pairs of times, we can test the model by looking at triplets. Specifically, we can consider

$$\begin{aligned} C_3(K, K', K'') &= P(K_t = K, K_{t+1} = K', K_{t+2} = K'') \\ &\quad - P_N(K)P_N(K')P_N(K''), \end{aligned} \quad (115)$$

which measures (connected) correlations among the numbers of active neurons in three successive time bins. The number of distinct triplets becomes large, so this was tested on a subgroup of  $N = 61$  cells, as shown in Fig. 30(a); the agreement between theory and experiment is excellent (Mora, Deny, and Marre, 2015).

Models that capture temporal correlation also give us a chance to look more deeply at the trade-off between probability and numerosity of network states. Again, we generalize to vary the inverse temperature  $\beta$ ,

$$P_{\text{dyn}\beta}(\{K_t\}) = \frac{1}{Z_{\text{dyn}}(\beta)} \exp[-\beta E(\{K_t\})], \quad (116)$$

with  $E_{\text{dyn}}(\{K_t\})$  the same function as in Eq. (114). We expect the mean energy to be proportional both to the number of

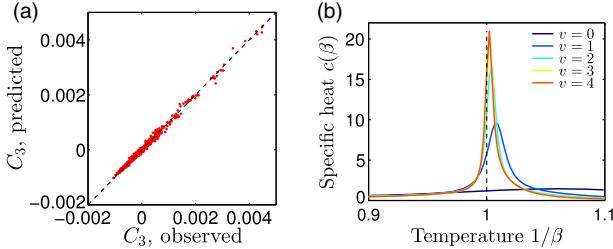


FIG. 30. Maximum entropy models for the dynamics of summed activity across  $N = 185$  neurons in the retina. (a) Correlations among three successive time bins from Eq. (115) computed for a subgroup of  $N = 61$  cells. (b) Specific heat as a function of a fictitious temperature for the full population of  $N = 185$  cells. Different curves are for models that match pairwise temporal correlations in different numbers of time bins. Note the higher, sharper peak close to the real system at  $\beta = 1$  as the model matches more of the data. From Mora, Deny, and Marre, 2015.

neurons  $N$  and to the duration of our observations  $\tilde{T}$ , so we can define a specific heat

$$c_{\text{dyn}}(\beta) = \frac{1}{N\tilde{T}} \beta^2 \langle (\delta E_{\text{dyn}})^2 \rangle, \quad (117)$$

with results shown in Fig. 30(b). The large variance in log probability occurs only when  $\beta$  is within a few percent of the value  $\beta = 1$  that describes the real network. This becomes clearer as we move to more accurate models, increasing the range  $v$  over which we match the temporal correlations. Quantitatively, the specific heat is  $\sim 50$  times larger than if neurons were uncorrelated. We can think of different values of  $\beta$  as describing possible networks with different levels of correlation, and the sharp peak in specific heat at  $\beta = 1$  means that the real network has collective behavior that is very different from other possible networks, even those that differ subtly.

The analysis of neural avalanches focuses on the summed activity of the network, the same collective variable  $K$  considered here. In simple branching models (Beggs and Plenz, 2003), one can again estimate the specific heat, and it diverges exactly at the critical value of the branching parameter that allows for a power-law distribution of avalanche sizes and durations (Mora, Deny, and Marre, 2015). This suggests that the thermodynamics of trajectories is capturing the same critical behavior as the dynamical analyses but without adjustable parameters in the definition of avalanche events.

#### D. Alternatives

For physicists criticality is an evocative concept. The rich phenomenology of critical points inspired the deep ideas of scaling, culminating the modern formulation of the renormalization group. It is exciting that something of this flavor arises in the complex context of living systems, whether in networks of neurons or swarms of midges (Attanasi *et al.*, 2014c). For biologists, in contrast, it can seem that invoking criticality is an example of imposing physics concepts onto a biological system, and we should worry about this too.

An essential tool in the experimental investigation of critical phenomena is the ability to tune the control parameters, pushing the system toward or away from the critical point and exploring the entire critical region. In addition, we usually have experimental probes that couple directly to the order parameter, whether it is the magnetization in a ferromagnet, the density of a fluid, or the degree of molecular alignment in a liquid crystal. For networks of neurons, these tools are largely absent.

An interesting exception is provided by culturing networks of neurons in a dish, where one can manipulate the microscopic parameters. By changing the mix of excitatory and inhibitory neurons, one can see transitions in the behavior of the network: changes of just a few percent in the relative populations of the two cell types produce dramatic qualitative effects that are reminiscent of phase transitions (Chen and Dzakpasu, 2010).

More generally, modern experiments provide us with data analogous to the record of a Monte Carlo simulation, the simultaneous trajectories of all the microscopic elements (neurons) over time. The challenge is to draw inferences from these data about where the real network is poised in the phase diagram of possible networks. As explained, the construction of maximum entropy models provides us with one way of doing this.

The maximum entropy model consistent with pairwise correlations is an Ising model, with the activity of neurons in the role of spins, and it thus is tempting to think of the coefficients  $J_{ij}$  as interactions between the neurons. This language seems natural for physicists, but we should be careful. Even in magnets we know that these are effective interactions, often mediated by fluctuations in additional degrees of freedom that we do not account for directly. In the extreme, a magnetic dipole interaction can be thought of as arising from each spin interacting independently only with the local magnetic field, rather than directly with other spins.

To make these connections explicit, it is useful to change from  $s_i = \{0, 1\}$  to the more familiar Ising variable  $s_i = 2\sigma_i - 1 = \pm 1$ , and to change sign conventions for the fields and couplings. Then the conventional Ising model with pairwise interactions,

$$P(s) = \frac{1}{Z} \exp \left[ \sum_i h_i s_i + \frac{1}{2} \sum_{ij} J_{ij} s_i s_j \right], \quad (118)$$

can be rewritten as

$$P(s) = \frac{1}{Z} \left[ \frac{\det J}{(2\pi)^N} \right]^{1/2} \int d^N \phi \exp \left[ -\frac{1}{2} \sum_{ij} \phi_i (J^{-1})_{ij} \phi_j + \sum_i (h_i + \phi_i) s_i \right]. \quad (119)$$

But because the spins appear linearly in the exponential, this can be factorized as

$$P(s) = \int d^N \phi \mathcal{P}(\phi) \prod_{i=1}^N P_{\text{ind}}(s_i | h_i + \phi_i), \quad (120)$$

where the distribution of field  $\phi$  is given by

$$\mathcal{P}(\phi) = \frac{2^N}{Z} \left[ \frac{\det J}{(2\pi)^N} \right]^{1/2} \exp[-\mathcal{H}(\phi)], \quad (121)$$

$$\mathcal{H}(\phi) = \frac{1}{2} \sum_{ij} \phi_i (J^{-1})_{ij} \phi_j - \sum_i \ln \cosh(\phi_i + h_i), \quad (122)$$

and the conditional distribution for each neuron (or spin) responding independently is, as always,

$$P_{\text{ind}}(s|\psi) = \frac{e^{\psi s}}{2 \cosh \psi}. \quad (123)$$

As an aside, one might worry that the matrix  $J$  is not invertible, or that it has negative eigenvalues that cause  $\mathcal{P}(\phi)$  to be ill defined. But with  $s = \pm 1$  we can always add terms to the diagonal of  $J$  that serve only to shift the zero of energy but will solve these problems.

Thus, as in the textbook derivations of mean-field theory (Sethna, 2021; Kivelson, Jiang, and Chang, 2024), we can trade interactions of neurons with one another for a picture in which they respond independently to fluctuating fields. Models with the structure of Eq. (120) often are referred to as latent variable models (Everitt, 1984) since the behavior that we observe  $\{s_i\}$  is controlled by some underlying hidden or latent variables  $\{\phi_i\}$ . Latent variable models are popular in the neuroscience literature, where they sometimes are presented as an alternative to the physicists' models for interacting neurons. We see that this is a false dichotomy since the different models are mathematically equivalent.

Ultimately, we want to understand whether the latent variable description changes our interpretation of the evidence for critical behavior. But first we should ask whether this description is a compelling alternative, regardless of where real networks are in their phase diagram. The latent variable or effective field description is especially useful if it simplifies the model, and indeed advocates of this description emphasize that it is simpler than the Ising model or, more precisely, that simple versions of the latent variable approach do as well as the Ising model. One clear possibility for neural systems is that the effective fields have a direct meaning for the brain, perhaps as the variables that neural activity is encoding, or are genuinely external to the network, such as sensory inputs.

In the hippocampus, for example, we might imagine that the latent variable is the position of the animal since we know that this is represented by the population of place cells; see Secs. IV.C and V. But in a sufficiently long recording, we can estimate the probability of each cell being active as a function of position  $\mathbf{x}$ , which is the classical place field  $F_i(\mathbf{x})$  as in Eq. (53). If position is the only latent variable, then cells are independent given the position and the joint distribution becomes

$$P_{\text{ind place}}(\sigma) = \int d\mathbf{x} \mathcal{P}(\mathbf{x}) \prod_{i=1}^N F_i(\mathbf{x})^{\sigma_i} [1 - F_i(\mathbf{x})]^{1-\sigma_i}, \quad (124)$$

where  $\sigma_i = \{0, 1\}$  as in previous sections and  $\mathcal{P}(\mathbf{x})$  is the distribution of positions seen in the experiment. In this model

all correlations are inherited from the animal's movement through the space  $\mathbf{x}$ . Importantly, this construction involves no free parameters.

We can test the independent place cell model of Eq. (124) in the same way that we have tested the maximum entropy models. If we compute the mean activity of each cell, it will be correct by construction. The pairwise correlations are a nontrivial prediction, and the matrix  $C_{ij}$  looks roughly correct element by element, but the eigenvalue spectrum is qualitatively incorrect, as noted in Sec. IV.C. Triplet correlations are significantly underestimated [see Fig. 31(a)], and in the bulk of small correlations there is essentially zero correlation between the data and the predictions of an independent place cell model [see Fig. 31(b)]. These results should be compared with the success of the pairwise maximum entropy model in Fig. 20. We conclude that in the hippocampus an extreme version of the latent variable scheme—in which the only latent variable is position—fails dramatically (Meshulam *et al.*, 2017).

For the retina there has been the explicit suggestion that any successes of the maximum entropy approach should be understood in a latent variable model where the latent variables are determined by the visual stimulus itself (Aitchison, Corradi, and Latham, 2016). This is the stimulus-dependent maximum entropy model of Eqs. (61) and (62), but with all interactions  $J_{ij} = 0$  and a complicated relation between the visual input and the time-dependent local fields  $h_i(t)$  allowed. No matter how complex this relation, the model predicts that if we show the same movie to the retina many times, then at a fixed moment in the movie there should be no correlations among the neurons, since the latent variables are fixed. The challenge in testing this prediction is that the probability of complete silence in the network is significant, even when  $N > 100$  cells, and in these silent moments, one cannot compute the correlations.

If we compute conditional correlations only in moments where both cells in the pair generate more than a handful of spikes, then we can indeed find examples where the correlations are near zero, but there also are many examples where the conditional correlations are *larger* than the overall

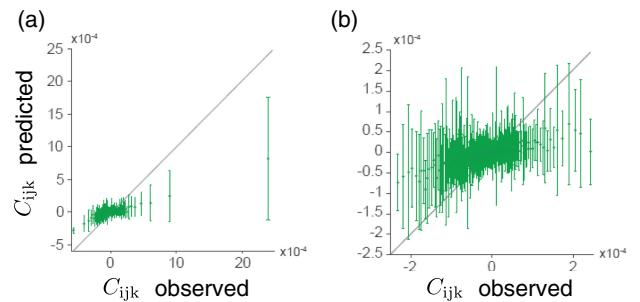


FIG. 31. Failure of a latent variable model for  $N \sim 100$  cells in the mouse hippocampus. Predicted vs observed triplet correlations [see Eq. (83)] calculated in a model where the latent variable is the position of the mouse [see Eq. (124)]. (a) Full dynamic range of the data, binned along the  $x$  axis as in Fig. 20. (b) Expanded view of the small correlations, which constitute the bulk of the data. From Meshulam *et al.*, 2017.

correlations, which is the opposite of the prediction of the latent variable model. There even are many pairs of neurons whose overall correlation is near zero, but at particular moments in a repeated movie the correlations are strong with either sign. These results eliminate a model in which the visual inputs serve as the latent variables to explain the correlation structure of the activity in the retina (Tkačik *et al.*, 2015).<sup>20</sup>

Thus, in two cases that have been studied carefully, we cannot find a description in which latent variables correspond to external stimuli. But if the latent variables are hidden from us, then it is not clear whether these variables are external to the network or emergent from the network dynamics itself. As an example, it has been suggested that the summed activity of all the cells in a network can serve as a latent variable (Aitchison, Corradi, and Latham, 2016), which would be like saying that the magnetization of the Ising model is a latent variable. In the mean-field limit this almost works (the natural latent variable is actually the conjugate field), but there is no doubt that the magnetization is emergent. One also can verify that, in the systems that we have discussed, different neurons are not conditionally independent given the summed activity; see, in the case of the retina, Tkačik *et al.* (2015).

Latent variable models would be especially attractive if one could achieve an accurate description with a small number of these variables. If the Ising model description is accurate, then a small number of latent variables requires the rank of the coupling matrix  $J_{ij}$  to be small. Even better would be a case where we could identify the latent variables with measurable quantities, but we have seen that it does not work to identify these variables with quantities that are genuinely external, such as a sensory stimulus. There are popular models for the encoding of low-dimensional sensory or environmental variables in which the “latent” variable that represents these signals in fact emerges from network interactions (Ben-Yishai, Bar-Or, and Sompolinsky, 1995; Tsodyks and Sejnowski, 1995; Zhang, 1996). If we ask about the distribution over observable network states, then the mathematical description is the same no matter whether the latent variable is external or emergent.

There is a simple but compelling argument for how the seemingly mysterious linearity of entropy versus energy, as well as the associated signatures of criticality, can arise from fluctuating fields (Schwab, Nemenman, and Mehta, 2014). It is useful to place this discussion in the context of the mean-field ferromagnet (Sethna, 2021; Kivelson, Jiang, and Chang, 2024).

The mean-field model is a collection of spins  $s \equiv \{s_i\}$  governed by the energy function

<sup>20</sup>To be clear, the visual inputs do generate correlations among neurons in the retina. The point is that these are not the only source of correlations, and the separation into externally driven and internally generated correlations does not provide an immediate simplification. We also emphasize that this is not a separation that is available to the brain under natural conditions. Further, the retina adapts to the distribution of its inputs, so there is no fixed mapping from correlations in the stimulus to correlations among neurons.

$$E_{MF}(s) = h \sum_{i=1}^N s_i - \frac{J}{2N} \sum_{i,j=1}^N s_i s_j \quad (125)$$

$$= N[hm - (J/2)m^2], \quad (126)$$

where the magnetization

$$m = \frac{1}{N} \sum_{i=1}^N s_i. \quad (127)$$

This describes a system in which all spins experience the same magnetic field and all pairs of spins interact equally. The factor of  $1/N$  in the interactions insures that energy and entropy are proportional to  $N$  as  $N \rightarrow \infty$ . Now we can follow the same arguments that lead from the general pairwise Ising model in Eq. (118) to the latent variable description in Eqs. (120) and (122), but this case is easier because there is only one latent field. The result is that the partition function can be written as

$$Z_{MF} \equiv \sum_{\{s_i\}} e^{-\beta E_{MF}(s)} \quad (128)$$

$$= \sqrt{\frac{2\pi}{J}} \int d\phi \exp [-Nf_{MF}(\phi, h)], \quad (129)$$

$$f_{MF}(\phi, h) = \frac{\phi^2}{2J} - \ln \cosh(\phi + h), \quad (130)$$

where for simplicity we choose units where the thermal energy  $1/\beta = 1$ .

At large  $N$  the integral in Eq. (129) is dominated by a single value of the latent field  $\phi = \phi_*$  that minimizes the free energy, that is,

$$\left. \frac{\partial f_{MF}(\phi, h)}{\partial \phi} \right|_{\phi=\phi_*} = 0. \quad (131)$$

If the second derivative

$$\left. \frac{\partial^2 f_{MF}(\phi, h)}{\partial \phi^2} \right|_{\phi=\phi_*} = \kappa \quad (132)$$

is of order unity, then fluctuations in the latent variable will be on a scale  $\delta\phi \sim 1/N^{1/2}$ . Because all spins couple equally to the latent field, these fluctuations produce correlations among all spins, but the scale of these correlations is small. The result is that covariance matrix elements  $C_{ij} \sim 1/N$ . The critical point is the place where the second derivative  $\kappa \rightarrow 0$ , and fluctuations in the latent field become anomalously large,  $\delta\phi \sim 1/N^{1/4}$ . The idea of Schwab, Nemenman, and Mehta (2014) was to turn this around: since criticality is marked by large fluctuations in the latent field, then if external signals drive large fluctuations in the latent variable, they could also generate the signatures of criticality, generically.

To make this concrete, consider a collection of Ising spins that all couple to the same magnetic field  $h$ , but this field itself is drawn from a distribution  $Q(h)$ . Crucially, this

distribution is imposed on the system by external inputs rather than being an emergent property of the interactions. Then the joint distribution for the state of all the spins is

$$P_{\text{latent}}(s) = \int dh Q(h) \prod_{i=1}^N P(s_i|h) \quad (133)$$

$$= \int dh Q(h) \prod_{i=1}^N \frac{e^{hs_i}}{2 \cosh(h)}. \quad (134)$$

Equations (133) and (134) become

$$P_{\text{latent}}(s) = \frac{1}{2^N} \int dh Q(h) \exp[-Nf(m, h)] \quad (135)$$

$$f(m, h) = -hm + \ln \cosh(h), \quad (136)$$

where as before the magnetization

$$m(s) = \frac{1}{N} \sum_{i=1}^N \sigma_i. \quad (137)$$

Once again, when  $N$  is large, the integral over fields is dominated by the value that minimizes the free energy density  $f(m, h)$ ,

$$h_*(s) = h_*(m) = \tanh^{-1}(m), \quad (138)$$

as long as  $Q(h_*)$  is nonzero. In making this argument, the distribution of fields must be externally imposed and thus cannot have an  $N$  dependence. The result is that the probability of any state  $s$  depends only on the magnetization  $m(s)$ ,

$$P_{\text{latent}}(s) = \exp[-E(m)], \quad (139)$$

$$E(m)/N = -h_*(m)m + \ln \cosh[h_*(m)] + \dots, \quad (140)$$

where we drop terms that are independent of  $m$  or vanish as  $N \rightarrow \infty$ . The entropy at fixed energy is then the entropy at fixed magnetization,

$$S(m)/N = -\frac{1+m}{2} \ln\left(\frac{1+m}{2}\right) - \frac{1-m}{2} \ln\left(\frac{1-m}{2}\right). \quad (141)$$

After some algebra Eqs. (140) and (141) can be combined to give  $S(m)/N = E(m)/N$ , as with the data in Fig. 29. More generally we see that  $d^2S(E)/dE^2 = 0$ , which is equivalent to the divergence of the specific heat, a core signature of criticality.

This argument generalizes beyond the case of a single fluctuating field coupled to the spins. Not only can one have multiple fields, but they can couple to more complex functions of the system state. What is required is that a mean-field approximation be valid so that at large  $N$  each state  $\{s_i\}$  picks a single value for all the latent variables out of some broad distribution (Schwab, Nemenman, and Mehta, 2014). This generality is striking, and it is natural to ask whether this

explains the signatures of criticality that we have seen experimentally.

The emergence of critical phenomenology from latent variable models cannot be completely generic. As an example, the previously discussed independent place cell model is one model in the broad class considered by Schwab, Nemenman, and Mehta (2014), but it is not an arbitrary model. We already know that this model does not explain the correlation structure that we see in populations of  $N \sim 100$  cells in the hippocampus, and it also does not predict the observed  $S/N = E/N$ ; see Fig. S7 of Tkačik *et al.* (2015). In this sense a biologically plausible version of the latent field model evades the conditions for the generic emergence of critical behavior at reasonable  $N$ .

Similarly, we can try to account for the large fluctuations in summed activity that we see in recordings from  $N \sim 1500$  hippocampal neurons using models where all cells are driven by a common field, as in Eqs. (135) and (136). The regime where we have a generic prediction of  $S/N = E/N$  is where the stiffness of the free energy restricts the fluctuations,

$$\delta h_f \sim \left[ N \frac{\partial^2 f(m, h)}{\partial h^2} \right]^{-1/2}, \quad (142)$$

to be much smaller than the range of fields  $\delta h_Q$  spanned by the distribution  $Q(h)$ . If we use this approach to look at the data analyzed in Sec. V, we find that  $\delta h_f \sim \delta h_Q$  within a factor of 2. While not conclusive, since the correct model surely is more complex, this result suggests that this network is not in the regime where fluctuating external fields explain apparent criticality.

Closely related to the utility of latent variable models is the question as to whether networks of real neurons are in a mean-field limit. For the mean-field ferromagnet [see Eq. (125)] there is one order parameter  $m$  and one (emergent) latent variable  $\phi$ . For the mean-field spin glass, by contrast, there are infinitely many order parameters (Mézard, Parisi, and Virasoro, 1987). An intermediate case is the Hopfield model of associative memory near its storage capacity. Now the number of latent variables is proportional to the number of neurons, but mean-field methods still are useful in the  $N \rightarrow \infty$  limit (Amit, Gutfreund, and Sompolinsky, 1987). These examples emphasize that mean-field theory with its emergent latent variables can succeed even if the system is not in any sense low dimensional. There is a genuine breakdown of mean-field theory as we approach critical points, but it is not clear whether the more sophisticated “high-dimensional” mean-field theories could account for critical phenomenology without such tuning. What we need is a test for the “mean-fieldness” of a system, and this is a deeper theoretical problem.

The prediction of critical behavior in the maximum entropy approach emerges, with no adjustable parameters, in models that account in detail for the correlation structure among neurons. While low-dimensional latent variable models have a regime in which they can generate signatures of criticality without fine-tuning, there is no example that we know of where such models account for all the observed correlation structure. Taking seriously what the maximum entropy

principle is doing—building *minimally structured* models—it seems that the observed correlation structure implies criticality but that networks could be critical without this correlation structure.

The models that we study also give us ways of generating surrogate data, for example, a network in which every neuron has the same mean activity as in the real network but correlations that are weaker or stronger, as in Eqs. (110)–(112). As we tune away from the real network, we lose signatures of criticality such as the linearity of entropy versus energy; see Fig. 29. Relatedly, models that capture more of the real correlation structure have stronger signs of criticality, as in Fig. 30. Thus, while there surely are critical networks with different correlation structures, plausible changes in correlations drive the predicted behavior of the network away from criticality.

Taken together, these observations suggest that the signatures of criticality that we see in networks of neurons are not a generic consequence of the system being driven by external fields. Instead, it really does seem that these systems are tuned to a special point in their parameter space. Ordinarily, such fine-tuning is worrisome, but networks of neurons have an array of mechanisms for adaptation and learning that allow stabilization of nongeneric behaviors. One clear example is that the oculomotor integrator (see Sec. VI.A) is tuned, continuously, based on visual feedback, holding it close to a bifurcation point and thereby allowing for long, emergent timescales (Major, Baker, Aksay, Mensh *et al.*, 2004; Major, Baker, Aksay, Seung, and Tank, 2004).

It is useful to compare the problem of criticality in networks of neurons to the corresponding problem in flocks of birds and swarms of insects; see Sec. A.2. In these animal groups, there are good reasons to think that interactions are local, so it is tempting to think that observation of long-range spatial correlations would be *prima facie* evidence for critical behavior, but this is not quite correct. First, even in equilibrium systems, the breaking of a continuous symmetry generates Goldstone modes, and fluctuations along these modes will generate long-range correlations. Maximum entropy models that match local correlations provide an example of this idea, which offers a quantitative description of the directional fluctuations in flocks with no free parameters; see Figs. 41(g) and 41(h). Second, nonequilibrium effects in animal groups can generate effectively nonlocal interactions, and this is central to theories of active matter (Toner and Tu, 1995, 1998; Marchetti *et al.*, 2013). But there are arguments that these effects are smaller than expected in real flocks (Mora *et al.*, 2016), so the observed long-range correlations in speed fluctuations may indeed provide evidence of critical behavior. In natural swarms one sees finite size and dynamical scaling behaviors that provide more direct evidence for criticality, regardless of the model (Attanasi *et al.*, 2014c; Cavagna *et al.*, 2017). While each example must stand on its own, it is an old dream that tuning to criticality might unify our understanding of disparate living systems.

## VII. RENORMALIZATION GROUP FOR NEURONS

Physicists are known for our appreciation of simplified models, perhaps even to the point of oversimplification

(Devine and Cohen, 1992). The complexity of living systems is in obvious tension with this drive for simplification. We can perhaps sympathize with biologists who worry that our theoretical impulses may be mismatched to the richness of life’s molecular details. A useful response is that there is nothing special about biology: in condensed matter physics and statistical mechanics, we routinely describe the macroscopic behavior of materials using models that are much simpler than the underlying microscopic mechanisms. These simplified models succeed not because we are lucky but because of the renormalization group (Wilson, 1979, 1983).

The central idea of the RG is to ask how our description of a system changes, systematically, as we change our scale. The crucial qualitative result is that many different microscopic mechanisms flow toward the same macroscopic behavior as we “zoom out” to look at longer length scales. This means that we can understand large-scale phenomena quantitatively if we can assign them to the correct universality class, even if we cannot get all the small scale details right. This gives us the license to write relatively simple models of complex systems (Anderson, 1984). We want to exercise this license in the context of the brain. To do this, we need to understand how to implement the RG when many of our usual guides (locality, symmetry, etc.) are absent. We then can ask whether there is any sign that simplification emerges from the data as we zoom out from individual neurons to more coarse-grained variables.

### A. Taking inspiration from the RG

The development of the renormalization group is one of the great chapters of theoretical physics from the second half of the 20th century, with origins in efforts to understand matter at both short and long distances (Gell-Mann and Low, 1954; Kadanoff, 1966). These ideas crystallized in the early 1970s and played a central role in revolutionizing our understanding of the strong interaction among elementary particles, critical phenomena at second-order phase transitions, the transition to chaos, etc. (Wilson, 1983). How can these ideas help us to think about networks of neurons?

In the standard formulation of the RG for statistical physics, we start with a set of variables  $z_{\ell_0} \equiv \{z_i(\ell_0)\}$  defined on some microscopic length scale  $\ell_0$ . Our description of these variables is given by a Hamiltonian that in turn specifies the Boltzmann distribution  $P_{\ell_0}(z)$ , or perhaps we will be interested in the dynamics generated by this Hamiltonian. We then imagine coarse graining the variables to average out the details on length scales below some  $\ell > \ell_0$ . The result is a new set of variables  $z_\ell$ , and we ask for the effective Hamiltonian that governs these variables. If we think of the Hamiltonian as being built from different kinds of interactions, then the effective strengths of these interactions change as we change scale from  $\ell_0$  to  $\ell$ , and the RG invites us to follow this flow as we vary  $\ell$ .

Seen from the most macroscopic perspective, it is natural to think of the scale  $\ell$  as varying continuously, but neurons are discrete entities. Even in the conventional examples from statistical mechanics, variables often live on a lattice, so we are restricted to making discrete steps along the flow,  $\ell \rightarrow b\ell$  with integer  $b$ . If we start with  $\ell_0$  equal to the lattice spacing, then at the first step of coarse graining, we average together  $b^d$

variables, where  $d$  is the dimensionality of the lattice. If we iterate  $k$  times, we arrive at coarse-grained variables constructed from  $K = b^{kd}$  of the original variables. Although we usually think of the RG flow as a function of length scale  $\ell$ , we can also think of the scale as being set by  $K$ , opening the possibility that “scale” in a network of interacting variables is more general than the spatial scale. Jona-Lasinio (1975) emphasized that one can go even further, thinking directly about flow in the space of probability distributions without any reference to Hamiltonians.

Renormalization group flows typically drive many different starting distributions  $P_{\ell_0}(z)$  toward the same  $P_\ell(z)$  as  $\ell$  becomes large. Along this trajectory parameters and predictions of the distribution exhibit simple scaling behaviors as a function of  $\ell$  or as a function of the number of variables  $K$  that are grouped together. A familiar example is the central limit theorem: if the original variables are sufficiently weakly correlated, then the distribution approaches a Gaussian at large  $K$ , and the variance of the coarse-grained variable scales as  $1/K$ . The RG predicts that more interesting starting points flow toward stable non-Gaussian distributions, with moments scaling as nontrivial powers of  $K$ .

The renormalization group approach provides a framework to understand how we can go from discrete Ising spins on a lattice to a description of smoothly varying local magnetization, or from the positions and momenta of individual molecules to the density of a fluid and the velocity of its flow. In these examples the coarse-graining operation is guided by symmetry and locality. Perhaps the most successful development of RG ideas in a biological context has been for flocks of birds and swarms of insects, where the ideas of symmetry and locality continue to be useful; see Sec. A.2. For networks of neurons, where connections can span distances encompassing thousands of cells, the principle of locality is less of a guide, and there are no obvious symmetries. How then do we choose a coarse-graining strategy?

Perhaps a more serious problem in taking inspiration from the renormalization group is that the RG is formulated as an approach to understanding theories or models, and not as an approach to data analysis. There has been work on renormalization group approaches to continuum models of neurons with local connectivity (Steyn-Ross and Steyn-Ross, 2016) and the Wilson-Cowan model (Tiberi *et al.*, 2022). Path integral methods strongly suggest that spiking networks have different collective behavior (Keup *et al.*, 2021), and these tools have been used to develop an RG for networks with moderately realistic spiking neurons (Brinkman, 2025). We hope there will be more work in these directions. But, keeping to the spirit of the discussion thus far, we ask the following question: How can we use the RG to guide the analysis of emerging experiments on large populations of real neurons?

To address these questions we rely on two key ideas. First, as emphasized, modern experiments on the electrical activity in networks of neurons give us access to something analogous to the trajectory of a Monte Carlo simulation on a statistical physics model, albeit a model that we do not know how to write. Thus, we can follow the approach used in now classical analyses of such simulations, for example, that by Binder (1981). We start with raw data on the most microscopic scale,

construct coarse-grained variables, and follow various features of the distribution of these variables as we change the scale of coarse graining.

Second, we use the measured pairwise correlations as a guide to which neurons are “neighbors” in the absence of locality (Bradde and Bialek, 2017). In one version (see Sec. VII.B), this involves averaging together the activities of the most correlated cells, building clusters of neurons that are analogous to block spins (Kadanoff, 1966). In another version (see Sec. VII.C), we successively filter out linear combinations of the population activity that make small contributions to the overall variance, and this is analogous to the momentum shell construction (Wilson, 1983). We later see that both these approaches uncover simple, precise, and reproducible scaling behaviors that now have been confirmed in multiple brain areas from multiple organisms. We then discuss the implications of these results and some future directions in Sec. VII.D.

## B. By analogy with real-space methods

Renormalization group methods in statistical physics rest on a notion of coarse graining, averaging over microscopic details. If we start with variables  $\{z_i\}$  that live on a regular lattice, it is natural to do this by combining variables with their neighbors, as in Fig. 32. Formally, we can write

$$z_i \rightarrow \tilde{z}_i = f \left( \sum_{j \in \mathcal{N}_i} z_j \right), \quad (143)$$

where  $\mathcal{N}_i$  is a neighborhood surrounding site  $i$ . If the function  $f(\cdot)$  is linear, then we are just averaging over a neighborhood, and, for example, this will lead from discrete Ising-like variables to a more continuous local magnetization if we iterate. If  $f(\cdot)$  is a threshold function, then we can implement majority rule such that clusters of Ising-like variables are mapped into Ising-like variables on the sparser lattice, as in the original block spin construction (Kadanoff, 1966).

In a system with local interactions, the variables in the neighborhood typically are the most strongly correlated with

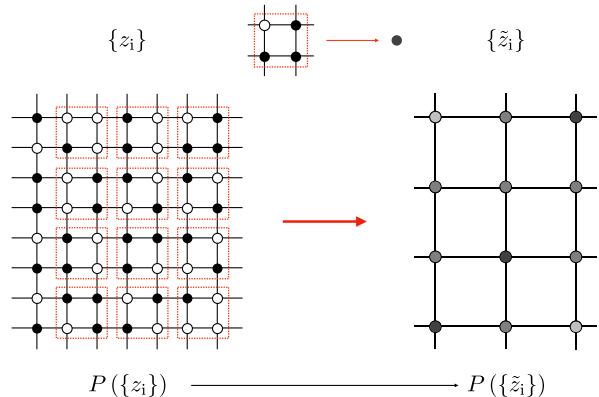


FIG. 32. Coarse graining on a regular lattice. We start with binary (black and white) variables  $\{z_i\}$  and replace  $2 \times 2$  blocks with the average of these variables  $\{\tilde{z}_i\}$ , which are shown as gray levels. The interesting question is, What happens to the joint distribution as we coarse grain, not just once but iteratively?

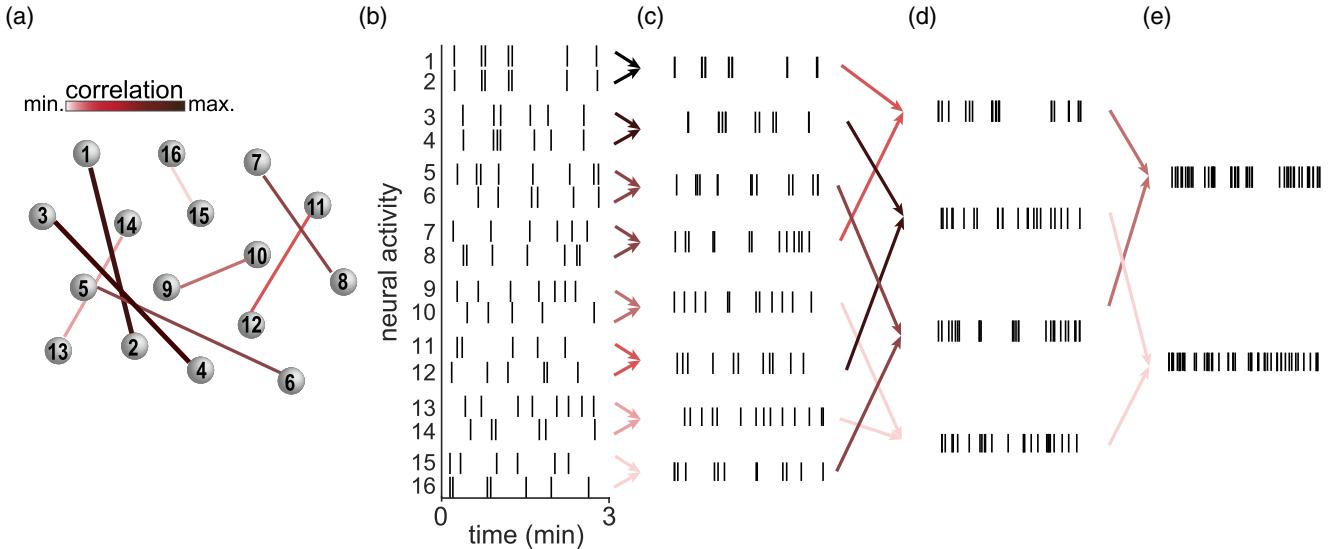


FIG. 33. Coarse graining neural activity. (a) A small group of neurons with links indicating the most strongly correlated pairs and the strength of these correlations. (b) Schematic sequence of action potentials from these cells. (c) Coarse graining by summing the activity in highly correlated pairs. (d) Finding the most strongly correlated pairs of coarse-grained variables in (c) and coarse graining again by summing. The strengths of the correlations are color coded as in (a). (e) One more iteration of this “real-space” coarse graining. From Meshulam *et al.*, 2018.

one another. This suggests that even if we do not have a notion of neighborhood, we can make progress by searching for the most correlated variables and using these to build the clusters that we use in coarse graining. A schematic of how this can work for neural activity is shown in Fig. 33.

We start with variables  $\{\sigma_i\}$  as before, describing the patterns of activity ( $\sigma_i = 1$ ) and silence ( $\sigma_i = 0$ ) across all the neurons  $i = 1, 2, \dots, N$  in a small window of time. To emphasize that this is the most microscopic description, we write this as  $\sigma_i = \sigma_i^{(1)}$ . Then as before, we can compute the means, covariance, and correlation matrices,

$$m_i^{(1)} = \langle \sigma_i^{(1)} \rangle, \quad (144)$$

$$C_{ij}^{(1)} = \left\langle \left[ \sigma_i^{(1)} - m_i^{(1)} \right] \left[ \sigma_j^{(1)} - m_j^{(1)} \right] \right\rangle, \quad (145)$$

$$c_{ij}^{(1)} = \frac{C_{ij}^{(1)}}{\sqrt{C_{ii}^{(1)} C_{jj}^{(1)}}}. \quad (146)$$

Now we search for the maximal nondiagonal element in the matrix  $c_{ij}$  of correlation coefficients. Because  $c$  is symmetric, this defines a pair of cells that we can write as  $i_1, j_*(i_1)$ . If we take all elements associated with these two cells out of the correlation matrix and search again for the maximal off-diagonal element, we find another pair  $i_2, j_*(i_2)$ . We can repeat this process until we have identified  $N/2$  pairs of cells  $\{i, j_*(i)\}$ . We then define coarse-grained variables by averaging these pairs together,

$$\sigma_i^{(2)} = \sigma_i^{(1)} + \sigma_{j_*(i)}^{(1)}, \quad (147)$$

which we can relabel so that  $i = 1, 2, \dots, N/2$ .

Importantly, we can iterate this process across scales. We compute the correlation matrix of the variables  $\{\sigma_i^{(2)}\}$  and search again for the maximally correlated pairs  $\{i, j_*(i)\}$ , then define

$$\sigma_i^{(3)} = \sigma_i^{(2)} + \sigma_{j_*(i)}^{(2)}, \quad (148)$$

etc. At each stage we have  $N_k = \lfloor N/2^{k-1} \rfloor$  variables remaining. This coarse graining produces clusters of  $K = 2, 4, \dots, 2^{k-1}$  neurons, and the variable  $\sigma_i^{(k)}$  is the summed activity of cluster  $i$ .

We emphasize that one could have different criteria for coarse graining and different ways of combining the variables. We return to some of these points in Sec. VII.D, but for now we explore what happens when we apply this simplest scheme to a network of real neurons. The first such example used the experiments on the activity of more than 1000 neurons described in Sec. V (Meshulam *et al.*, 2018, 2019).

We are interested in how the probability distributions transform and flow as we pass through successive scales of coarse graining. Looking at the joint distribution  $P(\{\sigma_i^{(k)}\})$  is essentially impossible. But much can be learned by looking at slices through this distribution, even the distribution of individual coarse-grained variables, as with the magnetization in the Ising model (Binder, 1981).

Since this coarse graining is based simply on adding the neighboring variables, the first moment of the distribution of the individual variables must scale linearly,

$$M_1(k) \equiv \frac{1}{N_k} \sum_{i=1}^{N_k} \langle \sigma_i^{(k)} \rangle = \frac{1}{N_k} \sum_{i=1}^{N_k} m_i^{(k)} = K M_1(1), \quad (149)$$

where again after  $k$  steps we have  $N_k$  clusters each involving  $K = 2^{k-1}$  of the original variables. The first nontrivial

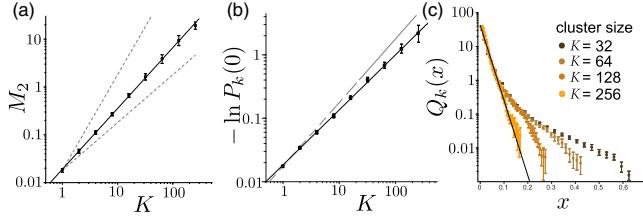


FIG. 34. Three slices through the distribution of coarse-grained variables. (a) Variance of the activity vs the real-space coarse-graining scale from Eq. (150). The solid line is  $M_2 \propto K^\alpha$ , where  $\alpha = 1.4 \pm 0.06$ . The dashed lines are predictions for independent ( $\alpha = 1$ ) or perfectly correlated neurons ( $\alpha = 2$ ). (b) Probability of silence vs the coarse-graining scale. The solid line is Eq. (151), with  $\beta = 0.88 \pm 0.01$ , while the dashed line is the expectation for independent neurons,  $\beta = 1$ . (c) Distribution of the normalized nonzero activity, as defined in Eq. (152). Adapted from Meshulam *et al.*, 2018, 2019.

question is about the second moment, or the variance in activity,

$$M_2(K) \equiv \frac{1}{N_k} \sum_{i=1}^{N_k} \langle (\sigma_i^{(k)} - m_i^{(k)})^2 \rangle. \quad (150)$$

Note that if neurons are independent, we expect  $M_2(K) \propto K$ , and many weakly correlated populations should approach this behavior at large  $K$ . If neurons are perfectly correlated, however, we expect  $M_2(K) \propto K^2$ . Looking at the data, in Fig. 34(a) we see that for neurons in the hippocampus  $M_2(K) \propto K^\alpha$ , with  $\alpha = 1.4 \pm 0.06$ . This nontrivial scaling is visible over more than two decades.

We can take another slice through the distribution by asking for the probability  $P_k(0)$  that the coarse-grained variable  $\sigma_i^{(k)} = 0$ . Since we started with variables  $\sigma_i = \{0, 1\}$ , this is the same as asking for the probability that all of the neurons inside the cluster of size  $K = 2^{k-1}$  are silent. If the neurons are independent, we expect a simple scaling  $P_k(0) \propto \exp(-aK)$  and once more expect to see this at large  $K$ , even if the cells are weakly correlated. Experimentally, we see in Fig. 34(b) that

$$P_k(0) = \exp(-aK^\beta), \quad (151)$$

with the exponent  $\beta = 0.88 \pm 0.01$ . Again, scaling is precise over more than two decades.

If we imagine making an explicit model for the joint activity of all the neurons inside one of the clusters, perhaps in the form of the pairwise models in Eq. (82), then the probability of complete silence is dependent only on the partition function,  $P_k(0) = 1/Z$ . This generalizes if we include higher-order terms such that Fig. 34(b) probes the effective free energy, which apparently behaves as  $F(K) = -aK^\beta$ . Since  $\beta < 1$ , the free energy is subextensive, and hence the free energy per neuron will vanish in the thermodynamic limit. This is consistent with the equality of entropy and energy that we saw for retinal neurons in Sec. VI.B; see Fig. 29.

More generally, if we define the normalized variable  $x = \sigma^{(k)}/K$ , then

$$P_k(x) = P_k(0)\delta_{x,0} + [1 - P_k(0)]Q_k(x). \quad (152)$$

Figure 34(c) shows the evolution of  $Q_k(x)$  as  $K$  increases. We see that the tail of the distribution is gradually absorbed into the bulk, which seems to approach a fixed form  $Q(x) \sim e^{-x/x_0}$ . If the neurons were independent, the central limit theorem would drive this distribution toward a Gaussian, but instead we see the emergence of a fixed non-Gaussian form.

In addition to looking at the distribution of single coarse-grained variables, we can look at the covariance matrix of the microscopic variables within each cluster of size  $K$ . The eigenvalue spectrum of this covariance matrix depends on the rank scaled by  $K$ , and there is a substantial region over which the spectrum is a power  $\lambda \sim (K/\text{rank})^\mu$ , with  $\mu = 0.71 \pm 0.06$ , although this is less crisp than the other examples of scaling.

Our discussion thus far has focused on the distribution of variables at a single moment in time. In the applications of the RG that we understand, however, we can often observe dynamic scaling (Hohenberg and Halperin, 1977). Intuitively, fluctuations on longer length scales take longer to relax because the underlying interactions are local. What is non-trivial is that correlation functions for variables coarse grained to different length scales collapse to a universal form if we measure time in units of the correlation time, and this correlation time itself varies as a power of the length scale. An elegant example of these ideas in a fully biological context is provided by dynamic scaling of the velocity fluctuations in natural swarms of insects (Cavagna *et al.*, 2017).

With networks of neurons we do not expect locality to be a good guide, but it still is plausible that more strongly coarse-grained variables will have slower dynamics, and we can search for dynamic scaling. Concretely, we define the correlation function for individual variables at coarse-graining scale  $k$ ,

$$\tilde{C}_i^{(k)}(t) = \left\langle [\sigma_i^k(t_0) - m_i^{(k)}][\sigma_i^k(t_0 + t) - m_i^{(k)}] \right\rangle, \quad (153)$$

and we can then normalize and average over the clusters to give

$$C^{(k)}(t) = \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{\tilde{C}_i^{(k)}(t)}{\tilde{C}_i^{(k)}(0)}. \quad (154)$$

Dynamic scaling is the hypothesis that the dependence on scale is captured by a single correlation time,

$$C^{(k)}(t) = C[t/\tau_c(k)], \quad (155)$$

with  $\tau_c(k) \propto K^z$ . In Fig. 35 we see that all of this works for the population of hippocampal neurons. We note that the dynamic range of correlation times accessed in this experiment is limited: at short times by the dynamics of the indicator molecules and at long times by the small value of the exponent  $z = 0.16 \pm 0.02$ .

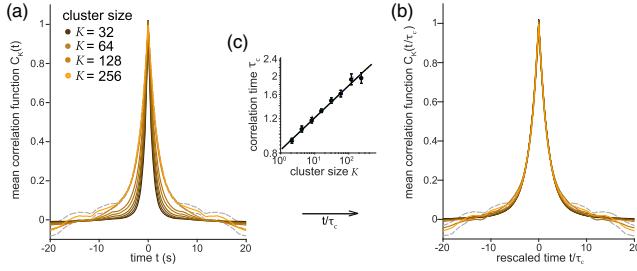


FIG. 35. Dynamic scaling across more than 1000 neurons in the hippocampus. (a) Mean correlation functions for coarse-grained variables [see Eq. (154)] in clusters of  $K = 2, 4, \dots, 256$  neurons (lightest orange corresponds to the largest cluster), with larger clusters exhibiting slower dynamics. The dashed gray line depicts  $\pm 1$  standard deviation across the  $K = 256$  neuron clusters. (b) Collapse under scaling of the time axis in Eq. (155). (c) Correlation time vs cluster size fit to  $\tau_c \propto K^z$ , with  $z = 0.16 \pm 0.02$ . From [Meshulam et al., 2018](#).

It is important that these scaling behaviors are not somehow driven by our choice to describe neural activity with binary variables. In these experiments neural activity was recorded by imaging of fluorescence from indicator molecules that provide a continuous signal as in Figs. 6 and 16. We can follow the same steps of coarse graining for these continuous signals, and the results are the same ([Meshulam et al., 2019](#)).

In the full theoretical structure of the RG, scaling exponents are signatures of universality classes. Before we can ask about universality, we have to ask about reproducibility, especially in such complex systems. As a first step, the same analyses have been done with data from experiments on multiple mice. Because scaling is precise across more than two decades, the error bars in determining the exponents in individual mice are small, which sets a high standard for reproducibility.<sup>21</sup> For example, the exponent describing the scaling of the free energy [see Fig. 34(b)] is  $\tilde{\beta} = 0.87 \pm 0.014 \pm 0.015$  for the mean, the rms error in single experiments, and the standard deviation across experiments in three mice. This holds out hope that we have uncovered features of the emergent behavior that are reproducible in the second decimal place.

A more ambitious search for universality was undertaken by [Morales, di Santo, and Muñoz \(2023\)](#). They analyzed experiments that are part of a large effort at the Allen Institute for Brain Science, in this case using multiple neuropixels probes (see Fig. 5) to record more than 100 neurons from each of many different areas of the mouse brain, simultaneously. Note that in addition to exploring many different brain regions, the technique for recording activity is completely different than in the hippocampal imaging data analyzed in Figs. 34 and 35. Nonetheless, all aspects of scaling are reproduced across all these brain areas. Examples include the scaling of the variance in coarse-grained activity [see Fig. 36(a)] and dynamic scaling [Fig. 36(b)]. There are even hints that the scaling exponents may be common across areas.

<sup>21</sup>To be clear, multiple experiments could “agree” simply because the error bars on the individual experiments are large. This would be much less compelling.

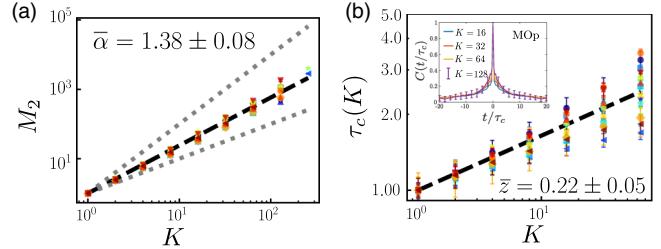


FIG. 36. Scaling in multiple distinct areas of the mouse brain. Neuronal data after the real-space (direct-correlation) coarse-graining procedure. (a) Variance of the coarse-grained activity vs cluster size for neurons in 16 different brain regions (depicted as different markers); cf. Fig. 34(a). (b) Dynamic scaling for the same brain areas. Correlation time vs cluster size; cf. Fig. 35. Inset: Decay of the autocorrelation function for the neurons in one brain region (primary motor cortex) showing the collapse once time is rescaled. From [Morales, di Santo, and Muñoz, 2023](#).

As larger-scale experiments become more common, we hope that scaling will be tested not only in multiple areas but also in multiple organisms ([Munn et al., 2024](#)).

### C. By analogy with momentum shell methods

In problems where *scale* really is a length scale, coarse graining is a gradual blurring out of spatial detail, much as what happens when we look through a microscope and defocus. In that analogy the spatial pattern is Fourier transformed and then reconstructed using only a limited range of wavelengths. Concretely, if we start with variables  $\phi(\vec{x})$  in a  $d$ -dimensional space with coordinates  $\vec{x}$ , the coarse-graining operation becomes

$$\phi(\vec{x}) \rightarrow \phi_\Lambda(\vec{x}) = z_\Lambda \int_{|\vec{k}| < \Lambda} \frac{d^d k}{(2\pi)^d} e^{i\vec{k}\cdot\vec{x}} \tilde{\phi}(\vec{k}), \quad (156)$$

$$\tilde{\phi}(\vec{k}) = \int d^d x e^{-i\vec{k}\cdot\vec{x}} \phi(\vec{x}), \quad (157)$$

where  $\Lambda = \pi/\ell$  cuts off contributions below a length scale  $\ell$  and  $z_\Lambda$  serves to (re)normalize the variables. In the microscopic analogy, this compensates for the loss of contrast as we defocus. As in real space, we are interested in how the probability distribution  $P_\Lambda[\phi_\Lambda]$  evolves as a function of the cutoff  $\Lambda$ . Since the Fourier variables are continuous (in the limit of a large system), we can make infinitesimal changes  $\Lambda \rightarrow \Lambda - d\Lambda$ . In quantum mechanics a wave with wave vector  $\vec{k}$  describes particles with momentum  $\vec{p} = \hbar\vec{k}$ , so that average over the details in a range  $\Lambda - d\Lambda < |\vec{k}| < \Lambda$  is equivalent to integrating out a momentum shell ([Wilson and Kogut, 1974](#)).

Momentum is conserved in systems with translation invariance. Even if we do not have a meaningful notion of momentum, spatial translation invariance privileges the Fourier transform. As an example, if variables  $z_i$  live on a lattice of points  $\vec{x}_i$ , translation invariance means that the covariance matrix elements  $C_{ij}$  can depend only on the difference in positions,

$$C_{ij} = C(\vec{x}_i - \vec{x}_j), \quad (158)$$

this matrix is diagonalized in a Fourier basis,

$$\sum_{j=1}^N C_{ij} u_{jr} = \lambda_r u_{ir}, \quad (159)$$

$$u_{jr} \propto \exp(i\vec{k}_r \cdot \vec{x}_j), \quad (160)$$

where we can put the modes in order by the rank  $r$  of the eigenvalue.

In the usual applications of the RG, large momenta correspond to small eigenvalues of the covariance matrix. This suggests that we can construct coarse-grained variables by filtering out the “modes” that correspond to small eigenvalues, without reference to space or momenta (Bradde and Bialek, 2017). This connects coarse graining to a more familiar data analysis technique, principal component analysis (Shlens, 2014).

Concretely, if we start with microscopic variables  $\{\sigma_i\}$ , we can compute the covariance matrix as usual,

$$C_{ij} = \langle (\sigma_i - \langle \sigma_i \rangle)(\sigma_j - \langle \sigma_j \rangle) \rangle, \quad (161)$$

and we then have eigenvalues and eigenvectors as in Eq. (159). We choose the rank  $r$  such that  $\lambda_1 \geq \lambda_2 \dots \lambda_N$ . We can define a projection onto the  $\hat{K}$  modes that make the largest contribution to the variance,

$$\hat{P}_{ij}(\hat{K}) = \sum_{r=1}^{\hat{K}} u_{ir} u_{jr}, \quad (162)$$

$$\phi_{\hat{K}}(i) = z_i(\hat{K}) \sum_j \hat{P}_{ij}(\hat{K}) [\sigma_i - \langle \sigma_i \rangle], \quad (163)$$

with the normalization  $z_i(\hat{K})$  such that  $\langle [\phi_{\hat{K}}(i)]^2 \rangle = 1$ .

As before, we want to follow the distribution of the individual coarse-grained variables  $P_{\hat{K}}(\phi_{\hat{K}})$ ; the results are shown in Fig. 37(a). To be sure that we have control over the full matrix  $C_{ij}$ , we look at clusters of  $N = 128$  neurons identified through the aforementioned real-space coarse graining. We can then filter out half of the modes such that  $\hat{K} = 64$ , resulting in a distribution  $P_{\hat{K}}(\phi_{\hat{K}})$  that still has some fine structure. If we reduce to  $\hat{K} = 32$ , these wiggles disappear, but the distribution remains asymmetric with long tails. This pattern continues as we reduce to  $\hat{K} = 16$  and then  $\hat{K} = 8$ , and in these last steps the distribution hardly changes. This suggests that as we coarse grain, the distribution flows toward a fixed form. Importantly, this form is very different from the Gaussian that would be guaranteed by the central limit theorem if correlations were weak.

The intuition behind dynamic scaling is that fluctuations on larger length scales relax more slowly, and we have seen that this generalizes to a network of neurons even though the meaning of scale now is more abstract; see Fig. 35. By transforming to the basis that diagonalizes the covariance matrix, we have isolated the modes of fluctuation that are

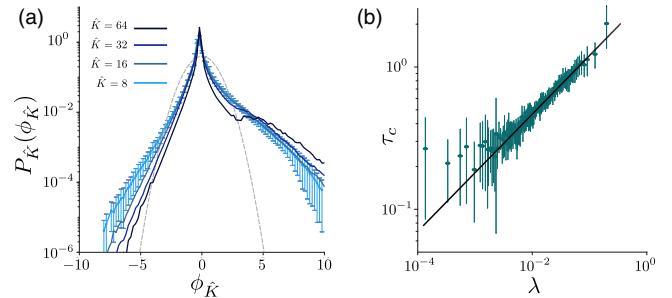


FIG. 37. Coarse graining in groups of  $N = 128$  neurons via momentum shells. (a) Following the distribution of individual coarse-grained variables from Eq. (163). Different colors correspond to keeping different numbers of modes  $\hat{K}$ , as in the legend; the dashed line is a Gaussian for comparison. (b) Dynamic scaling of the correlation time for fluctuations in mode  $r$  [see Eq. (165)] vs the associated eigenvalue of the covariance matrix,  $\tau_c(r) \propto \lambda_r^{z'}$ , with  $z' = 0.37 \pm 0.04$ . From Meshulam *et al.*, 2018.

independent at second order, and it is natural to ask how fluctuations along these modes relax. Variations along mode  $r$  are defined as

$$\tilde{\phi}_r = \sum_{i=1}^N [\sigma_i - \langle \sigma_i \rangle] u_{ir}, \quad (164)$$

and the correlation function is

$$C_r(t) = \langle \tilde{\phi}_r(t_0) \tilde{\phi}_r(t_0 + t) \rangle. \quad (165)$$

Dynamic scaling is the statement that all these correlations collapse when time is scaled by a single correlation time, and that this correlation time itself has a power-law dependence of scale. In the usual examples, this means that  $\tau_c \propto |\vec{k}|^z$  (Hohenberg and Halperin, 1977), but near a critical point the eigenvalues of the covariance matrix also have a power-law dependence on  $|\vec{k}|$ , so we can test directly for  $\tau_c \propto \lambda^{z'}$  as shown in Fig. 37(b). As before, the shortest correlation times are limited by the response of the fluorescent proteins that report on electrical activity, and the longest times are limited by the magnitude of the dynamic scaling exponent. Nonetheless, we observe reasonably precise scaling across two decades in  $\lambda$ .

The dynamic exponent  $z'$  that one finds by looking at the correlation times of the modes should be related to the one we see via coarse graining in real space,  $z$  [see Fig. 35(c)], through the exponent  $\mu$  that describes the decay of the eigenvalues of the covariance matrix,  $\tilde{z} = \mu z'$ . This works, although error bars are large (Meshulam *et al.*, 2018). More importantly, these results indicate that the network has no single characteristic timescale, but rather a continuum of timescales that can be accessed by probing on different scales.

#### D. RG as a path to understanding

If we believe there is an underlying simplicity to be found amid the complexity of neural network function and activity,

we might want to pause for a moment to convince ourselves that the RG can actually lead us there. This quest now feels attainable given the explosive experimental progress in obtaining datasets with increasing number of neurons, as in the aforementioned examples. While we may not know how to manipulate temperature or magnetization in the brain, we are gaining decades in the sheer number of monitored neurons.

The renormalization group is a powerful theoretical structure. Because we do not have a microscopic model for neural dynamics, we are not yet able to exploit this structure. What we have done instead is to adopt an RG-inspired approach to data analysis that has been described as a “phenomenological renormalization group” (Nicoletti, Suweis, and Maritan, 2020). The most interesting outcome would be the flow of probability distributions toward some fixed, non-Gaussian form and the appearance of power-law scaling along this trajectory, as happens at a critical point in well understood equilibrium statistical mechanics problems. This is what has been found, first in the initial application to the hippocampus and now in other systems; scaling exponents are reproducible and perhaps even universal. It is tempting to conclude that the underlying network dynamics must be described by a theory that is at a nontrivial fixed point of the renormalization group.

We should be cautious. Is it possible that some of the behaviors under coarse graining that we associate with RG fixed points could emerge more generically in nonequilibrium systems? Nicoletti, Suweis, and Maritan (2020) addressed this in an analysis of the contact process where binary variables are turned on with a probability per unit time proportional to the density of active variables at neighboring sites and then deactivate with a fixed probability per unit time. This model has one parameter, the proportionality constant in the activation rate, and there is a critical value that depends on the geometry of the network (Marro and Dickman, 1999). Below the critical point, the fully inactive state is absorbing, so the question is whether the phenomenological RG can distinguish the critical point from supercritical behaviors.

Perhaps surprisingly, one can see (weakly) nontrivial scaling behavior in some quantities even away from the critical point, as with the variance in activity shown in Fig. 38(a). But other quantities show clear deviations from scaling, even very close to criticality, as with the correlation times in Fig. 38(b). What is unambiguous is that the probability distributions of coarse-grained variables flow toward a nontrivial fixed form at the critical point, and toward a Gaussian otherwise. We can see this by coarse graining in real space [see Fig. 38(c)] or via momentum shells [see Fig. 38(d)]. Nicoletti, Suweis, and Maritan (2020) emphasize that the phenomenological RG can identify critical points unambiguously, but only if we check the full range of behaviors.

As with the related discussion of criticality in Sec. VI.D, it has been suggested that some of the phenomena uncovered by iterative coarse graining can be reproduced in a model where neurons respond independently to latent fields (Morrell, Sederberg, and Nemenman, 2021). In this view scaling and the flow toward fixed distributions are approximate, and it is not clear why scaling exponents should be reproducible across animals. A broader notion of universality, which is hinted at in Fig. 36 (Morales, di Santo, and Muñoz, 2023), would be even more difficult to understand.

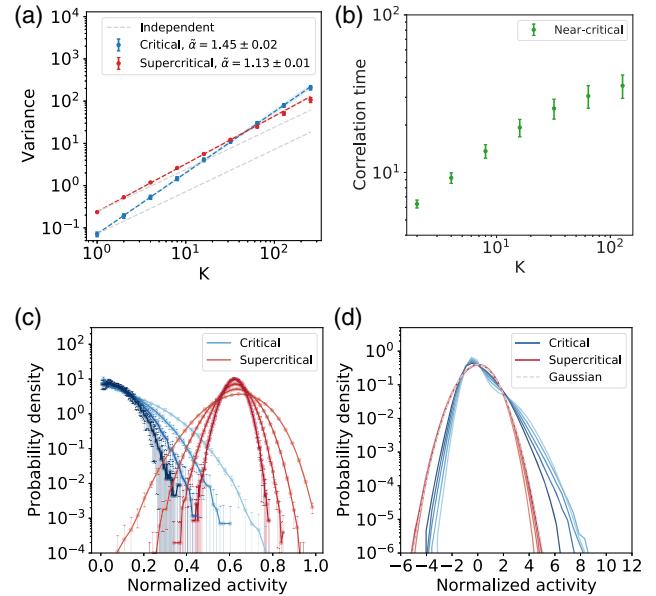


FIG. 38. Coarse graining of the contact process. (a) Variance of activity vs the scale of coarse graining in real space, as in Fig. 34(a). Behavior at criticality (blue line) is clearly different from the supercritical case (red line), which departs systematically but weakly from the expectations for independent variables (dashed lines). (b) Correlation time vs the scale of coarse graining in real space, as in Fig. 35. The control parameter is set close to its critical value, and we see hints of scaling at small  $K$  but clear departures at large  $K$ . (c) Distribution of individual coarse-grained variables for  $K = 32, 64, 128$ , and  $256$  at criticality (blue lines) and away from criticality (red lines). In both cases we see flow toward a fixed distribution, but away from criticality this is Gaussian, as expected from the central limit theorem. (d) As in (c) but with coarse graining via momentum shells, keeping  $N/8, N/16, N/32, N/64$ , and  $N/128$  of the modes. From Nicoletti, Suweis, and Maritan, 2020.

The suggestion that scaling behaviors emerge generically from latent variable models is incorrect. Consider models in which the effective field acting on each neuron  $i$  is a linear combination of  $K$  latent variables drawn from a Gaussian distribution. If the fields are weak, then the covariance matrix of neural activity has the same rank as the covariance matrix of the fields. This simple result breaks down at stronger fields, but even in the limit of infinitely strong fields there remains a gap in the eigenvalue spectrum of the covariance matrix, at least for typical choices of parameters, such that it is impossible to recover precise scaling behaviors.

We note that a concrete, biologically motivated model of latent fields (the independent place cell model discussed in Sec. VI.D) fails to exhibit scaling (Meshulam *et al.*, 2018). This result perhaps should not be surprising. In a population of place cells, there are two length scales, the approximate width of the place fields and the mean distance between place field centers. In the one-dimensional (virtual) environment that provides the background for the hippocampal experiments analyzed here, the ratio of these lengths gives us a characteristic number of neurons,  $K_c \sim 18$ . Indeed, analyses of the independent place cell model corresponding to Figs. 34(a) and 34(b) show “breaks” at  $K \sim K_c$ . While these are

approximate statements, they highlight the fact that, in the presence of such obvious scales, the observation of rather precise power-law scaling in both static and dynamic quantities is surprising.

Faced with high-dimensional observations, a natural reaction is to search for a lower-dimensional description. In some sense the renormalization group is the opposite approach ([Bradde and Bialek, 2017](#)). Rather than looking for the correct number of dimensions onto which to project the data, the RG invites us to examine how our description changes as we move the boundary between details that we ignore and features that we keep. Things simplify not because we have fewer degrees of freedom but because the model describing these degrees of freedom flows toward something simpler and more universal. The evidence thus far points toward the existence of such a simplified description. From the theoretical side, initial efforts at an RG analysis of models for networks of more realistic neurons suggest that these are described via new universality classes ([Brinkman, 2025](#)).

What we have not emphasized here is the connection of coarse graining to more functional behaviors. In the hippocampus how is position represented in the coarse-grained variables? More generally, do fine-grained and coarse-grained variables implement different principles for the encoding of the sensory world ([Munn et al., 2024](#))? Can local networks of neurons access different scaling trajectories as the brain switches among different global states ([Castro et al., 2024](#))? As coarse graining becomes a more commonly used tool for the analysis of large-scale neural recordings, we expect these questions to be sharpened and answered.

The most detailed tests of scaling in equilibrium critical phenomena span six decades with better than 1% precision ([Lipa et al., 1996](#)). As described in Secs. [III.B](#) and [III.C](#), the experimental frontier is moving toward recording from  $\sim 10^6$  neurons simultaneously. This opens the possibility of following coarse-graining trajectories across five decades with single cell resolution and of driving error bars down to the 1% level across more limited ranges. The extension of existing tools to organisms with larger brains also means that we will see simultaneous recordings from more neurons in single brain areas, within which scaling seems more likely. We already see signs that quantities that emerge from these analyses can be reproducible to the second decimal place. One possibility is that new, larger experiments will reveal crossovers between different regimes on different scales. Alternatively, the scaling behaviors seen thus far might prove to be essentially exact. Whatever the outcome, it is extraordinary to think that experiments on real, functioning brains could soon reach a precision comparable to those on equilibrium critical phenomena. The corresponding challenge to theory should be clear.

### VIII. OUTLOOK

Statistical physics has long been a source of useful metaphors for emergent behaviors in living systems. All the birds in a flock agreeing to fly in the same direction is like the alignment of spins in a magnet. Recalling memories in the brain is like a spin glass settling into one of many locally stable states. Quietly, examples have emerged that are more

than metaphors. Thus, experiments on single DNA molecules provide the most detailed tests of predictions for the random flight polymer, one of the classical models discussed in statistical mechanics courses ([Bustamante et al., 1994](#); [Marko and Siggia, 1995](#)). The explosion of data on networks of real neurons similarly offers the opportunity to move beyond metaphor.

We have seen that relatively simple statistical physics models—Ising models with pairwise interactions and modest generalizations—provide detailed quantitative descriptions of real networks, from the retina deep into the cortex and hippocampus. Correct predictions are not limited to a few macroscopic or thermodynamic quantities, but include detailed patterns of higher-order correlations and the way in which the activity of each neuron depends on the collective state of the others. Again, it is not only some trends in these quantities but also precise numerical values within experimental error that are being captured. The theories we are discussing may be swept away by the next generation of data, but these results set a standard for what we should demand in comparing theory with experiment.

The state of the field is such that our examples of success still are scattered, and each network that has been studied is different, described, for example, via a different matrix of interactions  $J_{ij}$ . While networks are different, there may be something common about the distribution out of which  $J_{ij}$  are drawn. As neural recordings at large  $N$  become more common and these analysis methods are applied more widely, we hope to see a compelling characterization of this distribution. This would be a major step beyond models for particular networks toward a theory of these networks more generally.

The experimental frontier is moving rapidly. Recordings from  $N \sim 10^3$  cells soon will be routine and  $N \sim 10^6$  will be possible with higher time resolution and a higher signal-to-noise ratio. We have emphasized that one cannot simply carry existing models to larger  $N$  unless the duration of experiments increases in proportion. This is not impossible, as stable recording methods allow visiting the same population of neuron day after day, not only to study nonstationary processes such as learning but also to increase the volume of data from which we can estimate the correlation structures in the network. At the same time, there are new ideas about how to build statistical physics models for networks from sparse data. The goal is to discover an underlying simplicity that allows fewer measurements to characterize the global dynamics, and this will represent real theoretical insight.

Perhaps the most fundamental question that we can sharpen by moving to larger  $N$  is the construction of a thermodynamics for networks of neurons—guided by theory but built from data. Can we convince ourselves that real networks are understandable in the thermodynamic limit? What are the relevant order parameters? Where are real networks in the phase diagram of possible networks? We have glimpsed possible answers to these questions at  $N \sim 100$ . We hope there will be a concerted effort to sharpen these answers through analyses of emerging experiments at larger  $N$ .

The idea that networks of neurons might be poised near a critical point or critical surface has been a continuing source of fascination and controversy. Much of the literature is about

why this would be a good idea or why it cannot be right rather than about the evidence, and we have tried to avoid these more ideological discussions here. What we have seen is that populations of  $N \sim 100$  neurons are described accurately by relatively simple statistical physics models, and that if we change the temperature or the relative strength of different terms in the model, then the parameter settings that describe the real system are close to criticality.

While one can construct models that capture various aspects of critical phenomenology without the underlying structure of a critical point, it is not easy to do this *and* engage with the detailed correlation structure of the real networks. In contrast, critical behavior emerges naturally from the simplest models that are consistent with this structure. Importantly, models off criticality describe plausible networks—for example, with slightly weaker or stronger correlations—but not the ones that we see experimentally. A concrete goal is to draw phase diagrams directly in the space of these measurable quantities.

In our modern understanding, saying that a system is at a critical point means that it is described by a theory that is a fixed point of the RG. More generally, the RG invites us to ask how our description of a system changes as we include more or less detail, and this suggests new approaches to data analysis. It is striking that the first efforts in this direction showed that coarse-grained variables flow to nontrivial distributions, that one can see precise scaling over more than two decades, and that exponents can be reproducible to the second decimal place, with hints of universality. These results should be tested more broadly in the coming years, and it will be important to confront the full structure of the data rather than focusing on one or two features that could by themselves be misleading.

As data collection moves to ever-larger scales, coarse graining becomes a more attractive approach to visualizing system behavior. In conventional applications of the renormalization group, the coarse-graining step is constrained by symmetries and the associated conservation laws, as well as by locality, but these are absent in networks of neurons. A first attempt was to average together the activity of neurons that are the most strongly correlated, and much remains to be explored using this idea. A deeper theoretical challenge is to reexamine the formulation of the RG itself.

Coarse graining is an example of lossy data compression (Cover and Thomas, 1991), and in general one can choose what is preserved in making such compressions, for example, the intelligibility of speech in the compression of acoustic waveforms. The fundamental trade-off is between the bits of information that coarse-grained variables carry about microscopic variables and the bits of information that they carry about “relevant” variables (Tishby, Pereira, and Bialek, 1999). Can we construct RG transformations with different choices for what is relevant?

Neural systems offer many possibilities for alternative formulations of coarse graining as selective compression. We could compress the states of multiple neurons to preserve the information that the coarse-grained activity provides about other neurons in the network, about the future states of the same neurons, or about external quantities such as sensory inputs and motor outputs. This combination of information-theoretic and renormalization group ideas could give new

perspectives on classical results, but it is most exciting to search for new fixed points and hence, by definition, new physics (Koch-Janusz and Ringel, 2018; Gordon *et al.*, 2021; Kline and Palmer, 2022).

One basic question that the combination of information theory with the RG might help answer is how to identify order parameters. In many contexts once we know the order parameter, we can almost immediately write an effective field theory, and the technical apparatus of the RG tells us which terms in this theory are relevant or irrelevant.<sup>22</sup> But finding the order parameter currently relies on inspiration rather than constructive calculation. In some contexts it seems clear that order parameters have an information-theoretic interpretation, for example, as the most compressed variable that provides information about the states of other variables at large spatial separations. A concrete goal is to turn this around and give an information-theoretic *definition* of order parameters that would allow their systematic discovery.

What do the observed scaling behaviors say about the function of neural networks in the life of the organism? Dynamic scaling suggests that we can access dynamics on different timescales simply by averaging together the activity of different combinations of neurons, in the spirit of ideas about “reservoir computing” (Maass, Natschläger, and Markram, 2002). Although attention often is focused on how to learn the correct readout scheme for a particular task, it also is essential for the reservoir to be sufficiently rich. Dynamic scaling means that there is a continuous range of available timescales, out to a longest time set only by the size of the network. We need a theory that links these results with the ability of the brain to make predictions and drive behavior across a wide range of timescales.

Since the brain drives behavior, scaling in neural dynamics suggests that we might find scaling in behavioral correlations across time (Bialek and Shaevitz, 2024), although the search for these correlations is challenging. If we can coarse grain the activity of neurons, we should also be able to coarse grain behavior itself, and this has been used to make notions of hierarchy in behavior more precise (Berman, Bialek, and Shaevitz, 2016). There has been significant progress in mapping high-resolution video of animal behavior into descriptions of postural trajectories or behavioral states (Berman *et al.*, 2014; Mathis *et al.*, 2018; Pereira *et al.*, 2019), and perhaps we can see this as a first step in coarse graining, one that should be unified with subsequent analysis of the state sequences in time. Put more simply, is there an RG for animal behavior?

In a different direction, models of network dynamics already are coarse grained compared to more microscopic models of electrical dynamics in single cells and synapses. Can we build a unified coarse-graining scheme that goes from the molecular description of ion channel dynamics through cellular interactions to collective network dynamics?

If networks of neurons are described via a nontrivial fixed point of the renormalization group, the central theoretical

---

<sup>22</sup>The notion of relevance in the renormalization group is different from the notion of relevance in information theory. But perhaps they are related. In this spirit, see Machta *et al.* (2013).

question is to identify this fixed point theory (or theories if different networks scale differently). Even if scaling eventually breaks down, the fact that we see this behavior over several decades implies that important aspects of network function will be controlled by the underlying fixed point. The understanding of flows toward these fixed points will make precise our intuition that different cellular details map to different classes of macroscopic network behavior.

Not long ago, everything that we have discussed in this section would have seemed like physicists' fantasies that are disconnected from real brains (perhaps there are a few remnants of this). What has changed dramatically is that all these ideas—Ising models and correlation functions, scaling behaviors and the RG, etc.—are connected to quantitative experiments on networks of real neurons. Old worries that experiments on living systems are irreducibly messy have been overcome by demonstrating the levels of precision and reproducibility that we expect in physics. Not all systems are equally accessible to this kind of exploration, but these results set an example for what is possible. We can connect all the way from abstract physics concepts to the details of particular neurons in specific brain regions. Our experimentalist friends will continue to move the frontier, combining tools from physics and biology to make more and more of the brain accessible in this way. The outlook for theory is bright.

## ACKNOWLEDGMENTS

Our exploration of these ideas has been shaped by our interactions with a wonderful group of experimentalist friends: D. Amodei, C. D. Brody, M. J. Berry II, J. L. Gauthier, M. Ioffe, S. A. Koay, A. M. Leifer, O. Marre, R. Segev, J. W. Shaevitz, and D. W. Tank. Our understanding of theoretical issues has been aided enormously by discussions with many collaborators and colleagues: F. Azhar, S. Bradde, A. Cavagna, X. Chen, L. Di Carlo, I. Giardina, C. M. Holmes, K. Krishnamurthy, C. W. Lynn, F. Mignacco, T. Mora, I. Nemenman, R. Pang, S. E. Palmer, L. Ramirez, C. Sarra, D. J. Schwab, E. Schneidman, S. Still, M. Tikhonov, G. Tkačik, and A. M. Walczak. We offer special thanks to those who gave advice about references, sometimes in extremis: E. R. F. Aksay, E. J. Chichilnisky, M. S. Goldman, G. Goodhill, F. M. Rieke, D. J. Schwab, and I. H. Stevenson. Thanks also to the three reviewers of the manuscript, whose deep engagement pointed to many places where we could do better. This work was supported in part by the National Science Foundation through the Center for the Physics of Biological Function (Grant No. PHY-1734030); by the National Institutes of Health through the BRAIN Initiative (Grant No. R01EB026943-01); by the Burroughs Wellcome Fund through a CASI award to L. M.; and by fellowships from the Swartz Foundation, the Simons Foundation, and the John Simon Guggenheim Memorial Foundation. W. B. also thanks the Center for Studies in Physics and Biology at Rockefeller University for its hospitality during part of this work.

## APPENDIX A: SEQUENCES, FLOCKS, AND MORE

Part of what makes maximum entropy models exciting is that they can be used in a wide variety of contexts, perhaps

pointing toward a more general statistical physics of biological networks. Examples range from the evolution of protein families and its connection to protein structure to the propagation of order in flocks of birds, and more. The exchange of methods and ideas among applications to these very different systems has been productive also for thinking about networks of neurons, so we now give a brief (and, one hopes, not too idiosyncratic) survey.

### 1. Protein families

Proteins are polymers of amino acids, and there are 20 amino acids to choose from at each site along the chain. To a large extent, this sequence determines the folded structure and function of the protein. The explosion of data on sequences has been even more dramatic than the explosion of data on networks of neurons, but thoughtful analysis of sequences began as soon as there were a handful to look at, and this played a crucial role in working out the genetic code (Brenner, 1957).

By the late 1970s, it was clear that proteins form families with similar functions and structures (Stroud, 1974), and eventually the sequence data would become plentiful enough that these relationships could be detected without structural or functional measurements (Finn *et al.*, 2014). Proteins are densely packed, and there is a strong intuition that evolutionary changes in one amino acid need to be compensated for by changes in neighboring amino acids (Göbel *et al.*, 1994). By the early 1990s, there were a few families of proteins with enough sequences that one could see signatures of these correlated pairwise substitutions (Neher, 1994).

To be concrete, define a variable  $s_i^\alpha = 1$  if the amino acid at site  $i$  along the chain is of type  $\alpha$ , and  $s_i^\alpha = 0$  otherwise. The full amino acid sequence of one protein then is  $\{s_i^\alpha\}$ , with  $i = 1, 2, \dots, N$  and  $\alpha = 1, 2, \dots, 20$ . If we have  $K$  proteins in a family, we have a larger set of variables  $\{s_i^\alpha(n)\}$ , with  $n = 1, 2, \dots, K$ . This is called a multiple sequence alignment.<sup>23</sup> We can measure the expectation values at each site

$$m_i^\alpha = \frac{1}{K} \sum_{n=1}^K s_i^\alpha(n), \quad (\text{A1})$$

which is the probability that amino acid  $\alpha$  is used at site  $i$  in the family. We can also define the joint probability of amino acids  $\alpha$  and  $\beta$  at sites  $i$  and  $j$ ,

$$C_{ij}^{\alpha\beta} = \frac{1}{K} \sum_{n=1}^K s_i^\alpha(n)s_j^\beta(n). \quad (\text{A2})$$

To synthesize a new family of proteins  $\{\tilde{s}_i^\alpha(n)\}$ , we could ask how similar these one- and two-body statistics are to the original family by computing

---

<sup>23</sup>It is useful to introduce a blank state at  $\alpha = 21$  allowing one protein to have two segments that overlap strongly with others in the family but a small gap in between.

$$\chi^2 = \sum_{i\alpha} W_i^\alpha \left[ \frac{1}{K} \sum_{n=1}^K \tilde{s}_i^\alpha(n) - m_i^\alpha \right]^2 + \frac{1}{2} \sum_{i\alpha} \sum_{j\beta} W_{ij}^{\alpha\beta} \left[ \frac{1}{K} \sum_{n=1}^K \tilde{s}_i^\alpha(n) \tilde{s}_j^\beta(n) - C_{ij}^{\alpha\beta} \right]^2. \quad (\text{A3})$$

In this formulation we can give each term a different weight, perhaps in proportion to the accuracy with which we can estimate each expectation value.

In the early 2000s, [Russ et al. \(2005\)](#) and [Socolich et al. \(2005\)](#) realized that one could use the similarity measure  $\chi^2$  as an energy function and generate new families of proteins from known families via Monte Carlo simulations. Most importantly, rather than just drawing samples out of the distribution, they actually synthesized the proteins and asked whether they fold and function like the naturally occurring members of the family. The short but compelling answer is that if one constrains only the one-body terms (i.e., set  $W_{ij}^{\alpha\beta} = 0$ ), then none of the many proteins synthesized in this way fold. With the two-body terms included, however, a reasonable fraction of all the new proteins synthesized do fold. This was startling, suggesting that pairwise correlations were sufficient to capture the essence of the mapping from the protein structure back to the amino acid sequence.

What was missing from the original analysis was an explicit construction of the underlying probability distribution. As it turns out, in the limit in which families are large ( $K \rightarrow \infty$ ) and the temperature of the Monte Carlo simulation is low, using  $\chi^2$  in Eq. (A3) as an energy function is equivalent to sampling the maximum entropy distribution consistent with one-one- and two-body statistics ([Bialek and Ranganathan, 2007](#)). This distribution has the form

$$P(\{s_i^\alpha\}) = \frac{1}{Z} \exp[-E_p(\{s_i^\alpha\})], \quad (\text{A4})$$

$$E_p(\{s_i^\alpha\}) = \sum_{i,\alpha} h_i^\alpha s_i^\alpha + \frac{1}{2} \sum_{i\alpha} \sum_{j\beta} J_{ij}^{\alpha\beta} s_i^\alpha s_j^\beta, \quad (\text{A5})$$

where the fields  $\{h_i^\alpha\}$  and couplings  $\{J_{ij}^{\alpha\beta}\}$  are adjusted to match the means  $\{m_i^\alpha\}$  and joint probabilities  $\{C_{ij}^{\alpha\beta}\}$ . The subscript in  $E_p$  in Eqs. (A4) and (A5) reminds us that these are Potts-like models for proteins.

Note that in this formulation the amino acid sequences are analogous to the patterns of spiking and silence in a network of neurons. The idea that proteins form families might correspond to these patterns forming a small number of globally structured clusters or brain states. Synthesizing proteins with sequences drawn from a model distribution would correspond to imposing patterns of activity onto the network, which still is a bit beyond the reach of today's experiments (though perhaps not for long).

One of the first modern applications of the maximum entropy approach was to families formed by pairs of interacting proteins that convey signals across the membrane of bacterial cells ([Weigt et al., 2009](#)). The crucial observation was that if one estimates the strength of correlation between amino acid choices at different sites, then this is only weakly

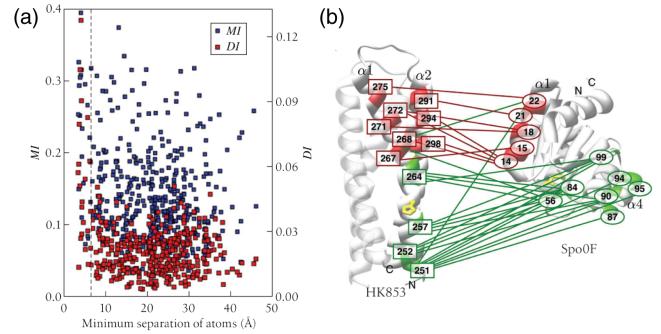


FIG. 39. Correlations vs effective interactions in protein sequence and structure for pairs of bacterial signaling proteins. (a) Mutual information (MI) between amino acid substitutions at pairs of sites  $i$  and  $j$  compared with direct information (DI) between these sites, calculated by allowing for the interaction  $J_{ij}$  in Eq. (A5) but eliminating the interactions with all other sites. (b) Lines connect sites above some threshold level of MI. Pairs in red share large DI and are in contact. Pairs in green have smaller DI, suggesting that correlations are indirect, and correspondingly they are not in contact. Adapted from [Bialek, 2012](#).

correlated with the distance between these sites in the three-dimensional structure; see Fig. 39(a). But we can imagine turning off all the interactions except those between sites  $i$  and  $j$  and then recomputing the mutual information. This “direct information” is strongly correlated with the distance between the sites, and a simple threshold allows us to identify the sites that are in contact at the interface between the two proteins; see Fig. 39(b). Subsequent work showed that this same principle also could be used to identify the correct interacting pairs of proteins ([Bitbol et al., 2016](#)).

The lesson of Fig. 39 is that, as in many statistical mechanics problems, spatially extended correlations can arise from much more local interactions. In this case the interactions are not real microscopic physical interactions, but instead effective interactions that describe the basic dependencies of amino acid substitutions on one another. This picture was presented quite clearly well before there were large enough datasets to make inference practical ([Lapedes et al., 1998](#)), but this seems to have been lost in conference proceedings that were not widely cited ([Lapedes, Giraud, and Jarzynski, 2012](#)). We note that large gaps between the range of interactions and the range of correlations, as seen here, are not generic.

If the statistics of amino acid substitutions in protein families are described via a set of spatially local effective interactions, then we should be able to predict the three-dimensional structure of these molecules from the sequence families alone. Remarkably, this works ([Marks et al., 2011](#); [Sulkowska et al., 2012](#)). These successes provided a foundation for the dramatic development of AlphaFold, a deep network that achieves unprecedented accuracy in structure prediction ([Jumper et al., 2021](#)).

The emphasis on structure prediction perhaps detracted from some of the more basic questions about the use of pairwise models. One exciting idea is that the effective energy function in Eq. (A5) might actually be related to the physical stability of the folded state. More ambitiously, if we build models for sequence variations across large populations of

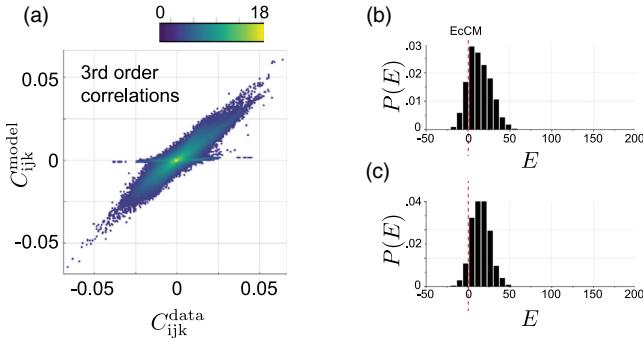


FIG. 40. Maximum entropy models for the ensemble of sequences in the AroQ family of chorismate mutases, enzymes involved in the synthesis of amino acids. (a) Probability density of triplet correlations predicted by the model given the correlation observed in the data. (b) The distribution of energies, from Eq. (48), across the sequences found in the data. The red dashed line marks the energy of a particular enzyme found in *E. coli*. (c) Distribution of energies predicted by the maximum entropy model cooled to a temperature  $T = 0.66$ . From Russ *et al.*, 2020.

viruses such as HIV, the energy might predict the fitness of different sequences in the environment provided by the patient's immune system (Chakraborty and Barton, 2017).

As with the discussion of patterns of activity in networks of neurons, we want to know whether the pairwise maximum entropy models correctly capture higher-order correlations across sequence variations. The first effort in this direction was focused on short, highly variable sequences in antibody molecules (Mora *et al.*, 2010). This was perhaps more influential as an introduction to the idea that one could use modern sequencing data to describe the full distribution of antibody diversity, fitting into a larger stream of work on physics problems motivated by immunology (Altan-Bonnet, Mora, and Walczak, 2020).

More recent work has shown that pairwise models can capture the three-point correlations among amino acid substitutions for a family of more than 1000 sequences for an enzyme involved in the synthesis of amino acids, as shown in Fig. 40(a) (Russ *et al.*, 2020). We emphasize that this test of the model is completely analogous to the tests in networks of neurons shown in Figs. 9 and 20. New sequences drawn from the effective Boltzmann distribution again are functional, and this is enhanced by “cooling” the distribution to lower temperature, which is consistent with the idea that the effective energy is a surrogate for functional behavior. Interestingly, the distribution of effective energies seen in the data [see Fig. 40(b)] is closer to the distribution seen at lower temperatures [see Fig. 40(c)] than at  $T = 1$ , where the model was learned. Even at these lower temperatures, the model generates large numbers of distinct functional sequences, thus providing input for the further design of new proteins.

Note that pairwise models for proteins are much more complicated than those for neurons. While neurons can be described well by binary variables (active or inactive, spiking or silent), each site along the amino acid chain has 20 possible amino acids, so there are  $\sim 200$  elements of the matrix  $J_{ij}^{\alpha\beta}$  for each pair of sites  $ij$ . At the same time, it is not easy to find

families with a number of sequences much larger than the typical number of samples in a neural recording lasting tens of minutes. More subtly, there is a correlation structure in these samples imposed by human choices to sequence certain organisms, or even particular proteins, and not others. Thus, the literature on maximum entropy models for sequence families involves much more discussion of sampling problems (Cocco, Monasson, and Sessak, 2011; Cocco, Monasson, and Weigt, 2013a, 2013b).

Perhaps the most fundamental prediction of maximum entropy models is the entropy of the sequence family itself (Barton *et al.*, 2016). In addition to providing a practical guide on how many sequences will fold into structures close to a target, the entropy gives us a sense for how to locate at these parts of living systems on a continuum from the generic to the particular. At one extreme we might have imagined that the interactions among amino acids are so complex that they might as well be random,<sup>24</sup> but this is incorrect because random sequences typically do not fold into compact or functional proteins. At the opposite extreme, we might have imagined that every detail of the sequence matters, but this is incorrect because proteins can tolerate many amino acid substitutions and remain functional.

The explicit construction of the probability distribution for sequences in a family provides a nuanced and quantitative response to these extreme views: proteins are not generic heteropolymers, but functionality persists in an ensemble of sequences that has substantial entropy rather than being confined to particular points in sequence space. This emphasis on the entropy of sequences associated with a single family and hence a particular structure also connects to much earlier work on global features of the sequence-structure mapping and the “designability” of structures (Li *et al.*, 1996). Although not usually phrased in this language, widely used descriptions of the variation in DNA sequences at protein binding sites can also be seen as maximum entropy models (von Hippel and Berg, 1986). Proteins with repeated segments provide an especially useful testing ground for these models (Espada *et al.*, 2015; Marchi *et al.*, 2019; Galpern *et al.*, 2022). We emphasize that writing explicit and quantitative models for the distribution of sequences is much more ambitious than the conventional use of highly simplified but tractable probabilistic models as a guide to data analysis, as in much of bioinformatics (Durbin *et al.*, 1998).

## 2. Collective behavior

At the other extreme of length scales is the use of statistical physics concepts to describe the behavior of animal groups, such as flocks of birds, schools of fish, and swarms of insects. The qualitative phenomenology of flocks, schools, and swarms is familiar. These collective behaviors are dramatic and have long been interesting to biologists because they provide a testing ground for ideas about the evolution of

<sup>24</sup>The idea that “complex equals random” was especially popular in the years immediately after the solution of the mean-field spin glass, which gave us many new tools for analyzing such random systems (Mézard, Parisi, and Virasoro, 1987).

cooperation. In the mid-1990s, there were efforts to write dynamical models for populations of self-propelled particles that could control their motion in relation to that of their neighbors (Vicsek *et al.*, 1995).<sup>25</sup> This work immediately caught the attention of the physics community, in part because these models exhibited directional ordering—the emergence of a well-defined direction of motion for all the “birds” in the flock—even in two dimensions, where this is forbidden for equilibrium systems.

The simulations of self-propelled particles have the flavor of a molecular dynamics simulation, but with microscopic entities that expend energy to keep moving on their own and with “social” rather than Newtonian forces. A large step forward was to ask whether there is a more macroscopic fluid mechanics that emerges as we coarse grain these molecular (ish) dynamics. More abstractly, what is the effective field theory that describes the long-distance, long-time behavior of a large collection of such self-propelled particles? The answer to this question (Toner and Tu, 1995, 1998) laid the foundations for the field of active matter (Marchetti *et al.*, 2013). As with the statistical mechanics of neural networks, this field now has a life independent of its origins as an effort to understand real flocks and swarms.

The early models and their field-theoretic development captured the qualitative phenomenon of flocking. As in other statistical physics problems, plausible local interactions lead to global ordering and the ordered state is separated from a disordered state by a phase transition. But these theories make quantitative predictions, and at the time there were essentially no large-scale, quantitative data with which to test these predictions.<sup>26</sup>

New analysis methods made it possible to reconstruct the three-dimensional positions of every individual in large, naturally occurring animal collectives, first in flocks of thousands of birds as in Figs. 41(a) and 41(b) (Ballerini *et al.*, 2008a; Cavagna, Giardina, Orlandi, Parisi, and Procaccini, 2008; Cavagna, Giardina, Orlandi, Parisi, Procaccini *et al.*, 2008) and then in swarms of hundreds of insects (Attanasi *et al.*, 2014b). Beyond new methods for image analysis, this work brought the conceptual framework of statistical physics to bear on the analysis of correlations in these animal groups (Cavagna, Giardina, and Grigera, 2018). It is particularly noteworthy that these analyses revealed precise and reproducible behaviors of animal groups out in the wild rather than in the laboratory.

Flocks of starlings are highly polarized, but there are measurable fluctuations around this mean velocity. A

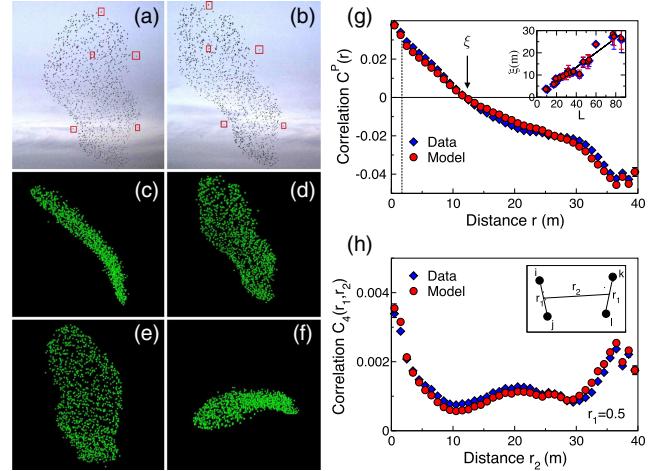


FIG. 41. Positions, velocities, and correlations in a flock of birds. (a),(b) A flock with  $N = 1246$  starlings captured at a single moment in time by two cameras separated by 25 m. The red boxes highlight corresponding birds in the two images and are used for reconstructing positions in three dimensions. (c)–(f) Three-dimensional reconstruction of the flock under four different points of view. (g) Two-point correlations of directional fluctuations,  $C^P(r)$  from Eq. (A8), comparing predictions of the maximum entropy model (red circles) to measurements on the real flock (blue diamonds). The typical radius of the neighborhoods  $\mathcal{N}_i$  is shown as a vertical dashed line. If we define the correlation length through  $C^P(\xi) = 0$ , then  $\xi$  is proportional to the linear size  $L$  of the flock, and this also is captured by the maximum entropy models, as shown in the inset. (h) Four-point correlations  $C_4(r_1, r_2)$  from Eq. (A9), with distances defined in the inset, again comparing theory (red circles) and experiment (blue diamonds). (a)–(f) From Ballerini *et al.*, 2008b. (g),(h) From Bialek *et al.*, 2012.

comparison of flocks of different sizes and densities demonstrates that correlations among these fluctuations depend not on the physical or metric distance between birds but on the ranking of neighbors, called the topological distance (Ballerini *et al.*, 2008b). Correlations between fluctuations in both direction and speed have a scale-invariant form, with no characteristic length scale other than the linear dimensions of the flock itself (Cavagna *et al.*, 2010). Swarms of midges are not polarized (the mean velocity is zero), but one can still see correlations in the velocity fluctuations, and these are again scale invariant (Attanasi *et al.*, 2014c). Analysis of events where flocks turn shows that information propagates ballistically rather than diffusively (Attanasi *et al.*, 2014a), and in swarms one can see dynamic scaling of the fluctuations with an exponent  $z = 1.37 \pm 0.11$ , far below the diffusive  $z = 2$  (Cavagna *et al.*, 2017). None of these quantitative results agree with predictions from the original models of self-propelled particles.

Some features of the correlation structure in flocks can be captured by surprisingly simple maximum entropy models (Bialek *et al.*, 2012, 2014). We start by writing the velocity of each bird  $i$  as  $\vec{v}_i = v_i \hat{s}_i$ , where  $\hat{s}_i$  is a unit vector pointing in the flight direction and  $v_i$  is the flight speed. As a first step, we focus on the flight directions and ignore the fluctuations in speed. We expect individual birds to orient relative to the

<sup>25</sup>Although the questions addressed were quite different, there was prior work in the computer science community on a model of “boids” (Reynolds, 1987). This in turn had precursors in the biological literature (Aoki, 1982).

<sup>26</sup>There were fascinating early efforts to characterize small schools of fish in the laboratory (Cullen, Shaw, and Baldwin, 1965). The state of the art circa 2000 was reviewed in edited volumes (Parrish and Hammer, 1997; Camazine *et al.*, 2001; Krause and Ruxton, 2002). Note that there was considerable progress in quantifying more complex collective behaviors, such as nest building by social insects, a subject that is as yet largely untouched by statistical physics methods.

average of their near neighbors, and we can measure the strength of this effect through the correlation

$$C_{\text{local}} = \frac{1}{N} \sum_{i=1}^N \hat{s}_i \cdot \left( \frac{1}{n_c} \sum_{j \in \mathcal{N}_i} \hat{s}_j \right), \quad (\text{A6})$$

where  $\mathcal{N}_i$  denotes the neighborhood of bird  $i$ , and from the observations on the topological character of the correlations, we take this neighborhood to include the  $n_c$  nearest neighbors. We treat all birds as equivalent,<sup>27</sup> so  $C_{\text{local}}$  is defined as an average over the flock. Given a measured  $\langle C_{\text{local}} \rangle_{\text{expt}}$ , the maximum entropy distribution for all flight directions in the flock is

$$P(\{\hat{s}_i\}) = \frac{1}{Z(J)} \exp \left[ J \sum_{i=1}^N \hat{s}_i \cdot \left( \frac{1}{n_c} \sum_{j \in \mathcal{N}_i} \hat{s}_j \right) \right], \quad (\text{A7})$$

where the value of  $J$  has to be adjusted to match  $\langle C_{\text{local}} \rangle$ . Because real flocks are highly polarized, all the relevant calculations can be done in a spin-wave approximation, and we can verify at the end that the inferred value of  $J$  is consistent with the validity of this approximation and with the measured polarization. Finally, we can find the best neighborhood size  $n_c$  by the maximum likelihood. Note that once we have chosen  $J$  to match the observed  $\langle C_{\text{local}} \rangle$ , all other quantities are predicted with no free parameters.

We can further decompose the unit vector  $\hat{s}$  into a longitudinal component along the mean velocity of the flock and a two-dimensional component  $\hat{\pi}$  perpendicular to the mean. Then there is a natural two-point correlation function

$$C^P(r) = \langle \hat{\pi}_i \cdot \hat{\pi}_j \rangle_{r_{ij}=r}, \quad (\text{A8})$$

where the average is over all pairs of birds separated by a distance  $r$ . Note that since the density of a flock is relatively uniform at a single moment in time, there is little difference between topological and metric distance in a single snapshot. Figure 41(g) compares this correlation function with the prediction from the maximum entropy model in Eq. (A7), and we see that the agreement is excellent from the scale of the neighborhoods  $\mathcal{N}_i$  out to the size of the flock as a whole. The behavior is featureless, suggesting that there is no characteristic scale. If we define a correlation length  $\xi$  as the distance at which  $C^P(r)$  changes sign, then  $\xi$  is proportional to the size of the flock, confirming the scale invariance, and this is correctly predicted by the maximum entropy models; see the inset of Fig. 41(g). We can go even further and estimate a four-point function

$$C_4(r_1, r_2) = \langle (\hat{\pi}_i \cdot \hat{\pi}_j)(\hat{\pi}_k \cdot \hat{\pi}_l) \rangle, \quad (\text{A9})$$

where the average is over four birds with relative positions shown in the inset of Fig. 41(h). Theory and experiment again agree very well, even though these effects are quite small.

The maximum entropy model in Eq. (A7) is equivalent to a equilibrium Heisenberg model with local interactions. Thus, when  $J$  is large enough to generate an ordered flock, scale-invariant fluctuations are a consequence of Goldstone's theorem. But this theorem does not guarantee *quantitative* agreement with the data, as observed. Instead of matching the average correlation of birds with their nearest  $n_c$  neighbors as in Eq. (A6), we can try matching the correlation with the nearest neighbor, the second nearest neighbor, etc. (Cavagna *et al.*, 2015). Each time that we add a constraint on the  $n$ th neighbor, we introduce a coupling  $J(n)$  into a generalization of Eq. (A7), leading to

$$P(\{\hat{s}_i\}) = \frac{1}{Z(J)} \exp \left[ \sum_{i=1}^N \sum_{j=1}^N J(k_{ij}) \hat{s}_i \cdot \hat{s}_j \right], \quad (\text{A10})$$

where bird  $j$  is the  $(k_{ij})$ th neighbor of bird  $i$ . The result of this exercise is that  $J(n) \sim \exp(-n/n_0)$ , with a range  $n_0 \sim 6$ . Thus, even if we try to match the longer-distance correlations explicitly, the structure of these correlations drives the model toward short-range effective interactions. These few short-range terms then are sufficient to quantitatively predict the observed longer-range and higher-order correlations, as in Figs. 41(g) and 41(h).

The equivalence to an equilibrium model might be surprising. The essence of the original models was that active systems generate behaviors that are not accessible in equilibrium, such as the breaking of a continuous symmetry in two dimensions. Alternatively, in the active system, the dynamics generates long-range effective interactions in the steady-state distribution. We now see explicitly that these effects are minimal in real flocks, which we can understand because the timescales for individual birds to align with their neighbors are shorter than the timescales for neighbors to exchange places, leading to a local equilibrium (Mora *et al.*, 2016). Thus, in the case of flocks, we not only can see that the simplest maximum entropy models work but also can verify explicitly through testing that more complex models are not needed: the extra effective interactions are driven to zero by the data.

In flocks one sees scale-invariant fluctuations not only in flight direction but also in flight speed (Cavagna *et al.*, 2010). In this case there is no Goldstone theorem to help us understand the origin of this behavior. If we try to build maximum entropy models that match the strength of local correlations as before, the parameters of these models are driven close to a point where the correlation length diverges, and predictions match the observed long-range correlations (Bialek *et al.*, 2014). If we restrict ourselves to local models, then there is a much more general argument that the effective potential that holds the speed of individual birds near the mean must be very “soft” near the minimum (Cavagna *et al.*, 2022). The maximum entropy approach thus suggests, strongly, that real flocks tune themselves to some nongeneric point in their parameter space. Swarms also seem to be poised at a special point in parameter space, although disordered (Attanasi *et al.*, 2014c; Cavagna *et al.*, 2017). There is as yet no maximum

<sup>27</sup>More precisely, we treat all birds in the interior of the flock as equivalent. The birds at the surface are special because all their neighbors are to one side of them. In what follows we take velocities of birds on the boundary of the flock as given and study the response of the bulk to this boundary condition.

entropy model for swarms, but there have been sophisticated renormalization group calculations to predict the observed dynamic scaling exponent (Cavagna *et al.*, 2019, 2023).

Flocks and swarms provide a useful touchstone for thinking about networks of neurons. In connecting theory to experiment, networks of neurons have the advantage that datasets are larger. But in animal groups the interactions are plausibly local, and it seems reasonable to treat all individuals as equivalent. Both of these considerations drive us toward a simpler set of constraints for the construction of maximum entropy models. As with neurons, there are a number of good reasons why these models of flocks might not have worked. The detailed, quantitative successes thus encourage us to think that statistical physics approaches can provide theories of real living systems, not simply metaphors that capture qualitative behaviors.

### 3. Ecology and metabolism

Maximum entropy methods have been used in ecology for many years, often in simple form, searching for models that match the mean abundances of species or their energetic load on the environment (Harte *et al.*, 2008; Banavar, Maritan, and Volkov, 2010; Harte and Newman, 2014). An important feature of these applications is that the chosen constraints are sums over contributions from each species, so the resulting models are noninteracting. In the context of neural networks, we have emphasized that maximum entropy provides an alternative path to connecting with statistical physics, not making assumptions about the underlying dynamics but instead pointing to particular experimental facts that we insist our models must match. More recent work in ecology has taken this point of view even further, noting that simplifying mechanistic hypotheses suggest particular quantities as the ones that we should use to constrain the maximum entropy construction (O'Dwyer, Rominger, and Xiao, 2017).

The project of building models for the distribution of species abundances has not yet felt the impact of dramatic improvements in the ability to measure the abundances of hundreds of species in microbial ecologies. The most famous examples are from the bacterial communities that inhabit humans and influence our health, but there are precise measurements in marine environments (Ward *et al.*, 2017), in hot springs (Rosen *et al.*, 2015; Birzu *et al.*, 2023), in soil (Lee *et al.*, 2024), and on synthetic communities constructed in the laboratory (Cheng *et al.*, 2022). In a different direction, classical ecological surveys, for example, of trees and shrubs in a small forested region observed over several years (Condit *et al.*, 2014), can be analyzed with ideas from statistical physics to discover unexpected structures (Villegas *et al.*, 2021, 2024).

As a final example, we consider maximum entropy approaches to cellular metabolism. There is a long tradition of abstracting from the frighteningly complex map of all the interlocking biochemical reactions to a stoichiometric matrix  $S_{i\mu}$  that connects the flux through the reaction  $i$  to the change in concentration of the molecule or metabolite  $\mu$ ,

$$\frac{dc_\mu}{dt} = \sum_i S_{i\mu} \nu_i. \quad (\text{A11})$$

If a bacterial cell is in a phase of steady-state growth, then  $dc_\mu/dt = 0$ , defining a null space for the set of fluxes  $\{\nu_i\}$ . For example, the core bacterial metabolism of  $N = 86$  reactions among  $M = 63$  metabolites leaves a space of  $D = 23$  in which the fluxes can vary; lower and upper bounds on the fluxes mean that this space is a convex polytope. It is sensible to take these fluxes as variables that can be controlled by the cell since each reaction is catalyzed by an enzyme whose expression level and activity can be regulated. To reproduce, the cell must make a copy of itself, which requires the synthesis of a particular combination of metabolites. Plausibly, then, the growth rate is

$$\lambda(\{\nu_i\}) = \sum_{i=1}^N \xi_i \nu_i \quad (\text{A12})$$

and the coefficients  $\xi_i$  are known, at least approximately.

Being in steady state means that fluxes balance, and this “flux balance analysis” (Orth, Thiele, and Palsson, 2010) often is supplemented by the hypothesis that fluxes are adjusted to maximize the growth rate (Ibarra, Edwards, and Palsson, 2002). This is an extreme hypothesis and would require infinite information to tune each flux to its optimal value. An alternative is to ask for the ensemble of fluxes that achieves an observed mean growth rate but otherwise is as random (minimally tuned) as possible; this is the maximum entropy distribution

$$P(\{\nu_i\}) = \frac{1}{Z(\beta)} \exp[-\beta \lambda(\{\nu_i\})]. \quad (\text{A13})$$

As  $\beta \rightarrow 0$ , we have a completely unregulated system, allowing fluxes to vary uniformly over all allowed values. As we increase  $\beta$ , the entropy in this space of fluxes is reduced and the mean growth rate increases, ultimately converging on the optimal growth rate  $\lambda_{\max}$  as  $\beta \rightarrow \infty$ . Another way of saying this is that achieving a certain mean growth rate requires specifying a certain amount of information about the fluxes relative to the unregulated, uniform distribution. This trade-off, which is illustrated in Fig. 42(a), is an example of rate-distortion theory (Cover and Thomas, 1991).<sup>28</sup>

The maximum entropy model in Eq. (A13) has one parameter  $\beta$  that must be set to match the average growth rate. The model then predicts the mean fluxes for all the individual reactions, many of which can be measured. Under conditions where *E. coli* achieve  $\sim 80\%$  of their maximal growth rate, theory and experiment agree within error bars for 20 independently measured reaction fluxes. Reaching this growth rate requires  $\sim 40$  bits of information about the fluxes, so the cell must have roughly two bits of bandwidth for controlling each degree of freedom in the metabolic network. The maximum entropy model predicts that fluxes and even the growth rate itself should be variable. Measurements on the lineages of individual cells grown in the presence of low doses of antibiotics make it possible to confirm the predicted dependence of the standard deviation in growth rate on the mean, as shown in Fig. 42(b). Richer behaviors are possible in

<sup>28</sup>See also the discussion in Sec. 6.3 of Bialek (2012).

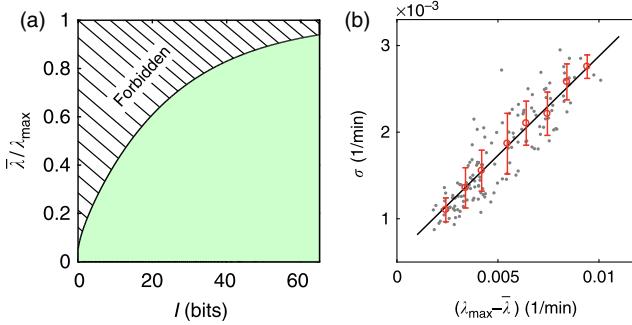


FIG. 42. Maximum entropy models for metabolism. (a) Achieving a particular growth rate  $\bar{\lambda}$  requires the entropy of the joint distribution of fluxes to be reduced by at least  $I$  bits below the entropy of the uniform distribution (green region). Points in the hashed (forbidden) region are not achievable. (b) The scaling of the standard deviation in growth rates with the growth rate itself. Each gray point is measured from a single lineage of cells. The red points with errors are the mean and standard deviations in equally spaced bins. The maximum entropy model predicts a linear relation, as shown by the solid line. From De Martino *et al.*, 2018.

models that address the spatial structure of the metabolic networks (Narayananakutty *et al.*, 2024).

#### 4. Coda

In summary, maximum entropy methods have proved to be productive in describing emergent behaviors of biological systems on all scales, from protein molecules to ecology and from bacterial metabolism to flocks of birds.<sup>29</sup> As with networks of neurons, the key idea is that these methods connect general statistical physics models directly to experimental data on particular systems, resulting in detailed—and often successful—quantitative predictions. This emphasizes once more that statistical physics descriptions of living systems should not be just metaphorical. Different systems are in different regimes with respect to dataset size and the complexity of the simplest plausible models, giving us the opportunity to test our algorithmic tools more extensively. In each case we learn something about the particular system, but we also see common themes. Notably, the parameters of the maximum entropy models that match basic facts about these systems seem to be nongeneric.

#### APPENDIX B: INFERENCE

In small systems we can perform an “exact” maximum entropy construction, but once  $N$  is large we need approximate

<sup>29</sup>Another natural target for this analysis is the covariation of gene expression levels in cells. Early work used measurements averaged over many cells, but with variations across time in response to perturbations (Lezon *et al.*, 2006). Dramatic developments in experimental technique now make it possible to literally count almost every molecule of messenger RNA in single cells, labeled by the gene from which it was transcribed. Recent work built maximum entropy models to describe these data (Sarra *et al.*, 2024; Skinner, Lamine, and Mani, 2024).

numerical methods for solving the inverse problem. To understand the general strategy, it is useful to place the maximum entropy models into context.

From a physics point of view, the maximum entropy models are special because they are the solution to a variational problem, constrained by experimental observations. But one could also take the view that they constitute an interesting family of models, and we want to fit them to the data. Assume that we have made  $M$  independent observations of the state  $\vec{\sigma}$ , which we index as  $\vec{\sigma}^{(n)}$ , with  $n = 1, 2, \dots, M$ . If our model of the probability distribution is, as in Eqs. (20) and (21),

$$P(\vec{\sigma}|\{\lambda_\mu\}) = \frac{1}{Z} \exp \left[ - \sum_{\mu=1}^K \lambda_\mu f_\mu(\vec{\sigma}) \right], \quad (\text{B1})$$

where we note explicitly the dependence on the parameters, then the probability or likelihood of observing the data is

$$P(\{\vec{\sigma}^{(n)}\}|\{\lambda_\mu\}) = \frac{1}{Z^M} \exp \left[ - \sum_{\mu=1}^K \lambda_\mu \sum_{n=1}^M f_\mu(\vec{\sigma}^{(n)}) \right]. \quad (\text{B2})$$

A conventional strategy for estimating the parameters  $\{\lambda_\mu\}$  is maximum likelihood, optimizing the probability that our model will generate the observed data.<sup>30</sup> To do this, we differentiate the log probability with respect to each  $\lambda_\mu$ , being careful that the partition function  $Z$  depends on these parameters,

$$\frac{1}{M} \frac{\partial \ln P(\{\vec{\sigma}^{(n)}\}|\{\lambda_\mu\})}{\partial \lambda_\mu} = -\frac{\partial \ln Z}{\partial \lambda_\mu} - \frac{1}{M} \sum_{n=1}^N f_\mu(\vec{\sigma}^{(n)}). \quad (\text{B3})$$

We have the usual identities from statistical mechanics,

$$\frac{\partial \ln Z}{\partial \lambda_\mu} = -\langle f_\mu(\vec{\sigma}) \rangle_P, \quad (\text{B4})$$

and we recognize the average over experimental data,

$$\frac{1}{M} \sum_{n=1}^N f_\mu(\vec{\sigma}^{(n)}) = -\langle f_\mu(\vec{\sigma}) \rangle_{\text{expt}}. \quad (\text{B5})$$

Thus, we have

$$\frac{1}{M} \frac{\partial \ln P(\{\vec{\sigma}^{(n)}\}|\{\lambda_\mu\})}{\partial \lambda_\mu} = -[\langle f_\mu(\vec{\sigma}) \rangle_{P_\lambda} - \langle f_\mu(\vec{\sigma}) \rangle_{\text{expt}}]. \quad (\text{B6})$$

This derivative vanishes and the likelihood is maximized when we satisfy the constraints in Eq. (17), matching the predicted and observed expectation values of the observable on which we choose to focus.

Equation (B6) shows us that the likelihood of the data is maximized when the constraints are satisfied, but it tells us

<sup>30</sup>We can also think of this as finding the model that allows us to construct the shortest code for the data (Cover and Thomas, 1991; Mézard and Montanari, 2009; Bialek, 2012).

more: if we adjust each  $\lambda_\mu$  in proportion to the difference between the theoretical and experimental expectation values, then we are climbing the gradient in likelihood toward the point where the constraints are satisfied. This suggests the following algorithm for learning the parameters  $\{\lambda_\mu\}$  or, equivalently, for solving the constraint in Eq. (17):

- (1) Choose some set of parameters  $\{\lambda_\mu\}$ .
- (2) Perform a Monte Carlo simulation to generate samples from the distribution  $P(\vec{\sigma}|\{\lambda_\mu\})$ .
- (3) From these samples estimate the expectation values  $\langle f_\mu(\vec{\sigma}) \rangle_{P_\lambda}$ .
- (4) Update the parameters

$$\lambda_\mu \rightarrow \lambda_\mu - \eta [\langle f_\mu(\vec{\sigma}) \rangle_{P_\lambda} - \langle f_\mu(\vec{\sigma}) \rangle_{\text{expt}}], \quad (\text{B7})$$

where  $\eta$  is a small “learning rate.”

- (5) Return to (2) or proceed to (6).
- (6) End when constraints are satisfied within the error bars on the experimental estimates of the expectation values.

This approach has a long history, dating back to at least the work of Ackley, Hinton, and Sejnowski (1985). Once the maximum entropy models for neurons were introduced, these tools were pushed quickly from  $N = 10$  up to  $N = 40$  neurons in the retina (Tkačik *et al.*, 2006, 2009), and they continue to be at the core of most applications of the maximum entropy idea.

The brute force Monte Carlo methods can be improved. One idea is to add some “inertia” to the updating of the parameters in Eq. (B7) or to allow the learning rate  $\eta$  to slow with time as the algorithm gets closer to the final answer, as in simulated annealing (Kirkpatrick, Gelatt, and Vecchi, 1983). More fundamentally, when the parameters change by only a small amount, one might be able to reuse the same Monte Carlo samples with new weights (Broderick *et al.*, 2007), as in histogram Monte Carlo simulations (Ferrenberg and Swendsen, 1988), thus increasing efficiency. There is also some artistry involved in choosing the length of the Monte Carlo simulations to balance errors in estimating expectation values versus the efficiency of moving through parameter space, and again some sort of annealing can be useful. Many of these and other issues were described by Lee and Daniels (2019), who also provided PYTHON code.

As noted, the construction of maximum entropy models is the inverse of the usual statistical mechanics problem: rather than being given the coupling constants and being asked to compute expectation values, we are given the expectation values and asked to estimate the coupling constants. A related problem is to find the coupling constants at the fixed points of the RG using Monte Carlo methods (Swendsen, 1984). The reappearance of this problem in the context of neural data analysis has led to new algorithms and the exploration of different approximations.

If we focus on one neuron, we can write the probability that it is active as a function of the state of all the other neurons; see also Eq. (B8). In the purely pairwise models, Eq. (35) with  $\sigma_i = \{0, 1\}$ , this is

$$P(\sigma_i = 1 | \{\sigma_{i \neq j}\}) = \frac{1}{1 + \exp[-h_i^{\text{eff}}(\{\sigma_{i \neq j}\})]}, \quad (\text{B8})$$

where the effective field

$$h_i^{\text{eff}}(\{\sigma_{i \neq j}\}) = h_i + \sum_j J_{ij} \sigma_j. \quad (\text{B9})$$

We also have

$$P(\sigma_i = 0 | \{\sigma_{i \neq j}\}) = \frac{1}{1 + \exp[+h_i^{\text{eff}}(\{\sigma_{i \neq j}\})]}. \quad (\text{B10})$$

We can fit these expressions to the data in the usual way and thus determine one row of the  $J_{ij}$  matrix without confronting the real difficulties of the underlying statistical mechanics problem. For this one cell, the fitting problem has become a form of regression. In the “pseudolikelihood” method, we pretend that the fitting for each neuron is independent of all the others, so the log probability of the data is the sum of the terms from individual cells (Aurell and Ekeberg, 2012). There are interesting connections between this method and the Monte Carlo RG (Albert and Swendsen, 2014).

Since the maximum entropy construction can be done exactly at  $J_{ij} = 0$ , it is natural to ask how far we can get with perturbation theory, perhaps suitably resummed (Sessak and Monasson, 2009). Perturbation theory is interesting both because it may provide a path to solving the inverse problem and because it can give us a sense for the strength of correlations (Azhar and Bialek, 2010). An alternative to perturbation theory is a cluster expansion, instantiating the intuition that even in a large network interactions may be strongest among more limited groups of neurons (Cocco and Monasson, 2011, 2012). For a review of these and other methods, see Nguyen, Zecchina, and Berg (2017).

## REFERENCES

- Abdelfattah, A. S., *et al.*, 2019, *Science* **365**, 699.
- Ackley, D. H., G. E. Hinton, and T. J. Sejnowski, 1985, *Cognit. Sci.* **9**, 147.
- Adrian, E. D., 1928, *The Basis of Sensation: The Action of the Sense Organs* (W. W. Norton, New York).
- Adrian, E. D., and B. H. C. Matthews, 1934, *Brain* **57**, 355.
- Ahrens, M. B., M. B. Orger, D. N. Robson, J. M. Li, and P. J. Keller, 2013, *Nat. Methods* **10**, 413.
- Aidley, D. J., 1998, *The Physiology of Excitable Cells*, 4th ed. (Cambridge University Press, Cambridge, England).
- Aitchison, L., N. Corradi, and P. E. Latham, 2016, *PLoS Comput. Biol.* **12**, e1005110.
- Aksay, E., G. Gamkrelidze, H. S. Seung, R. Baker, and D. W. Tank, 2001, *Nat. Neurosci.* **4**, 184.
- Albert, J., and R. H. Swendsen, 2014, *Phys. Procedia* **57**, 99.
- Alhashim, M. G., K. Hausknecht, and M. P. Brenner, 2025, *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2403644122.
- Altan-Bonnet, G., T. Mora, and A. M. Walczak, 2020, *Phys. Rep.* **849**, 1.
- Amit, D. J., 1989, *Modeling Brain Function: The World of Attractor Neural Networks* (Cambridge University Press, Cambridge, England).

- Amit, D. J., H. Gutfreund, and H. Sompolinsky, 1985, *Phys. Rev. A* **32**, 1007.
- Amit, D. J., H. Gutfreund, and H. Sompolinsky, 1987, *Ann. Phys. (N.Y.)* **173**, 30.
- Anderson, P. W., 1984, *Basic Notions of Condensed Matter Physics* (Benjamin Cummings, Menlo Park, CA).
- Aoki, I., 1982, *Nippon Suisan Gakkaishi* **48**, 1081.
- Attanasi, A., *et al.*, 2014a, *Nat. Phys.* **10**, 691.
- Attanasi, A., *et al.*, 2014b, *PLoS Comput. Biol.* **10**, e1003697.
- Attanasi, A., *et al.*, 2014c, *Phys. Rev. Lett.* **113**, 238102.
- Aurell, E., and M. Ekeberg, 2012, *Phys. Rev. Lett.* **108**, 090201.
- Azhar, F., and W. Bialek, 2010, [arXiv:1012.5987](#).
- Bak, P., 1996, *How Nature Works: The Science of Self-Organized Criticality* (Copernicus, New York).
- Bak, P., C. Tang, and K. Wiesenfeld, 1987, *Phys. Rev. Lett.* **59**, 381.
- Ballerini, M., *et al.*, 2008a, *Anim. Behav.* **76**, 201.
- Ballerini, M., *et al.*, 2008b, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 1232.
- Banavar, J. R., A. Maritan, and I. Volkov, 2010, *J. Phys. Condens. Matter* **22**, 063101.
- Barson, D., A. S. Hamodi, X. Shen, G. Lur, R. T. Constable, J. A. Cardin, M. C. Crair, and M. J. Higley, 2020, *Nat. Methods* **17**, 107.
- Barton, J. P., A. K. Chakraborty, S. Cocco, H. Jacquin, and R. Monasson, 2016, *J. Stat. Phys.* **162**, 1267.
- Beggs, J. M., and D. Plenz, 2003, *J. Neurosci. Methods* **23**, 11167.
- Ben-Yishai, R., R. L. Bar-Or, and H. Sompolinsky, 1995, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 3844.
- Berman, G. J., W. Bialek, and J. W. Shaevitz, 2016, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 11943.
- Berman, G. J., D. M. Choi, W. Bialek, and J. W. Shaevitz, 2014, *J. R. Soc. Interface* **11**, 20140672.
- Bertschinger, N., and T. Natschläger, 2004, *Neural Comput.* **16**, 1413.
- Bialek, W., 2012, *Biophysics: Searching for Principles* (Princeton University Press, Princeton, NJ).
- Bialek, W., 2024, in *Les Houches Summer School Lecture Notes: Theoretical Biological Physics 2023*, SciPost Physics Lecture Notes Vol. 84, edited by A.-F. Bitbol, T. Mora, I. Nemenman, and A. M. Walczak (SciPost Foundation, Amsterdam).
- Bialek, W., A. Cavagna, I. Giardina, T. Mora, O. Pohl, E. Silvestri, M. Viale, and A. M. Walczak, 2014, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 7212.
- Bialek, W., A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walczak, 2012, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 4786.
- Bialek, W., S. E. Palmer, and D. J. Schwab, 2020, [arXiv:2008.12279](#).
- Bialek, W., and R. Ranganathan, 2007, [arXiv:0712.4397](#).
- Bialek, W., and J. W. Shaevitz, 2024, *Phys. Rev. Lett.* **132**, 048401.
- Binder, K., 1981, *Z. Phys. B* **43**, 119.
- Birzu, G., S. H. Muralidharan, D. Goudeau, R. R. Malmstrom, D. S. Fisher, and D. Bhaya, 2023, *eLife* **12**, RP90849.
- Bitbol, A.-F., R. S. Dwyer, L. J. Colwell, and N. S. Wingreen, 2016, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 12180.
- Bliss, T. V. P., and T. Lømo, 1973, *J. Physiol. (London)* **232**, 331.
- Block, H., 1962, *Rev. Mod. Phys.* **34**, 123.
- Block, H., B. W. Knight, Jr., and F. Rosenblatt, 1962, *Rev. Mod. Phys.* **34**, 135.
- Bradde, S., and W. Bialek, 2017, *J. Stat. Phys.* **167**, 462.
- Braun, J., J. R. Abney, and J. C. Owicki, 1984, *Nature (London)* **310**, 316.
- Brenner, S., 1957, *Proc. Natl. Acad. Sci. U.S.A.* **43**, 687.
- Brinkman, B. A. W., 2025, *PRX Life* **3**, 043003.
- Broderick, T., M. Dudík, G. Tkačik, R. E. Schapire, and W. Bialek, 2007, [arXiv:0712.2437](#).
- Bustamante, C., J. F. Marko, E. D. Siggia, and S. Smith, 1994, *Science* **265**, 1599.
- Camazine, S., J.-L. Deneubourg, N. R. Franck, J. Sneyd, G. Theraulaz, and E. Bonabeau, 2001, Eds., *Self-Organization in Biological Systems* (Princeton University Press, Princeton, NJ).
- Carleo, G., I. Cirac, K. Cranmer, L. Daudet, M. Schulz, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, 2019, *Rev. Mod. Phys.* **91**, 045002.
- Carmena, J. M., M. A. Lebedev, R. E. Crist, J. E. O'Doherty, D. M. Santucci, D. F. Dimitrov, P. G. Patil, C. S. Henriquez, and M. A. L. Nicolelis, 2003, *PLoS Biol.* **1**, e42.
- Castellana, M., and W. Bialek, 2014, *Phys. Rev. Lett.* **113**, 117204.
- Castro, D. M., T. Feliciano, N. A. P. de Vasconcelos, C. Soares-Cunha, B. Coimbra, A. J. Rodrigues, P. V. Carelli, and M. Copelli, 2024, *PRX Life* **2**, 023008.
- Cavagna, A., A. Cimarelli, I. Giardina, G. Parisi, R. Santagati, F. Stefanini, and M. Viale, 2010, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 11865.
- Cavagna, A., D. Conte, C. Creato, L. Del Castello, I. Giardina, T. S. Grigera, S. Melillo, L. Parisi, and M. Viale, 2017, *Nat. Phys.* **13**, 914.
- Cavagna, A., A. Culla, X. Feng, I. Giardina, T. S. Grigera, W. Kion-Crosby, S. Melillo, G. Pisegna, L. Postiglione, and P. Villegas, 2022, *Nat. Commun.* **13**, 2315.
- Cavagna, A., L. Del Castillo, S. Dey, I. Giardina, S. Melillo, L. Parisi, and M. Viale, 2015, *Phys. Rev. E* **92**, 012705.
- Cavagna, A., L. Di Carlo, I. Giardina, L. Grandinetti, T. S. Grigera, and G. Pisegna, 2019, *Phys. Rev. Lett.* **123**, 268001.
- Cavagna, A., L. Di Carlo, I. Giardina, T. S. Grigera, S. Melillo, L. Parisi, G. Pisegna, and M. Scandolo, 2023, *Nat. Phys.* **19**, 1043.
- Cavagna, A., I. Giardina, F. Ginelli, T. Mora, D. Piovani, R. Tavarone, and A. M. Walczak, 2014, *Phys. Rev. E* **89**, 042707.
- Cavagna, A., I. Giardina, and T. A. Grigera, 2018, *Phys. Rep.* **728**, 1.
- Cavagna, A., I. Giardina, A. Orlandi, G. Parisi, and A. Procaccini, 2008, *Anim. Behav.* **76**, 237.
- Cavagna, A., I. Giardina, A. Orlandi, G. Parisi, A. Procaccini, M. Viale, and V. Zdravkovic, 2008, *Anim. Behav.* **76**, 217.
- Chakraborty, A. K., and J. P. Barton, 2017, *Rep. Prog. Phys.* **80**, 032601.
- Chalfie, M., Y. Tu, G. Euskirchen, W. W. Ward, and D. C. Prasher, 1994, *Science* **263**, 802.
- Chayes, J. T., L. Chayes, and E. H. Lieb, 1984, *Commun. Math. Phys.* **93**, 57.
- Chen, T.-W., *et al.*, 2013, *Nature (London)* **499**, 295.
- Chen, X., and R. Dzakpasu, 2010, *Phys. Rev. E* **82**, 031907.
- Chen, X., F. Randi, A. M. Leifer, and W. Bialek, 2019, *Phys. Rev. E* **99**, 052418.
- Chen, X., M. Winiarski, A. Puścian, E. Knapska, A. M. Walczak, and T. Mora, 2023, *Phys. Rev. X* **13**, 041053.
- Cheng, A. G., *et al.*, 2022, *Cell* **185**, 3617.
- Chung, J. E., J. F. Magland, A. H. Barnett, V. M. Tolosa, A. C. Tooker, K. Y. Lee, K. G. Shah, S. H. Felix, L. M. Frank, and L. F. Greengard, 2017, *Neuron* **95**, 1381.
- Chung, J. E., *et al.*, 2019, *Neuron* **101**, 21.
- Cocco, S., and R. Monasson, 2011, *Phys. Rev. Lett.* **106**, 090601.
- Cocco, S., and R. Monasson, 2012, *J. Stat. Phys.* **147**, 252.
- Cocco, S., R. Monasson, and V. Sessak, 2011, *Phys. Rev. E* **83**, 051123.
- Cocco, S., R. Monasson, and M. Weigt, 2013a, *PLoS Comput. Biol.* **9**, e1003176.
- Cocco, S., R. Monasson, and M. Weigt, 2013b, *J. Phys. Conf. Ser.* **473**, 012010.

- Cohen, L. B., and B. M. Salzberg, 1978, *Rev. Physiol. Biochem. Pharmacol.* **83**, 35.
- Condit, R., S. Lao, A. Singh, S. Esufali, and S. Dolins, 2014, *For. Ecol. Manage.* **316**, 21.
- Cook, S. J., et al., 2019, *Nature (London)* **571**, 63.
- Cooper, L. N., 1973, in *Collective Properties of Physical Systems: Proceedings of the Twenty-Fourth Nobel Symposium, Lerum, Sweden, 1973*, edited by B. Lundqvist and S. Lundqvist (Academic Press, New York), pp. 252–264.
- Cover, T. M., and J. A. Thomas, 1991, *Elements of Information Theory* (John Wiley & Sons, New York).
- Cragg, B. G., and H. N. V. Temperley, 1954, *Electroencephalogr. Clin. Neurophysiol.* **6**, 85.
- Cullen, J. M., E. Shaw, and H. A. Baldwin, 1965, *Anim. Behav.* **13**, 534.
- Dayan, P., and L. F. Abbott, 2001, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems* (MIT Press, Cambridge, MA).
- De Martino, D., A. M. C. Andersson, T. Bergmiller, C. C. Guet, and G. Tkačik, 2018, *Nat. Commun.* **9**, 2988.
- Demas, J., J. Manley, F. Tejera, K. Barber, H. Kim, F. M. Traub, B. Chen, and A. Vaziri, 2021, *Nat. Methods* **18**, 1103.
- Derrida, B., 1981, *Phys. Rev. B* **24**, 2613.
- Devine, B., and J. E. Cohen, 1992, *Absolute Zero Gravity: Science Jokes, Quotes and Anecdotes* (Simon and Schuster, New York).
- Dhawale, A. K., R. Poddar, S. B. E. Wolff, V. A. Normand, E. Kopelowitz, and B. P. Ölveczky, 2017, *eLife* **6**, e27702.
- Dombeck, D. A., C. D. Harvey, L. Tian, L. L. Looger, and D. W. Tank, 2010, *Nat. Neurosci.* **13**, 1433.
- Dorkenwald, S., et al., 2024, *Nature (London)* **634**, 124.
- Durbin, R., S. R. Eddy, A. Krogh, and G. Mitchison, 1998, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press, Cambridge, England).
- Espada, R., R. G. Parra, T. Mora, A. M. Walczak, and D. U. Ferreiro, 2015, *BMC Bioinf.* **16**, 207.
- Everitt, B., 1984, *An Introduction to Latent Variable Models* (Chapman and Hall, London).
- Ferrari, U., S. Deny, M. Chalk, G. Tkačik, O. Marre, and T. Mora, 2018, *Phys. Rev. E* **98**, 042410.
- Ferrari, U., T. Obuchi, and T. Mora, 2017, *Phys. Rev. E* **95**, 042321.
- Ferrenberg, A. M., and R. H. Swendsen, 1988, *Phys. Rev. Lett.* **61**, 2635.
- Field, G. D., and E. J. Chichilnisky, 2007, *Annu. Rev. Neurosci.* **30**, 1.
- Finn, R. D., et al., 2014, *Nucleic Acids Res.* **42**, D222.
- Fontenele, A. J., J. S. Sooter, V. K. Norman, S. H. Gautam, and W. L. Shew, 2024, *Sci. Adv.* **10**, eadq9303.
- Fontenele, A. J., et al., 2019, *Phys. Rev. Lett.* **122**, 208101.
- Friedman, N., S. Ito, B. A. W. Brinkman, M. Shimono, R. E. L. DeVille, K. A. Dahmen, J. M. Beggs, and T. C. Butler, 2012, *Phys. Rev. Lett.* **108**, 208102.
- Galpern, E. A., J. Marchi, T. Mora, A. M. Walczak, and D. U. Ferreiro, 2022, *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2204131119.
- Gardner, E., 1988, *J. Phys. A* **21**, 257.
- Gardner, E., and B. Derrida, 1988, *J. Phys. A* **21**, 271.
- Gauthier, J. L., G. D. Field, A. Sher, M. Greschner, J. Shlens, A. M. Litke, and E. J. Chichilnisky, 2009, *PLoS Biol.* **7**, e1000063.
- Gell-Mann, M., and F. E. Low, 1954, *Phys. Rev.* **95**, 1300.
- Ghosh, K., P. D. Dixit, L. Agozzino, and K. A. Dill, 2020, *Annu. Rev. Phys. Chem.* **71**, 213.
- Göbel, U., C. Sander, R. Scheiner, and A. Valencia, 1994, *Proteins* **18**, 309.
- Gordon, A., A. Banerjee, M. Koch-Janusz, and Z. Ringel, 2021, *Phys. Rev. Lett.* **126**, 240601.
- Granot-Atedgi, E., G. Tkačik, R. Segev, and E. Schneidman, 2013, *PLoS Comput. Biol.* **9**, e1002922.
- Hampson, R. E., J. D. Simeral, and S. A. Deadwyler, 1999, *Nature (London)* **402**, 610.
- Harte, J., and E. A. Newman, 2014, *Trends Ecol. Evol.* **29**, 384.
- Harte, J., T. Zillio, E. Conlisk, and A. B. Smith, 2008, *Ecology* **89**, 2700.
- Harvey, C. D., F. Collman, D. A. Dombeck, and D. W. Tank, 2009, *Nature (London)* **461**, 941.
- Hebb, D. O., 1949, *The Organization of Behavior: A Neuropsychological Theory* (John Wiley & Sons, New York).
- Helmhchen, F., and W. Denk, 2005, *Nat. Methods* **2**, 932.
- Hertz, J., A. Krogh, and R. G. Palmer, 1991, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood City, CA).
- Hinton, G. E., 2024, <https://www.nobelprize.org/prizes/physics/2024/hinton/lecture/>.
- Hires, S. A., D. A. Gutnisky, J. Yu, D. H. O'Connor, and K. Svoboda, 2015, *eLife* **4**, e00619.
- Hochberg, L. R., et al., 2012, *Nature (London)* **485**, 372.
- Hodgkin, A. L., and A. F. Huxley, 1952, *J. Physiol. (London)* **117**, 500.
- Hohenberg, P. C., and B. I. Halperin, 1977, *Rev. Mod. Phys.* **49**, 435.
- Hopfield, J. J., 1982, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 2554.
- Hopfield, J. J., 1984, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 3088.
- Hopfield, J. J., 2024, <https://www.nobelprize.org/prizes/physics/2024/hopfield/lecture/>.
- Hopfield, J. J., and D. W. Tank, 1985, *Biol. Cybern.* **52**, 141.
- Hopfield, J. J., and D. W. Tank, 1986, *Science* **233**, 625.
- Hornik, K., M. Stinchcombe, and H. White, 1989, *Neural Netw.* **2**, 359.
- Hoshal, B. D., C. M. Holmes, K. Bojanek, J. Salisbury, M. J. Berry II, O. Marre, and S. E. Palmer, 2023, *bioRxiv*, 10.1101/2023.08.08.552526.
- Ibarra, R. U., J. S. Edwards, and B. O. Palsson, 2002, *Nature (London)* **420**, 186.
- James, W., 1904, *Psychology: Brief Course* (Henry Holt, New York).
- Jaynes, E. T., 1957, *Phys. Rev.* **106**, 620.
- Jaynes, E. T., 1982, *Proc. IEEE* **70**, 939.
- Jin, L., Z. Han, J. Platsis, J. R. A. Woolerton, L. B. Cohen, and V. A. Pieribone, 2012, *Neuron* **75**, 779.
- Johnson, F. H., O. Shimomura, Y. Saiga, L. C. Gershman, G. T. Reynolds, and J. R. Waters, Jr., 1962, *J. Cell. Comp. Physiol.* **60**, 85.
- Jona-Lasinio, G., 1975, *Nuovo Cimento Soc. Ital. Fis.* **26B**, 99.
- Jones, K. E., P. K. Campbell, and R. A. Normann, 1992, *Ann. Biomed. Eng.* **20**, 423.
- Joshua, M., and S. G. Lisberger, 2015, *Neuroscience* **296**, 80.
- Jumper, J., et al., 2021, *Nature (London)* **596**, 583.
- Jun, J. J., et al., 2017, *Nature (London)* **551**, 232.
- Kadanoff, L. P., 1966, *Physics* **2**, 263.
- Kaifosh, P., J. D. Zaremba, N. B. Danielson, and A. Losonczy, 2014, *Front. Neuroinf.* **8**, 80.
- Kandel, E. R., J. H. Schwartz, T. M. Jessell, S. A. Siegelbaum, and A. J. Hudspeth, 2012, *Principles of Neural Science*, 5th ed. (McGraw-Hill, New York).
- Keller, J. B., and B. Zumino, 1959, *J. Chem. Phys.* **30**, 1351.
- Keup, C., T. Kühn, D. Dahmen, and M. Helias, 2021, *Phys. Rev. X* **11**, 021064.

- Kirkpatrick, S., C. D. Gelatt, Jr., and M. P. Vecchi, 1983, *Science* **220**, 671.
- Kirkpatrick, S., and B. Selman, 1994, *Science* **264**, 1297.
- Kivelson, S. A., J. M. Jiang, and J. Chang, 2024, *Statistical Mechanics of Phases and Phase Transitions* (Princeton University Press, Princeton, NJ).
- Kline, A. G., and S. E. Palmer, 2022, *New J. Phys.* **24**, 033007.
- Koch-Janusz, M., and Z. Ringel, 2018, *Nat. Phys.* **14**, 578.
- Kochkov, D., et al., 2024, *Nature (London)* **632**, 1060.
- Krause, J., and G. D. Ruxton, 2002, *Living in Groups* (Oxford University Press, Oxford).
- Krishnamurthy, K., T. Can, and D. J. Schwab, 2022, *Phys. Rev. X* **12**, 011011.
- Kunkin, W., and H. W. Firsch, 1969, *Phys. Rev.* **177**, 282.
- Landau, L. D., and E. M. Lifshitz, 1977, *Statistical Physics* (Pergamon Press, Oxford).
- Lapedes, A., and R. Farber, 1988, in *Neural Information Processing Systems*, edited by D. Z. Anderson (American Institute of Physics, New York), pp. 442–456.
- Lapedes, A., B. Giraud, and C. Jarzynski, 2012, [arXiv:1207.2484](https://arxiv.org/abs/1207.2484).
- Lapedes, A. S., B. G. Giraud, L. C. Liu, and G. D. Stormo, 1998, Los Alamos National Laboratory Report No. LA-UR-98-1094.
- LeCun, Y., 1987, Ph.D. thesis (Université Pierre et Marie Curie).
- LeCun, Y., Y. Bengio, and G. Hinton, 2015, *Nature (London)* **521**, 436.
- Lee, E. D., and B. C. Daniels, 2019, *J. Open Res. Software* **7**, 3.
- Lee, K. K., S. Liu, K. Croker, D. R. Huggins, M. Tikhonov, M. Mani, and S. Kuehn, 2024, bioRxiv, [10.1101/2024.03.15.584851](https://doi.org/10.1101/2024.03.15.584851).
- Leifer, A. M., C. Fang-Yen, M. Gershoff, M. J. Alema, and A. D. T. Samuel, 2011, *Nat. Methods* **8**, 147.
- Levin, E., N. Tishby, and S. A. Solla, 1990, *Proc. IEEE* **78**, 1568.
- Lezon, T. R., J. R. Banavar, M. Cieplak, A. Maritan, and N. V. Fedoroff, 2006, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 19033.
- Li, H., R. Helling, C. Tang, and N. S. Wingreen, 1996, *Science* **273**, 666.
- Lipa, J. A., D. R. Swanson, J. A. Nissen, T. C. P. Chui, and U. E. Israelsson, 1996, *Phys. Rev. Lett.* **76**, 944.
- Litke, A. M., et al., 2004, *IEEE Trans. Nucl. Sci.* **51**, 1434.
- Little, W. A., 1974, *Math. Biosci.* **19**, 101.
- Little, W. A., and G. L. Shaw, 1975, *Behav. Biol.* **14**, 115.
- Little, W. A., and G. L. Shaw, 1978, *Math. Biosci.* **39**, 281.
- Liu, Y. S., C. F. Stevens, and T. O. Sharpee, 2009, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 16499.
- Loback, A., J. Prentice, M. Ioffe, and M. J. Berry II, 2017, *Neural Comput.* **29**, 3119.
- Lynn, C. W., C. M. Holmes, W. Bialek, and D. J. Schwab, 2022a, *Phys. Rev. Lett.* **129**, 118101.
- Lynn, C. W., C. M. Holmes, W. Bialek, and D. J. Schwab, 2022b, *Phys. Rev. E* **106**, 034102.
- Lynn, C. W., Q. Yu, R. Pang, W. Bialek, and S. E. Palmer, 2025, *Phys. Rev. Res.* **7**, L022039.
- Lynn, C. W., Q. Yu, R. Pang, S. E. Palmer, and W. Bialek, 2025, *Phys. Rev. E* **111**, 054411.
- Maass, W., T. Natschläger, and H. Markram, 2002, *Neural Comput.* **14**, 2531.
- Machta, B. B., R. Chachra, M. K. Transtrum, and J. P. Sethna, 2013, *Science* **342**, 604.
- Macke, J. H., I. Murray, and P. Latham, 2011, in *Advances in Neural Information Processing Systems*, Vol. 24, edited by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger (Curran Associates, Red Hook, NY).
- Macke, J. H., M. Opper, and M. Bethge, 2011, *Phys. Rev. Lett.* **106**, 208102.
- Maheswaranathan, N., et al., 2023, *Neuron* **111**, 2742.
- Major, G., R. Baker, E. Aksay, B. Mensh, H. S. Seung, and D. W. Tank, 2004, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 7739.
- Major, G., R. Baker, E. Aksay, H. S. Seung, and D. W. Tank, 2004, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 7745.
- Manley, J., S. Lu, K. Barber, J. Demas, H. Kim, D. Meyer, F. M. Traub, and A. Vaziri, 2024, *Neuron* **112**, 1694.
- Maoz, O., G. Tkacik, M. S. Esteki, R. Kiani, and E. Schneidman, 2020, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 25066.
- Marchetti, M. C., J. F. Joanny, S. Ramaswamy, T. B. Liverpool, J. Prost, M. Rao, and R. A. Simha, 2013, *Rev. Mod. Phys.* **85**, 1143.
- Marchi, J., E. A. Galpern, R. Espada, D. U. Ferreiro, A. M. Walczak, and T. Mora, 2019, *PLoS Comput. Biol.* **15**, e1007282.
- Marko, J. F., and E. D. Siggia, 1995, *Macromolecules* **28**, 8759.
- Marks, D. S., L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, 2011, *PLoS One* **6**, e28766.
- Marre, O., D. Amodei, K. Sadeghi, F. Soo, T. E. Holy, and M. J. Berry II, 2012, *J. Neurosci. Methods* **32**, 14859.
- Marro, J., and R. Dickman, 1999, *Nonequilibrium Phase Transitions in Lattice Models* (Cambridge University Press, Cambridge, England).
- Martignon, L., G. Deco, K. Laskey, M. Diamond, W. Freiwald, and E. Vaadia, 2000, *Neural Comput.* **12**, 2621.
- Martin, P. C., E. D. Siggia, and H. A. Rose, 1973, *Phys. Rev. A* **8**, 423.
- Maruyama, R., K. Maeda, H. Moroda, I. Kato, M. Inoue, H. Miyakawa, and T. Aonishi, 2014, *Neural Netw.* **55**, 11.
- Mathis, A., P. Mamićanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge, 2018, *Nat. Neurosci.* **21**, 1281.
- McCulloch, W. S., and W. Pitts, 1943, *Bull. Math. Biol.* **5**, 115.
- McNaughton, B. L., J. O’Keefe, and C. A. Barnes, 1983, *J. Neurosci. Methods* **8**, 391.
- Mead, C. A., 1989, *Analog VLSI and Neural Systems* (Addison-Wesley, Redwood City, CA).
- Mehta, P., M. Bukov, C.-H. Wang, A. G. R. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, 2019, *Phys. Rep.* **810**, 1.
- Meister, M., J. Pine, and D. A. Baylor, 1994, *J. Neurosci. Methods* **51**, 95.
- Merchan, L., and I. Nemenman, 2016, *J. Stat. Phys.* **162**, 1294.
- Meshulam, L., J. L. Gauthier, C. D. Brody, D. W. Tank, and W. Bialek, 2017, *Neuron* **96**, 1178.
- Meshulam, L., J. L. Gauthier, C. D. Brody, D. W. Tank, and W. Bialek, 2018, [arXiv:1812.11904](https://arxiv.org/abs/1812.11904).
- Meshulam, L., J. L. Gauthier, C. D. Brody, D. W. Tank, and W. Bialek, 2019, *Phys. Rev. Lett.* **123**, 178103.
- Meshulam, L., J. L. Gauthier, C. D. Brody, D. W. Tank, and W. Bialek, 2021, [arXiv:2112.14735](https://arxiv.org/abs/2112.14735).
- Mézard, M., and A. Montanari, 2009, *Information, Physics, and Computation* (Oxford University Press, Oxford).
- Mézard, M., G. Parisi, and M. A. Virasoro, 1987, *Spin Glass Theory and Beyond* (World Scientific, Singapore).
- MICrONS Consortium, 2025, *Nature (London)* **640**, 435.
- Minaee, S., T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao, 2024, [arXiv:2402.06196](https://arxiv.org/abs/2402.06196).
- Minsky, M., and S. Papert, 1969, *Perceptrons* (MIT Press, Cambridge, MA).
- Monasson, R., R. Zecchina, S. Kirkpatrick, B. Selman, and L. Troyansky, 1999, *Nature (London)* **400**, 133.
- Mora, T., and W. Bialek, 2011, *J. Stat. Phys.* **144**, 268.
- Mora, T., S. Deny, and O. Marre, 2015, *Phys. Rev. Lett.* **114**, 078105.
- Mora, T., A. M. Walczak, W. Bialek, and C. G. Callan, Jr., 2010, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 5405.

- Mora, T., A. M. Walczak, L. Del Castello, F. Ginelli, S. Melillo, L. Parisi, M. Viale, A. Cavagna, and I. Giardina, 2016, *Nat. Phys.* **12**, 1153.
- Morales, G. B., S. di Santo, and M. A. Muñoz, 2023, *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2208998120.
- Morrell, M. C., A. J. Sederberg, and I. Nemenman, 2021, *Phys. Rev. Lett.* **126**, 118302.
- Mukamel, E. A., A. Nimmerjahn, and M. J. Schnitzer, 2009, *Neuron* **63**, 747.
- Munn, B. R., E. Müller, I. Favre-Bulle, E. Scott, M. Breakspear, and J. M. Shine, 2024, [10.1101/2024.06.22.600219](https://doi.org/10.1101/2024.06.22.600219).
- Muñoz, M. A., 2018, *Rev. Mod. Phys.* **90**, 031001.
- Musallam, S., B. D. Corneil, B. Greger, H. Scherberger, and R. A. Andersen, 2004, *Science* **305**, 258.
- Narayanan, K., J. A. Pereiro-Morejon, A. Ferrero, V. Onesto, S. Forciniti, L. L. del Mercato, R. Mulet, A. DeMartino, D. S. Tourigny, and D. DeMartino, 2024, [arXiv:2405.13424](https://arxiv.org/abs/2405.13424).
- Neher, E., 1994, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 98.
- Nemenman, I., G. D. Lewen, W. Bialek, and R. R. de Ruyter van Steveninck, 2008, *PLoS Comput. Biol.* **4**, e1000025.
- Ngampruetikorn, V., and D. J. Schwab, 2023, [arXiv:2309.14047](https://arxiv.org/abs/2309.14047).
- Nguyen, H. C., R. Zecchina, and J. Berg, 2017, *Adv. Phys.* **66**, 197.
- Nguyen, J. P., F. B. Shipley, A. N. Linder, G. S. Plummer, J. W. Shaevitz, and A. M. Leifer, 2016, *Proc. Natl. Acad. Sci. U.S.A.* **113**, E1074.
- Nicoletti, G., S. Suweis, and A. Maritan, 2020, *Phys. Rev. Res.* **2**, 023144.
- Obuchi, T., S. Cocco, and R. Monasson, 2015, *J. Stat. Phys.* **161**, 598.
- O'Dwyer, J. P., A. Rominger, and X. Xiao, 2017, *Ecol. Lett.* **20**, 832.
- O'Keefe, J., and J. Dostrovsky, 1971, *Brain Res.* **34**, 171.
- O'Keefe, J., and L. Nadel, 1978, *The Hippocampus as a Cognitive Map* (Oxford University Press, Oxford).
- Orth, J., I. Thiele, and B. O. Palsson, 2010, *Nat. Biotechnol.* **28**, 245.
- Pachitariu, M., A. M. Packer, N. Pettit, H. Dalgleish, M. Hausser, and M. Sahani, 2013, in *Advances in Neural Information Processing Systems*, Vol. 26, edited by C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (Curran Associates, Red Hook, NY).
- Packer, A. M., L. E. Russell, H. W. P. Dalgleish, and M. Häusser, 2015, *Nat. Methods* **12**, 140.
- Parrish, J. K., and W. M. Hammer, 1997, Eds., *Animal Groups in Three Dimensions* (Cambridge University Press, Cambridge, England).
- Pascanu, R., T. Mikolov, and Y. Bengio, 2013, *Proc. Mach. Learn. Res.* **28**, 1310, <https://proceedings.mlr.press/v28/pascanu13.html>.
- Pereira, T. D., D. E. Aldarondo, L. Willmore, M. Kislin, S. S.-H. Wang, M. Murthy, and J. W. Shaevitz, 2019, *Nat. Methods* **16**, 117.
- Pine, J., and J. Gilbert, 1982, *Soc. Neurosci. Abstr.* **8**, 670.
- Platsis, J., X. Ye, A. M. Ahrens, C. Liu, I. A. Chen, I. G. Davison, L. Tian, V. A. Pieribone, and J. L. Chen, 2023, *Nat. Methods* **20**, 1095.
- Pontryagin, L. S., 1987, *L. S. Pontryagin Collected Works, Vol. 4: Mathematical Theory of Optimal Processes* (Routledge, London).
- Prasher, D. C., V. K. Eckenrode, W. W. Ward, F. G. Prendergast, and M. J. Cormier, 1992, *Gene* **111**, 229.
- Prentice, J. S., J. Homann, K. D. Simmons, G. Tkačik, V. Balasubramanian, and P. C. Nelson, 2011, *PLoS One* **6**, e19884.
- Pressé, S., K. Ghosh, J. Lee, and K. A. Dill, 2013, *Rev. Mod. Phys.* **85**, 1115.
- Radvansky, B. A., and D. A. Dombeck, 2018, *Nat. Commun.* **9**, 839.
- Ramirez, L., W. Bialek, S. E. Palmer, and D. J. Schwab, 2024, [arXiv:2112.14334v2](https://arxiv.org/abs/2112.14334v2).
- Ramón y Cajal, S., 1893, *La Cellule* **119**, 17.
- Ramón y Cajal, S., 1894, *Proc. R. Soc. London* **55**, 444.
- Randi, F., A. K. Sharma, S. Dvali, and A. M. Leifer, 2023, *Nature (London)* **623**, 406.
- Reynolds, C., 1987, in *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques, Anaheim, 1987*, edited by M. C. Stone (Association for Computing Machinery, New York), pp. 25–34.
- Rickgauer, J. P., K. Deisseroth, and D. W. Tank, 2014, *Nat. Neurosci.* **17**, 1816.
- Rieke, F., D. Warland, R. de Ruyter van Steveninck, and W. Bialek, 1997, *Spikes: Exploring the Neural Code* (MIT Press, Cambridge, MA).
- Roberts, D. A., 2021, [arXiv:2104.00008](https://arxiv.org/abs/2104.00008).
- Roberts, D. A., and S. Yaida, 2022, *The Principles of Deep Learning Theory: An Effective Theory Approach to Understanding Neural Networks* (Cambridge University Press, Cambridge, England).
- Rosen, M. J., M. Davison, D. Bhaya, and D. S. Fisher, 2015, *Science* **348**, 1019.
- Rosenblatt, F., 1961, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms* (Spartan Books, Washington, DC).
- Roudi, Y., S. Nirenberg, and P. E. Latham, 2009, *PLoS Comput. Biol.* **5**, e1000380.
- Roy, S., N. Y. Jun, E. L. Davis, J. Pearson, and G. D. Field, 2021, *Nature (London)* **592**, 409.
- Rumelhart, D. E., G. E. Hinton, and R. J. Williams, 1986, *Nature (London)* **323**, 533.
- Russ, W., D. M. Lowery, P. Mishra, M. B. Yaffe, and R. Ranganathan, 2005, *Nature (London)* **437**, 579.
- Russ, W. P., et al., 2020, *Science* **369**, 440.
- Sampaio Filho, C. I. N., L. de Arcangelis, H. J. Herrmann, D. Plenz, P. Kells, T. L. Ribeiro, and J. S. Andrade, Jr., 2024, *Sci. Rep.* **14**, 7002.
- Sarra, C., L. Sarra, L. Di Carlo, T. GrandPre, Y. Zhang, C. G. Callan, Jr., and W. Bialek, 2024, [arXiv:2408.08037](https://arxiv.org/abs/2408.08037).
- Schneidman, E., M. J. Berry II, R. Segev, and W. Bialek, 2006, *Nature (London)* **440**, 1007.
- Schneidman, E., S. Still, M. J. Berry II, and W. Bialek, 2003, *Phys. Rev. Lett.* **91**, 238701.
- Schnitzer, M. J., and M. Meister, 2003, *Neuron* **37**, 499.
- Schwab, D. J., I. Nemenman, and P. Mehta, 2014, *Phys. Rev. Lett.* **113**, 068102.
- Segev, R., J. Goodhouse, J. Puchalla, and M. J. Berry II, 2004, *Nat. Neurosci.* **7**, 1155.
- Serruya, M. D., N. G. Hatsopoulos, L. Paninski, M. R. Fellows, and J. P. Donoghue, 2002, *Nature (London)* **416**, 141.
- Sessak, V., and R. Monasson, 2009, *J. Phys. A* **42**, 055001.
- Sethna, J., 2021, *Statistical Mechanics: Entropy, Order Parameters, and Complexity* (Oxford University Press, Oxford).
- Seung, H. S., 1996, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 13339.
- Shannon, C. E., 1948, *Bell Syst. Tech. J.* **27**, 379.
- Shimomura, O., F. H. Johnson, and Y. Saiga, 1962, *J. Cell. Comp. Physiol.* **59**, 223.
- Shlens, J., 2014, [arXiv:1404.1100](https://arxiv.org/abs/1404.1100).
- Shlens, J., G. D. Field, J. L. Gauthier, M. I. Grivich, D. Petrusca, A. Sher, A. M. Litke, and E. J. Chichilnisky, 2006, *J. Neurosci. Methods* **26**, 8254.
- Skinner, D. J., P. Lamaire, and M. Mani, 2024, *bioRxiv*, [10.1101/2024.07.26.605398](https://doi.org/10.1101/2024.07.26.605398).
- Smith, S. L., and M. Häusser, 2010, *Nat. Neurosci.* **13**, 1144.
- Smolensky, P., 1986, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1—Foundations*,

- edited by D. E. Rumelhart and J. L. McClelland (MIT Press, Cambridge, MA), pp. 194–281.
- Socolich, M., S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner, and R. Ranganathan, 2005, *Nature (London)* **437**, 512.
- Sofroniew, N. J., D. Flickinger, J. King, and K. Svoboda, 2016, *eLife* **5**, e14472.
- Solovey, G., L. M. Alonso, T. Yanagawa, N. Fujii, M. O. Magnasco, G. A. Cecchi, and A. Proekt, 2015, *J. Neurosci. Methods* **35**, 10866.
- Srivastava, K. H., C. M. Holmes, M. Vellema, A. R. Pack, C. P. H. Elemans, I. Nemenman, and S. J. Sober, 2017, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 1171.
- Steinmetz, N. A., C. Koch, K. D. Harris, and M. Carandini, 2018, *Curr. Opin. Neurobiol.* **50**, 92.
- Steinmetz, N. A., et al., 2021, *Science* **372**, eabf4588.
- Stevenson, I. H., and K. P. Kording, 2011, *Nat. Neurosci.* **14**, 139.
- Steyn-Ross, M. L., and D. A. Steyn-Ross, 2016, *Phys. Rev. E* **93**, 022402.
- Stroud, R. M., 1974, *Sci. Am.* **231**, No. 1, 74.
- Sulkowska, J. I., F. Morcos, M. Weigt, T. Hwa, and J. N. Onuchic, 2012, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 10340.
- Swendsen, R. H., 1984, *Phys. Rev. Lett.* **52**, 1165.
- Tang, A., et al., 2008, *J. Neurosci. Methods* **28**, 505.
- Tang, C., K. Wiesenfeld, P. Bak, S. Coppersmith, and P. Littlewood, 1987, *Phys. Rev. Lett.* **58**, 1161.
- Tavoni, G., S. Cocco, and R. Monasson, 2016, *J. Comput. Neurosci.* **41**, 269.
- Tavoni, G., U. Ferrari, F. P. Battaglia, S. Cocco, and R. Monasson, 2017, *Network Neurosci.* **1**, 275.
- Taylor, D. M., S. I. H. Tillery, and A. B. Schwartz, 2002, *Science* **296**, 1829.
- Tian, L., J. Akerboom, E. R. Schreiter, and L. L. Looger, 2012, *Prog. Brain Res.* **196**, 79.
- Tiberi, L., J. Stappmanns, T. Kühn, T. Luu, D. Dahmen, and M. Helias, 2022, *Phys. Rev. Lett.* **128**, 168301.
- Tishby, N., F. C. Pereira, and W. Bialek, 1999, in *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing, Monticello, IL, 1999*, edited by B. Hajek and R. S. Sreenivas (University of Illinois Press, Champaign, IL), pp. 368–377 [[arXiv:physics/0004057](https://arxiv.org/abs/physics/0004057)].
- Tkačik, G., O. Marre, D. Amodei, E. Schneidman, W. Bialek, and M. J. Berry, 2014, *PLoS Comput. Biol.* **10**, e1003408.
- Tkačik, G., O. Marre, T. Mora, D. Amodei, M. J. Berry II, and W. Bialek, 2013, *J. Stat. Mech.* P03011.
- Tkačik, G., T. Mora, O. Marre, D. Amodei, S. E. Palmer, M. J. Berry II, and W. Bialek, 2015, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 11508.
- Tkačik, G., J. S. Prentice, V. Balasubramanian, and E. Schneidman, 2010, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 14419.
- Tkačik, G., E. Schneidman, M. J. Berry II, and W. Bialek, 2006, [arXiv:q-bio/0611072](https://arxiv.org/abs/q-bio/0611072).
- Tkačik, G., E. Schneidman, M. J. Berry II, and W. Bialek, 2009, [arXiv:0912.5409](https://arxiv.org/abs/0912.5409).
- Toner, J., and Y. Tu, 1995, *Phys. Rev. Lett.* **75**, 4326.
- Toner, J., and Y. Tu, 1998, *Phys. Rev. E* **58**, 4828.
- Trautmann, E. M., et al., 2023, bioRxiv, 10.1101/2023.02.01.526664.
- Treves, A., O. Miglino, and D. Parisi, 1992, *Psychobiology* **20**, 1.
- Tsien, R. Y., 2009, *Angew. Chem., Int. Ed.* **48**, 5612.
- Tsoar, A., R. Nathan, Y. Bartan, A. Vyssotski, G. Dell’Omo, and N. Ulanovsky, 2011, *Proc. Natl. Acad. Sci. U.S.A.* **108**, E718.
- Tsodyks, M., and T. Sejnowski, 1995, *Int. J. Neural Syst.* **6**, 81.
- Turing, A. M., 1937, *Proc. London Math. Soc.* s2-**42**, 230.
- Urai, A. E., B. Doiron, A. M. Leifer, and A. K. Churchland, 2022, *Nat. Neurosci.* **25**, 11.
- van Vreeswijk, C., and H. Sompolinsky, 1998, *Neural Comput.* **10**, 1321–1371.
- Varshney, L. R., B. L. Chen, E. Paniagua, D. H. Hall, and D. B. Chklovskii, 2011, *PLoS Comput. Biol.* **7**, e1001066.
- Vicsek, T., A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet, 1995, *Phys. Rev. Lett.* **75**, 1226.
- Villegas, P., A. Cavagna, M. Cencini, H. Fort, and T. S. Grigera, 2021, *R. Soc. Open Sci.* **8**, 202200.
- Villegas, P., T. Gili, G. Caldarelli, and A. Gabrielli, 2024, *Phys. Rev. E* **109**, L042402.
- Villette, V., et al., 2019, *Cell* **179**, 1590.
- Vishwanathan, A., et al., 2024, *Nat. Neurosci.* **27**, 2443.
- Vogels, T. P., K. Rajan, and L. F. Abbott, 2005, *Annu. Rev. Neurosci.* **28**, 357.
- von Hippel, P. H., and O. G. Berg, 1986, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 1608.
- Vorontsov, E., C. Trabelsi, S. Kadoury, and C. Pal, 2017, *Proc. Mach. Learn. Res.* **70**, 3570, <https://proceedings.mlr.press/v70/vorontsov17a.html>.
- Ward, C. S., C.-M. Yung, K. M. Davis, S. K. Blinebry, T. C. Williams, Z. I. Johnson, and H. D. E., 2017, *ISME J.* **11**, 1412.
- Watkin, T. L. H., A. Rau, and M. Biehl, 1993, *Rev. Mod. Phys.* **65**, 499.
- Weigt, M., R. White, H. Szurmant, J. Hoch, and T. Hwa, 2009, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 67.
- Weisenburger, S., et al., 2019, *Cell* **177**, 1050.
- White, J. G., E. Southgate, J. N. Thomson, and S. Brenner, 1986, *Phil. Trans. R. Soc. B* **314**, 1.
- Wiener, N., 1958, *Nonlinear Problems in Random Theory* (MIT Press, Cambridge, MA).
- Wiener, S. I., C. A. Paul, and H. Eichenbaum, 1989, *J. Neurosci. Methods* **9**, 2737.
- Willett, F. R., D. T. Avansino, L. R. Hochberg, J. M. Henderson, and K. V. Shenoy, 2021, *Nature (London)* **593**, 249.
- Wilson, K. G., 1979, *Sci. Am.* **241**, No. 2, 158.
- Wilson, K. G., 1983, *Rev. Mod. Phys.* **55**, 583.
- Wilson, K. G., and J. Kogut, 1974, *Phys. Rep.* **12**, 75.
- Wilson, M. A., and B. L. McNaughton, 1993, *Science* **261**, 1055.
- Wu, E. G., et al., 2023, bioRxiv, 10.1101/2023.11.06.565889.
- Yartsev, M. M., and N. Ulanovsky, 2013, *Science* **340**, 367.
- Zhang, K., 1996, *J. Neurosci. Methods* **16**, 2112.
- Zhang, Y., et al., 2023, *Nature (London)* **615**, 884.
- Zhang, Z., L. Bai, L. Cong, P. Yu, T. Zhang, W. Shi, F. Li, J. Du, and K. Wang, 2021, *Nat. Biotechnol.* **39**, 74.
- Zhao, Z., H. Zhu, X. Li, L. Sun, F. He, J. E. Chung, D. F. Liu, L. Frank, L. Luan, and C. Xie, 2023, *Nat. Biomed. Eng.* **7**, 520.
- Zhu, S. C., Y. N. Wu, and D. Mumford, 1997, *Neural Comput.* **9**, 1627.
- Ziv, Y., L. D. Burns, W. D. Cocker, W. O. Hamel, K. K. Ghosh, L. J. Kitch, A. El Gamal, and M. J. Schnitzer, 2013, *Nat. Neurosci.* **16**, 264.
- Zong, W., et al., 2017, *Nat. Methods* **14**, 713.