PNAS RESEARCH ARTICLE | PHYSICS NEUROSCIENCE OPEN ACCESS

On the dimensionality of behavior

William Bialek[a,b,c,1]

Contributed by William Bialek; received October 19, 2020; accepted March 4, 2022; reviewed by André Brown, Sandeep Robert Datta, and Samuel Sober

There is a growing effort in the "physics of behavior" that aims at complete quantitative characterization of animal movements under more complex, naturalistic conditions. One reaction to the resulting explosion of high-dimensional data is the search for low-dimensional structure. Here I try to define more clearly what we mean by the dimensionality of behavior, where observable behavior may consist of either continuous trajectories or sequences of discrete states. This discussion also serves to isolate situations in which the dimensionality of behavior is effectively infinite.

information | prediction | complexity

Significance

How do we characterize animal behavior? Psychophysics started with human behavior in the laboratory, and focused on simple contexts, such as the decision among just a few alternative actions in response to sensory inputs. In contrast, ethology focused on animal behavior in the natural environment, emphasizing that evolution selects potentially complex behaviors that are useful in specific contexts. New experimental methods now make it possible to monitor animal and human behaviors in vastly greater detail. This "physics of behavior" holds the promise of combining the psychophysicist's quantitative approach with the ethologist's appreciation of natural context. One question surrounding this growing body of data concerns the dimensionality of behavior. Here I try to give this concept a precise definition.

Observations on behavior provide a window into the dynamics of the brain and mind. This is an ancient idea, now receiving renewed attention because of the explosive growth of methods for quantitative measurements of behavior (1–8). These methods produce enormous quantities of raw data, such as high-resolution videos, so there is an obvious practical interest in data compression. This often involves searching for a low-dimensional description of the animal's configuration or posture at each moment in time. This search is grounded both by the observation that even large and complex animals have relatively small numbers of muscles or joints and by direct evidence that motor behaviors are described by low-dimensional models in organisms from the worm *Caenorhabditis elegans* to humans and nonhuman primates (1, 9–14).

Reducing great volumes of video data to time series for just a few degrees of freedom is a triumph. The fact that this now can be done more or less automatically with machine learning methods means that exhaustive and quantitative characterization of behavior is possible in a much wider range of organisms, under a wider range of conditions. But the classical literature on dynamical systems reminds us that the time series of even a single variable could encode a higher-dimensional structure (15, 16). Indeed, this seems natural: The brain that generates behavior has many degrees of freedom, and observations of behavior should be sensitive to these potentially high-dimensional dynamics. On the other hand, the dynamics of large neural networks might be confined to low-dimensional manifolds, perhaps to match the dimensionality of motor behaviors (17–20).

All of these developments point to the problem of defining the dimensionality of behavior. In the extreme, we can imagine that the observable behavior reduces to a single function of time, as with the opening angle of a clamshell. Can we analyze this time series to identify a well-defined dimensionality for the underlying dynamics? Is it possible that this dimensionality is effectively infinite?

A Context for Phenomenology

The quantitative analysis of behavior, including what follows here, is unapologetically phenomenological. The question is not "How does the brain generate behavior?" but rather "What is it about behavior that we would like to explain?" In an era of highly mechanistic biology, this emphasis on phenomenological description may seem odd. So, at the risk of repeating things that are well known, it is useful to remind ourselves of the long historical context for this approach.

If we want to explain why we look like our parents, a qualitative answer is that we carry copies of their DNA. But, if we want to understand the reliability with which traits pass from generation to generation, then DNA structure is not enough—the free energy differences between correct and incorrect base pairing are not sufficient to explain the reliability of molecular copying if the reactions are allowed to come to thermal equilibrium, and this problem arises not just in DNA replication but in every step of molecular information transmission. Cells achieve their observed reliability by holding these reactions away from equilibrium, allowing for proofreading or error correction (21, 22). In the absence of proofreading, the majority of proteins would contain at least one incorrect amino acid, and ∼10% of our genes would be different from those carried by either parent; these error rates are orders of magnitude larger than observed.

Author affiliations: [a]Joseph Henry Laboratories of Physics, Princeton University, Princeton, NJ 08544; [b]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544; and [c]Initiative for the Theoretical Sciences, The Graduate Center, City University of New York, New York, NY 10016

Author contributions: W.B. designed research, performed research, and wrote the paper.

Reviewers: A.B., MRC London Institute of Medical Sciences; S.R.D., Harvard Medical School; and S.S., Emory University.

The author declares no competing interest.

See online for related content such as Commentaries.

[1]Email: wbialek@princeton.edu.

Published April 29, 2022.

These quantitative differences are so large that life without proofreading would be qualitatively different.[*]

The example of proofreading highlights the importance of starting with a quantitative characterization of the phenomena we are trying to explain. For brains and behavior, this is an old idea. In the late nineteenth century, many people were trying to turn observations on seeing and hearing into quantitative experiments, creating a subject that would come to be called psychophysics (23). By ∼1910, these experiments were sufficiently well developed that Lorentz could look at data on the "minimum visible" and suggest that the retina is capable of counting single photons (24), and Rayleigh could identify the conflict between our ability to localize low-frequency sounds and the conventional wisdom that we are "phase deaf" (25). Both of these essentially theoretical observations, grounded in quantitative descriptions of human behavior, drove experimental efforts that unfolded over more than half a century.

Also ∼1910, von Frisch (26) was doing psychophysics experiments to demonstrate bees could, in fact, discriminate among the beautiful colors of the flowers that they pollinate.[†] But he took these experiments in a very different direction, focusing not on the discrete choices made by individual bees but on how these individuals communicated their sensory experiences to other residents of the hive, leading to the discovery of the "dance language" of bees. What grew out of the work by von Frisch and others was ethology (28), which emphasizes the richness of behavior in its natural context, the context in which it was selected for by evolution. Because ethologists wrestle with complex behaviors, they often resort to verbal description. In contrast, psychophysicists focus on situations in which subjects are constrained to a small number of discrete alternative behaviors, so it is natural to give a quantitative description by estimating the probabilities of different choices under various conditions.

The emergence of a quantitative language for the analysis of psychophysical experiments was aided by the focus on constrained behaviors, but was not an automatic consequence of this focus. For photon counting in vision, the underlying physics suggests how the probability of seeing vs. not seeing will depend on light intensity (29), but the observation that human observers behave as predicted points to profound facts about the underlying mechanisms (30). Attempts to formalize the problems of detection led to a more general view of the choices among discrete alternative behaviors being discriminations among signals in a background of noise (31), and, in the 1950s and 1960s, this view was exported to experimental psychology (23). Much of this now seems like an exercise in probability and statistics, something obviously correct, but the early literature records considerable skepticism about whether this (or perhaps any) mathematization of human behavior would succeed.

More generally, quantitative phenomenology has been foundational, certainly in physics and also in the mainstream of biology. Mendel's genetics was a phenomenological description of the patterns of inheritance, and the realization that genes are arranged linearly along chromosomes came from a more refined quantitative analysis of these same patterns (32). The work of Hodgkin and Huxley (33) led to our modern understanding of electrical activity in terms of ion channel dynamics, but explicitly eschewed mechanistic claims in favor of phenomenology. The idea that transmission across a synapse depends on transmitter molecules packaged into vesicles emerged from the quantitative analysis of voltage fluctuations at the neuromuscular junction (34).

Even when we are searching for microscopic mechanisms, it is not anachronistic to explore macroscopic descriptions. Time and again, the scientific community has leaned on phenomenology to imagine the underlying mechanisms, often taking literally the individual terms in a mathematical description as representing the actual microscopic elements for which we should be searching, whether these are genes, ion channels, synaptic vesicles, or quarks (35–37). What is anachronistic, in the literal sense of the word, is to believe that microscopic mechanisms were discovered by direct microscopic observations without guidance from phenomenology on a larger scale.

In this broad context, how can we construct a quantitative phenomenology of complex, naturalistic behaviors? When we do psychophysics, we characterize behaviors with numbers that are meaningfully comparable across situations and across species. To give but one example, we can discuss the accumulation of evidence for decisions that humans and nonhuman primates make based on visual inputs, but we can use the same mathematical language to discuss decisions made by rodents based on auditory inputs (38). A quantitative characterization of naturalistic behaviors requires, similarly, that we attach comparable numbers to very different kinds of time series. The dimensionality of behavior is a candidate for this sort of unifying mathematical language.

Two Examples

To work toward a sharper definition, consider the case in which the behavior we observe is just a single function of time, $x(t)$. Two examples of such trajectories are shown in Fig. 1, *Left*, and we will see that these correspond to one-dimensional (blue) and two-dimensional (red) systems. Qualitatively, the blue trajectory varies on one characteristic time scale, while the red trajectory involves rough movements on a short time scale superposed on smoother movements over a longer time scale. Our task is to make these observations precise.

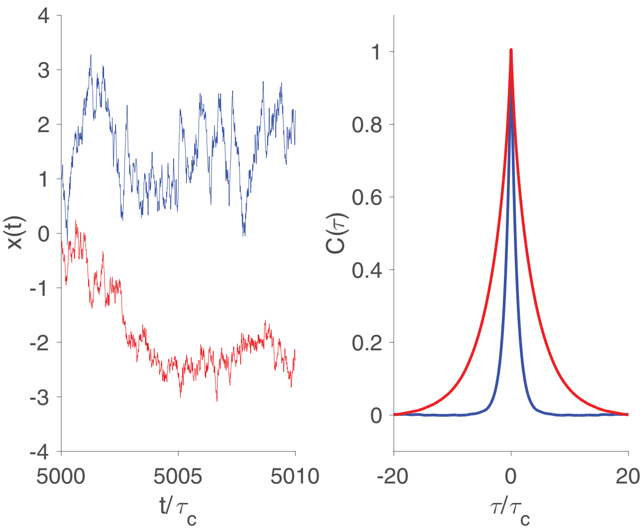

Fig. 1. Two examples of behavioral trajectories (*Left*) and their correlation functions (*Right*). One-dimensional example, from Eq. **1**, is shown in blue. Two-dimensional example, from Eq. **4**, is shown in red. Behavioral trajectories are offset for clarity, and time is measured in units of the correlation time τ_c.

[*] In retroviruses, including HIV, reproduction occurs without proofreading. The dramatically accelerated pace of evolution in these viruses gives a glimpse of how different life would be if the transmission of genetic information depended on base pairing alone.

[†] See also the remarkable early work from Turner, who studied both insect behavior and neuroanatomy in the decades straddling 1900 (27).

Let's work backward and start with a model for the behavior, a model in which it seems clear that the behavior really is one dimensional: The observed behavioral trajectory $x(t)$ is described completely by

$$\tau_c \frac{dx(t)}{dt} = -x(t) + \eta(t), \qquad [\mathbf{1}]$$

where $\eta(t)$ is white noise,

$$\langle \eta(t)\eta(t')\rangle = 2\tau_c \langle x^2\rangle \delta(t-t'). \qquad [\mathbf{2}]$$

It is important that the noise source is white; nonwhite noise sources, which themselves are correlated over time, are equivalent to having hidden degrees of freedom that carry these correlations. The blue trajectory in Fig. 1, *Left* is drawn from a simulation of Eq. **1** with $\langle x^2\rangle = 1$.

The observable consequences of the dynamics in Eqs. **1** and **2**. are well known: $x(t)$ will be a Gaussian stochastic process, with the two-point correlation function

$$C_1(\tau) = \langle x(t)x(t+\tau)\rangle = \langle x^2\rangle e^{-|\tau|/\tau_c}, \qquad [\mathbf{3}]$$

shown in Fig. 1, *Right*. We recall that, for a Gaussian process, once we specify the two-point function, there is nothing else to say about the system. Importantly, we can turn this around: If the observed behavior is a Gaussian stochastic process, and the correlations decay exponentially as in Eq. **3**, then Eqs. **1** and **2** are a complete description of the dynamics.

An example of a clearly two-dimensional system involves not only the observable $x(t)$ but also an internal variable $y(t)$,

$$\tau_c \frac{d}{dt}\begin{bmatrix} x(t) \\ y(t)\end{bmatrix} = -\begin{bmatrix} 1 & a \\ a & 1\end{bmatrix}\begin{bmatrix} x(t) \\ y(t)\end{bmatrix} + \begin{bmatrix} \eta_1(t) \\ \eta_2(t)\end{bmatrix}. \qquad [\mathbf{4}]$$

To keep things simple, we can assume that the driving noises are independent of one another, and, again, they should be white so that we are not hiding additional variables that carry correlations. Since *y* is hidden, its units are arbitrary, which allows us to have the strength of the noise driving each variable be the same without loss of generality, so that

$$\langle \eta_\mathrm{i}(t)\eta_\mathrm{j}(t')\rangle = 2\tau_c\langle x^2\rangle (1-a^2)\delta_\mathrm{ij}\delta(t-t'). \qquad [\mathbf{5}]$$

The choice to give each variable the same correlation time is just for illustration, as is the symmetry of the dynamical matrix; the red trajectory in Fig. 1, *Left* is drawn from a simulation of Eq. **4** with $\langle x^2\rangle = 1$ and $a = 0.75$. Again, $x(t)$ is Gaussian, but now the correlation function has two exponential decays,

$$C_2(\tau) = A_+ e^{-(1+a)|\tau|/\tau_c} + A_- e^{-(1-a)|\tau|/\tau_c} \qquad [\mathbf{6}]$$

$$A_\pm = \frac{1}{2}\langle x^2\rangle \frac{(1-a^2)}{1\pm a}, \qquad [\mathbf{7}]$$

shown in Fig. 1, *Right*. The short time scale $\tau_c/(1+a)$ corresponds to the rough movements seen in the trajectory, while $\tau_c/(1-a)$ corresponds to the smoother movements.

We see that a one-dimensional system generates behavior with a correlation function that has one exponential decay, while a two-dimensional system generates a correlation function with two exponential decays. We would like to turn this around, and say that, if we observe a certain structure in the behavioral correlations, then we can infer the underlying dimensionality.

Gaussian Processes More Generally

Analyzing behavioral trajectories by constructing explicit dynamical equations, as in Eq. **1** or **4**, may not be the best approach. In particular, if there are hidden dimensions, then there is no preferred coordinate system in the space of unmeasured variables, and hence no unique form for the dynamical equations. Let us think, instead, about the probability distribution of the observed trajectories $x(t)$. For Gaussian processes, this has the form

$$P[x(t)] = \frac{1}{Z} e^{-S[x(t)]} \qquad [\mathbf{8}]$$

$$S[x(t)] = \frac{1}{2}\int dt \int dt'\, x(t)K(t-t')x(t'), \qquad [\mathbf{9}]$$

where the integrals run over the interval of our observations, which should be long. The kernel $K(\tau)$ is inverse to the correlation function,

$$\int dt''\, K(t-t'')\langle x(t'')x(t')\rangle = \delta(t-t'). \qquad [\mathbf{10}]$$

We can divide the full trajectory $x(t)$ into the past, $\mathbf{x}_\mathrm{p}$, with $t \leq 0$, and the future, $\mathbf{x}_\mathrm{f}$, with $t > 0$. Schematically,

$$S[x(t)] = \frac{1}{2}\mathbf{x}_\mathrm{p}\cdot K_\mathrm{pp}\cdot \mathbf{x}_\mathrm{p} + \frac{1}{2}\mathbf{x}_\mathrm{f}\cdot K_\mathrm{ff}\cdot \mathbf{x}_\mathrm{f} + \mathbf{x}_\mathrm{p}\cdot K_\mathrm{pf}\cdot \mathbf{x}_\mathrm{f}, \qquad [\mathbf{11}]$$

where K_pf couples the past and future. More explicitly,

$$\mathbf{x}_\mathrm{p}\cdot K_\mathrm{pf}\cdot \mathbf{x}_\mathrm{f} = \int_0^\infty dt \int_0^\infty dt'\, x(-t)K(t+t')x(t'). \qquad [\mathbf{12}]$$

If K_pf is of finite rank, so that

$$K(t+t') = \sum_{\mu=1}^{D} a_\mu \phi_\mu(t)\phi_\mu(t'), \qquad [\mathbf{13}]$$

then everything that we can predict about future behavior given knowledge of past behavior is captured by D features,

$$P[\mathbf{x}_\mathrm{f}|\mathbf{x}_\mathrm{p}] = P[\mathbf{x}_\mathrm{f}|\{F_\mu\}] \qquad [\mathbf{14}]$$

$$F_\mu = \int_0^\infty dt\, \phi_\mu(t)x(-t). \qquad [\mathbf{15}]$$

Eq. **14** is telling us that the features $\{F_\mu\}$ provide "sufficient statistics" for making predictions. We recall that, in a dynamical system with D variables,

$$\frac{dy_\mathrm{i}}{dt} = g_\mathrm{i}(\{y_\mathrm{j}\}) + \eta_\mathrm{i}(t), \quad \mathrm{i} = 1, 2, \cdots, D, \qquad [\mathbf{16}]$$

predicting the future ($t > 0$) requires specifying D initial conditions (at $t = 0$). In this precise sense, the number of variables that we need to achieve maximum predictive power is the dimensionality of the dynamical system. To complete the argument, we need to show that K_pf has finite rank when correlations decay as a finite combination of exponentials; see *Appendix A*.

In the case of Gaussian stochastic processes, we thus arrive at a recipe for defining the dimensionality of the underlying dynamics. We estimate the correlation function, take its inverse to find the kernel, and isolate the part of this kernel which couples past and future. If this past–future kernel is of finite rank, then we can identify this rank with the dimensionality of the system. In Fig. 2, *Top* we see a sample trajectory (in red) from a system that is, by

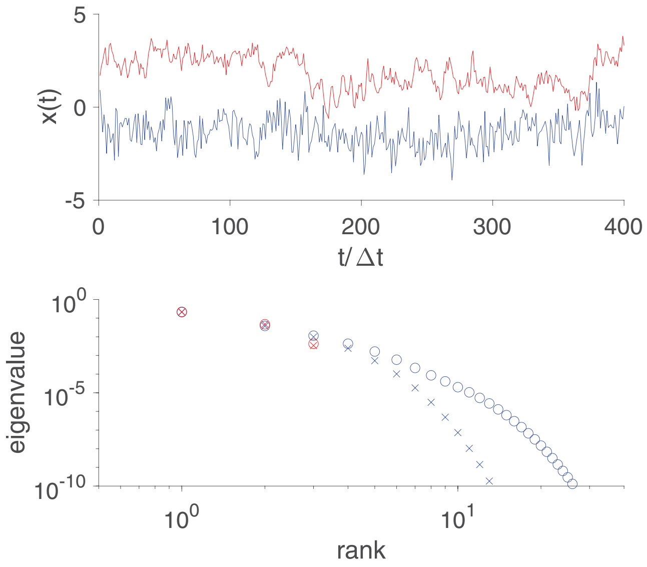

Fig. 2. Sample trajectories (*Top*) and spectra of the matrix K_{pf} (*Bottom*). In red is an example in which the underlying dynamics is three dimensional. In blue is an example with power law correlations, as in Eq. **17** with $\alpha = 1/2$, which is effectively infinite dimensional. Time is measured in discrete steps Δt, and spectra are computed in windows of duration $T = 100\Delta t$ (×) or $T = 1000\Delta t$ (○). Details are provided in *Appendix B*.

construction, three dimensional, with correlation times $4\times, 32\times$, and $256\times$ the discrete time step (Δt) of our observations. As explained in *Appendix B*, the coefficients a_μ in Eq. **13** can be found as the eigenvalues of a symmetric matrix, and these eigenvalues are plotted in Fig. 2, *Bottom* in rank order (in red). This numerical analysis yields three clearly nonzero eigenvalues, with other eigenvalues below 10^{-10}. Importantly, we find essentially the same three eigenvalues when the analysis is done in time windows of very different sizes—here $T = 100\Delta t$ and $T = 1,000\Delta t$, smaller and larger, respectively, than the longest correlation time.

The past–future coupling is not guaranteed to be of finite rank. More generally, if we analyze signals in a window of size T, then the rank can grow with T. This happens, for example, if behavioral correlations decay as a power of time,

$$\langle x(t)x(t')\rangle = \langle x^2\rangle \frac{t_0^\alpha}{t_0^\alpha + |t-t'|^\alpha}. \qquad [17]$$

Fig. 2, *Top* shows a sample trajectory from a Gaussian process with this correlation function (in blue), and Fig. 2, *Bottom* shows the associated spectrum of coefficients a_μ for $\alpha = 1/2$ (in blue). This illustrates both that there is no obvious cutoff to the spectrum and that the spectrum extends farther when the analysis is done in longer time windows [$T = 100\Delta t$ (×) vs. $T = 1000\Delta t$ (○)]. Indeed, the larger the window, the farther the spectrum extends, with no bound. Under these conditions, the dimensionality is effectively infinite.

The possibility that behavioral correlations decay as a power of time has a long and sometimes contentious history. It thus is worth noting that scaling of the correlation function implies an effectively infinite dimensionality, but it is not required. We can imagine situations in which the kernel K_{pf} has an arbitrarily large number of nonzero eigenvalues in the limit of long observation times even if the correlation is not precisely a power law.

While the relation of dimensionality to the spectrum of K_{pf} is attractive, estimating this spectrum from finite data can be challenging. Even if the true spectrum has only a finite number of nonzero eigenvalues, in matrices built from finite samples of data, the zero eigenvalues will be replaced by a continuous spectrum, and this could make it difficult, in practice, to distinguish finite from infinite dimensional processes. At the same time, it is important to emphasize that difficulty in resolving eigenvalues against a continuum generated by finite sample size is not evidence for low dimensionality, nor should a continuum be assigned as noise without further analysis. Random matrix theory provides quantitative predictions for spectral broadening in closely related contexts, including the dependence of spectra on sample size and matrix dimensionality, and these should provide a basis for identifying the contributions of noise to the observed eigenvalue spectra (39).

Discrete States

In many cases, it is natural to describe animal behavior as moving through a sequence of discrete states. We do this, for example, when we transcribe human speech to text, and when we describe a bacterium as running or tumbling (40). This identification of discrete states is not just an arbitrary quantization of continuous motor outputs, nor should it be a qualitative judgment by human observers. Discrete states should correspond to distinguishable clusters, or resolvable peaks in the distribution over the natural continuous variables, and the dynamics should consist of movements in the neighborhood of one peak that are punctuated by relatively rapid jumps to another peak (e.g., ref. 3). A "mechanism" for such discreteness is the existence of multiple dynamical attractors, with jumps driven by noise (e.g., refs. 1 and 13).

When behavioral states are discrete, how do we define dimensionality? Once again, it is useful to think about the simplest case, where there are just two behavioral states—perhaps "doing something" and "doing nothing"—and time is marked by discrete ticks of a clock. We can represent the two states at each time t by an Ising variable $\sigma_t = \pm 1$. If the sequence of behavioral states were Markovian, then σ_t depends only on σ_{t-1}, and, because $\sigma^2 = 1$, the only possible stationary probability distribution for the sequences $\sigma_1, \sigma_2, \cdots, \sigma_T$ is

$$P(\{\sigma_t\}) = \frac{1}{Z} \exp\left[h \sum_t \sigma_t + J \sum_t \sigma_{t-1}\sigma_t\right], \qquad [18]$$

which is the one-dimensional Ising model with nearest-neighbor interactions. Importantly, if we measure the correlations of the fluctuations in behavioral state around its mean,

$$C(t-t') \equiv \langle (\sigma_t - \langle\sigma\rangle)(\sigma_{t'} - \langle\sigma\rangle)\rangle, \qquad [19]$$

we find that these correlations decay exponentially,

$$C(t-t') = C(0)e^{-|t-t'|/\tau_c}, \qquad [20]$$

where we can express τ_c in terms of h and J (41). This reminds us of the exponential decays in the continuous case with Gaussian fluctuations.

Suppose that we have only two states, but observe correlations that do not decay as a single exponential. Then the probability distribution $P(\{\sigma_t\})$ must have terms that describe explicit dependences of σ_t on $\sigma_{t'}$ with $t - t' > 1$. This can be true only if there are some hidden states or variables that carry memory across the temporal gap $t - t'$. A sensible definition for the dimensionality of behavior then refers to these internal variables.

Imagine that we observe the mean of the behavioral variable, $\langle\sigma\rangle$, and the correlation function $C(t-t')$. What can we say about the probability distribution $P(\{\sigma_t\})$? There are infinitely many models that are consistent with measurements of just these (two-point) correlations, but there is one that stands out as having the minimal structure required to match these observations (42). Said another way, there is a unique model that predicts the observed correlations but otherwise generates behavioral sequences that are as random as possible. This minimally structured model is the one that has the largest possible entropy, and it has the form

$$P\left(\{\sigma_t\}\right) = \frac{1}{Z}\exp\left[h\sum_t \sigma_t + \frac{1}{2}\sum_{t,t'} J(t-t')\sigma_t\sigma_{t'}\right], \quad [21]$$

where the parameter h must be adjusted so that the model predicts the observed mean behavior $\langle\sigma\rangle$, and the function $J(t-t')$ must be adjusted so that the model predicts the observed correlation function $C(t-t')$.

Maximum entropy models have a long history, and a deep connection to statistical mechanics (42). As applied to temporal sequences, the maximum entropy models sometimes are referred to as maximum caliber (43). For biological systems, there has been interest in the use of maximum entropy methods to describe amino acid sequence variation in protein families (44–46), patterns of electrical activity in populations of neurons (47–51), velocity fluctuations in flocks of birds (52, 53), and more. There have been more limited attempts to use these ideas in describing temporal sequences, in neural populations (54) and in flocks (55–57).

To connect with the previous discussion, for continuous variables, a Gaussian process is the maximum entropy model consistent with the measured (two-point) correlations. In particular, if correlations decay as a combination of exponentials, then, in discrete time, the relevant Gaussian model has maximum entropy consistent with correlations among a finite number of neighboring time points. These models can also be written as autoregressive processes (58).

The maximum entropy model in Eq. **21** can be rewritten exactly as a model in which the behavioral state at time t depends only on some internal variable $x(t)$. As explained in *Appendix C*, $x(t)$ is not Gaussian, but the only coupling of past and future, again, is through a kernel $K(t)$. This kernel is not the inverse of the observed behavioral correlations but of the effective interactions between states at different times, $J(\tau)$. But, importantly, we are considering quantities that are determined by the correlation function, and hence the problem is conceptually similar to the Gaussian case: We analyze the correlations to derive a kernel, and the dimensionality of behavior is the rank of this kernel. The maximum entropy model plays a useful role because it is the least structured model consistent with the observed correlations.

If $x(t)$ is one dimensional in the sense defined above, then the interactions decay over some fixed time scale, $J(t) \sim J_0 e^{-|t|/\tau}$, and, at long times, the correlations also will decay exponentially. At the opposite extreme, if $x(t)$ has effectively infinite dimensionality, then we can have $J(t) \approx J_0|t|^{-\alpha}$. Ising models with such power-law interactions are the subject of a large literature in statistical physics; the richest behaviors are at $\alpha = 2$, where results presaged major developments in the renormalization group and topological phase transitions (59–62). It would be fascinating if these models emerged as effective descriptions of strongly non-Markovian sequences in animal behavior, as suggested recently (63).

Generalization

In both the continuous Gaussian case and the discrete case, dimensionality can be measured through the problem of prediction. To make this more general, consider observations of behavior in a time window $-T < t < T$; for simplicity, I will keep the notation $x(t)$ for the behavioral trajectory. Within each window, the trajectory $x(t<0)$ defines the past $\mathbf{x}_\mathrm{p}$, $x(t>0)$ defines the future $\mathbf{x}_\mathrm{f}$, and these are drawn from the joint probability distribution $P_T(\mathbf{x}_\mathrm{p}, \mathbf{x}_\mathrm{f})$. To characterize the possibility of making predictions, we can measure the mutual information between past and future,

$$I(\mathbf{x}_\mathrm{past}; \mathbf{x}_\mathrm{fut}) = \sum_{\mathbf{x}_\mathrm{p},\mathbf{x}_\mathrm{f}} P_T(\mathbf{x}_\mathrm{p}, \mathbf{x}_\mathrm{f}) \log\left[\frac{P_T(\mathbf{x}_\mathrm{p}, \mathbf{x}_\mathrm{f})}{P_T(\mathbf{x}_\mathrm{p})P_T(\mathbf{x}_\mathrm{f})}\right]. \quad [22]$$

This "predictive information" $I_\mathrm{pred}(T)$ can have very different qualitative behaviors as T becomes large (64).

For a time series that can be captured by a finite-state Markov process, or more generally described by a finite correlation time, then $I_\mathrm{pred}(T)$ is finite as $T \to \infty$. On the other hand, for Gaussian processes with correlation functions that decay as a power, as in Eq. **17**, the predictive information diverges logarithmically, $I_\mathrm{pred}(T \to \infty) \propto \log T$, and similarly for discrete time series with power-law correlations.[‡]

In the example of a dynamical system with D variables, as in Eq. **16**, all the predictive power available will be realized if we can specify D numbers, which are the initial conditions for integrating the differential equations. Thus we consider smooth mappings of the past into d features,

$$\mathcal{M}_d : \mathbf{x}_\mathrm{past} \to \{F_\mu\}, \quad \mu = 1, 2, \cdots, d. \quad [23]$$

For any choice of features, we can compute how much predictive information has been captured, and then we can maximize over the mapping, resulting in

$$I_\mathrm{pred}(T; d) = \max_{\mathcal{M}_d} I(\{F_\mu\}; \mathbf{x}_\mathrm{fut}), \quad [24]$$

which is the maximum predictive information we can capture with d features in windows of duration T.

If the system truly is D dimensional, then D features of the past are sufficient to capture all of the available predictive information. This means that a plot of $I_\mathrm{pred}(T; d)$ vs. d will saturate. To be precise, we are interested in what happens at large T, so we can define

$$\lim_{T\to\infty} \frac{I_\mathrm{pred}(T; d)}{I_\mathrm{pred}(T)} = f(d). \quad [25]$$

If $f(d \geq D) = 1$, then we can write the analog of Eq. **14**,

$$P[\mathbf{x}_\mathrm{f}|\mathbf{x}_\mathrm{p}] = P[\mathbf{x}_\mathrm{f}|\{F_\mu\}], \quad [26]$$

where the features F_μ now are more complex functions of the past. But there are only D of these features needed to make Eq. **26** true ($\mu = 1, 2, \cdots, D$), and so we conclude that the behavior has dimensionality D.

The equivalence of Eq. **26** to Eq. **14** immediately tells us that the general information theoretic definition of dimensionality agrees with the definition for Gaussian processes based on the spectrum of K_pf. In the Gaussian case, we see that the features F_μ are just linearly filtered versions of the past, as in Eq. **15**. The connection to the discussion of two-state variables is a bit more complicated, and exploits the equivalence to an internal or latent variable as described in *Appendix C*.

[‡] If we observe a continuous variable in continuous time, then smoothness generates a formal divergence in the mutual information between past and future. Modern analyses of behavior typically begin with video data, with time in discrete frames, evading this problem. Alternatively, if measurements include a small amount of white noise, then the predictive information becomes finite even without discrete time steps. Thanks go to A. Frishman for emphasizing the need for care here.

Conclusion

The arguments here define the dimensionality of behavior as the minimum number of features of the past needed to make the maximally informative predictions about the future. As we consider pasts of longer duration, the dimensionality can grow, potentially without bound. The connection between dimensionality and prediction is familiar from the now classical literature on dynamical systems, which also reminds us that, in its most general form, any such definition runs into all the well-known difficulties of estimating dimensions from finite data (65). More useful is the result that, in some cases, estimating this predictive dimensionality reduces to analyzing the spectrum of a matrix.

Data Availability. There are no data underlying this work.

Appendix

A. Past–Future Kernels, Explicitly. We are interested in the behavior of the kernel K when the correlation function C is a sum of exponentials. As noted above, we need to be a little careful to make this problem well posed. If we monitor a continuous variable in continuous time, then continuity leads to infinite mutual information between $x(t^-)$ and $x(t^+)$. We can solve this either by assuming that observations are made at discrete ticks of a clock (as in video recordings) or by assuming that observations are made in a background of white noise. Here I will take the second approach.

The statement that the correlation function is a sum of exponentials, but measurements are in a background of white noise, means that the observed correlation function

$$\langle x(t)x(0)\rangle \equiv C(t) = \sum_{\mu=1}^{M} A_\mu e^{-|t|/\tau_\mu} + \mathcal{N}\delta(t), \tag{27}$$

where $\mathcal{N}$ is the strength of the noise. We want to construct the kernel $K(t)$ that is the operator inverse to C, as in Eq. **10**. We recall that this can be done by passing to Fourier space,

$$G(\omega) = \int dt\, e^{+i\omega t} C(t) \tag{28}$$

$$K(t) = \int \frac{d\omega}{2\pi} e^{-i\omega t} \frac{1}{G(\omega)}. \tag{29}$$

From Eq. **27**, we can see that

$$G(\omega) = \int dt\, e^{+i\omega t}\left[\sum_{\mu=1}^{M} A_\mu e^{-|t|/\tau_\mu} + \mathcal{N}\delta(t)\right] \tag{30}$$

$$= \sum_{\mu=1}^{M} \frac{2A_\mu\tau_\mu}{1+(\omega\tau_\mu)^2} + \mathcal{N}. \tag{31}$$

Then, to find $K(t)$, we invert and transform back, being careful to isolate the contribution of the white noise term,

$$K(t) = \int \frac{d\omega}{2\pi} e^{-i\omega t}\left[\sum_{\mu=1}^{M} \frac{2A_\mu\tau_\mu}{1+(\omega\tau_\mu)^2} + \mathcal{N}\right]^{-1} \tag{32}$$

$$= \int \frac{d\omega}{2\pi} e^{-i\omega t}\left[\frac{1}{\mathcal{N}} - \frac{P_{M-1}(\omega^2)}{P_M(\omega^2)}\right], \tag{33}$$

where

$$P_{M-1}(\omega^2) = \sum_{\mu=1}^{M} 2A_\mu\tau_\mu \prod_{\nu\neq\mu}[1+(\omega\tau_\nu)^2] \tag{34}$$

is a $M-1$ st-order polynomial in ω^2, and

$$P_M(\omega^2) = \mathcal{N}\left(\mathcal{N}\prod_{\mu=1}^{M}[1+(\omega\tau_\mu)^2] + P_{M-1}(\omega^2)\right) \tag{35}$$

is a Mth-order polynomial in ω^2. Note that both polynomials have all real and positive coefficients.

We notice that $P_{M-1}(\omega^2)/P_M(\omega^2)$ vanishes at large $|\omega|$, and $e^{-i\omega t}$ vanishes for values of ω with a large negative (positive) imaginary part if $t>0$ ($t<0$). This means that we can do the integral over ω in Eq. **33** by closing a contour in the complex plane. Then we can use the fact that

$$P_M(\omega^2) = B\prod_{\rm n=1}^{M}(\omega^2 - \omega_{\rm n}^2), \tag{36}$$

where B is a constant and $\{\omega_{\rm n}^2\}$ are the roots of the polynomial. The simplest case is where all $\omega_{\rm n}^2$ are real, in which case they must be negative, and we can write $\omega_{\rm n} = -i\lambda_{\rm n}$, with $\lambda_{\rm n} > 0$. Then, for $t>0$, we close the contour in the lower half plane, picking out the poles at $\omega = \omega_{\rm n}$, while, for $t<0$, we close the contour in the upper half plane, picking out the poles at $\omega = -\omega_{\rm n}$. The result is that

$$K(t) = \frac{1}{\mathcal{N}}\delta(t) - \int \frac{d\omega}{2\pi} e^{-i\omega t}\frac{P_{M-1}(\omega^2)}{P_M(\omega^2)} \tag{37}$$

$$= \frac{1}{\mathcal{N}}\delta(t) - \frac{1}{B}\sum_{\rm n}\frac{P_{M-1}(\omega^2=-\lambda_{\rm n}^2)}{2\lambda_{\rm n}\prod_{\rm m\neq n}(\lambda_{\rm m}^2-\lambda_{\rm n}^2)} e^{-\lambda_{\rm n}|t|}. \tag{38}$$

If we look back at the derivation of Eq. **12**, we can see that a delta function term in $K(t)$ does not contribute to coupling past and future. Thus $K(t>0)$ collapses into the form of Eq. **13**,

$$K(t+t') = \sum_{\rm n=1}^{M} a_{\rm n}\phi_{\rm n}(t)\phi_{\rm n}(t') \tag{39}$$

$$a_{\rm n} = \frac{1}{B}\frac{P_{M-1}(\omega^2=-\lambda_{\rm n}^2)}{2\lambda_{\rm n}\prod_{\rm m\neq n}(\lambda_{\rm m}^2-\lambda_{\rm n}^2)} \tag{40}$$

$$\phi_{\rm n}(t) = e^{-\lambda_{\rm n}t}, \tag{41}$$

and the dimensionality $D = M$, as we hoped: If the observed behavioral variable is Gaussian, and the correlation function can be written as the sum of M exponentials, then the system has underlying dimensionality $D = M$.
It is useful to work out the case $M=1$. Then we have

$$\langle x(t)x(0)\rangle \equiv C(t) = Ae^{-|t|/\tau_c} + \mathcal{N}\delta(t), \tag{42}$$

And, after some algebra, we find

$$K(t) = a_1 e^{-\lambda_1|t|} \tag{43}$$

$$\lambda_1 = \frac{1}{\tau_c}\sqrt{1+2A\tau_c/\mathcal{N}}. \tag{44}$$

It is useful to think more explicitly about the fact that we have embedded a correlated signal in the background of white noise, so we can write

$$x(t) = y(t) + \eta(t) \tag{45}$$

$$\langle y(t)y(t')\rangle = Ae^{-|t-t'|/\tau_c} \tag{46}$$

$$\langle \eta(t)\eta(t')\rangle = \mathcal{N}\delta(t-t'). \tag{47}$$

Only $y(t)$ is predictable; the best predictions would be based on knowledge of $y(t=0)$. One can then show that the best estimate of this quantity, given observations on the noisy $x(t)$, is

$$y_{\mathrm{est}}(0) = \int_0^\infty dt\, K(t) x(-t), \tag{48}$$

with the same $K(t)$ as in Eq. **43**. Thus, asking for the optimal prediction is the same as asking for the optimal separation of the predictable signal from the unpredictable noise (66).

B. Details for Fig. 2. To generate Fig. 2, I start with some assumed correlation function $C(\tau) \equiv \langle x(t)x(t+\tau)\rangle$, sampled at discrete times $t_{\mathrm{n}} = \mathrm{n}\Delta t$. This defines a correlation matrix $C_{\mathrm{nm}} = C(t_{\mathrm{n}} - t_{\mathrm{m}})$, and then the kernel is the inverse of this matrix. One row of the matrix $K_{\mathrm{nm}} = (C^{-1})_{\mathrm{nm}}$ provides a sampled version of the function $K(t-t')$, from which we can construct $K_{\mathrm{pf}} = K(t+t')$ from Eq. **12**. Note that $\tilde{K}_{\mathrm{nm}} = K(t_{\mathrm{n}} + t_{\mathrm{m}})$ is a symmetric matrix, and, if we normalize the functions $\phi_n(t)$, then the coefficients a_{n} in Eq. **13** are the eigenvalues of this matrix; these eigenvalues are plotted in Fig. 2.

As an example that should illustrate a finite dimensionality, consider

$$C_3(\tau) = \frac{1}{3}\left[e^{-|\tau|/(4\Delta t)} + e^{-|\tau|/(32\Delta t)} + e^{-|\tau|/(256\Delta t)}\right]. \tag{49}$$

To be a bit more realistic, I add measurement noise with an amplitude of 10%, independent in each time bin, so that $C_{\mathrm{nm}} \rightarrow (1/1.01)C_{\mathrm{nm}} + (0.01/1.01)\delta_{\mathrm{nm}}$. Then the matrix K_{nm} is computed by inverting C_{nm} in a window of $T = 4,000\Delta t$. The symmetric $\tilde{K}_{\mathrm{nm}}$ is constructed in windows of $T = 100\Delta t$ or $T = 1000\Delta t$, and then diagonalized to find the eigenvalues. As an example that could illustrate infinite dimensionality, I consider the power-law correlation function in Eq. **17**, with $t_0 = \Delta t$ and $\alpha = 1/2$, then follow the same procedure as with C_3.

C. Interactions vs Latent Variables. Models where the observed degrees of freedom depend on hidden or latent variables, but not directly on one another, are sometimes set in opposition to statistical physics models, where it is more natural to think about direct interactions. But, as explained also in the supplementary material of ref. 67, this dichotomy is incorrect. In fact, interacting models can be rewritten as models where individual element responds independently to some hidden or latent variables. As an example, the maximum entropy model in Eq. **21** can be rewritten exactly as a model in which the behavioral state at time *t* depends only on some internal variable $x(t)$,

$$P\left(\{\sigma_t\}\right) = \int Dx\, P[x(t)] \prod_t P(\sigma_t | x(t) + h), \tag{50}$$

$$P(\sigma|x+h) = \frac{\exp\left[\sigma \cdot (x+h)\right]}{2\cosh(x+h)}, \tag{51}$$

and the distribution of the internal variable is

$$P[x(t)] = \frac{1}{Z'} e^{-S'[x(t)]} \tag{52}$$

$$S'[x(t)] = \frac{1}{2}\sum_{t,t'} x(t)K(t-t')x(t') - \sum_t \ln\cosh\left(x(t)+h\right),$$

where $K(t)$ is the matrix inverse of the function $J(t)$,

$$\sum_{t''} K(t-t'')J(t''-t') = \delta_{tt'}. \tag{53}$$

ACKNOWLEDGMENTS. This paper is based, in part, on a presentation at the Physics of Behavior Virtual Workshop (30 April 2020), organized by G. J. Berman and G. J. Stephens. Videos are available at https://www.youtube.com/watch?v=xSwWAgp2VdU. My sincere thanks go to Gordon, Greg, and all my fellow panelists for this wonderful event, especially in such difficult times. Thanks go to V. Alba, G. J. Berman, X. Chen, A. Frishman, K. Krishnamurthy, A. M. Leifer, C. W. Lynn, S. E. Palmer, J. W. Shaevitz, and G. J. Stephens for many helpful discussions. This work was supported, in part, by the NSF through the Center for the Physics of Biological Function (Award PHY-1734030) and Grant PHY-1607612, and by the NIH (Grant NS104889).

1. G. J. Stephens, B. Johnson-Kerner, W. Bialek, W. S. Ryu, Dimensionality and dynamics in the behavior of *C. elegans*. *PLOS Comput. Biol.* **4**, e1000028 (2008).
2. K. Branson, A. A. Robie, J. Bender, P. Perona, M. H. Dickinson, High-throughput ethomics in large groups of *Drosophila*. *Nat. Methods* **6**, 451–457 (2009).
3. G. J. Berman, D. M. Choi, W. Bialek, J. W. Shaevitz, Mapping the stereotyped behaviour of freely moving fruit flies. *J. R. Soc. Interface* **11**, 20146072 (2014).
4. A. B. Wiltschko *et al.*, Mapping sub–second structure in mouse behavior. *Neuron* **88**, 1121–1135 (2015).
5. A. Mathis *et al.*, DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289 (2018).
6. T. D. Pereira *et al.*, Fast animal pose estimation using deep neural networks. *Nat. Methods* **16**, 117–125 (2019).
7. S. R. Datta, D. J. Anderson, K. Branson, P. Perona, A. Leifer, Computational neuroethology: A call to action. *Neuron* **104**, 11–24 (2019).
8. M. W. Mathis, A. Mathis, Deep learning tools for the measurement of animal behavior in neuroscience. *Curr. Opin. Neurobiol.* **60**, 1–11 (2020).
9. A. d'Avella, E. Bizzi, Low dimensionality of supraspinally induced force fields. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 7711–7714 (1998).
10. M. Santello, M. Flanders, J. F. Soechting, Postural hand synergies for tool use. *J. Neurosci.* **18**, 10105–10115 (1998).
11. T. D. Sanger, Human arm movements described by a low-dimensional superposition of principal components. *J. Neurosci.* **20**, 1066–1072 (2000).
12. L. C. Osborne, S. G. Lisberger, W. Bialek, A sensory source for motor variation. *Nature* **437**, 412–416 (2005).
13. G. J. Stephens, M. Bueno de Mesquita, W. S. Ryu, W. Bialek, Emergence of long timescales and stereotyped behaviors in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 7286–7289 (2011).
14. T. Ahamed, A. C. Costa, G. J. Stephens, Capturing the continuous complexity of behavior in *C elegans*. *Nat. Phys.* **17**, 275–283 (2021).
15. N. H. Packard, J. P. Crutchfield, J. D. Farmer, R. S. Shaw, Geometry from a time series. *Phys. Rev. Lett.* **45**, 712–716 (1980).
16. H. D. I. Abarbanel, R. Brown, J. J. Sidorowich, L. S. Tsimring, The analysis of observed chaotic data in physical systems. *Rev. Mod. Phys.* **65**, 1331–1392 (1993).
17. S. Kato *et al.*, Global brain dynamics embed the motor command sequence of *Caenorhabditis elegans*. *Cell* **163**, 656–669 (2015).
18. A. L. A. Nichols, T. Eichler, R. Latham, M. Zimmer, A global brain state underlies *C. elegans* sleep behavior. *Science* **356**, eaam6851 (2017).
19. B. M. Yu *et al.*, Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *J. Neurophysiol.* **102**, 614–635 (2009).
20. J. A. Gallego, M. G. Perich, L. E. Miller, S. A. Solla, Neural manifolds for the control of movement. *Neuron* **94**, 978–984 (2017).
21. J. J. Hopfield, Kinetic proofreading: A new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proc. Natl. Acad. Sci. U.S.A.* **71**, 4135–4139 (1974).
22. J. Ninio, Kinetic amplification of enzyme discrimination. *Biochimie* **57**, 587–595 (1975).
23. D. M. Green, J. A. Swets, *Signal Detection Theory and Psychophysics* (Wiley, New York, 1966).
24. M. A. Bouman, "History and present status of quantum theory in vision" in *Sensory Communication*, W. Rosenblith, Ed. (MIT Press, Cambridge, MA, 1961), pp. 377–401.
25. X. I. I. Lord Rayleigh, On our perception of sound direction. *Philos. Mag. Series 6* **13**, 214–232 (1907).
26. K. von Frisch, Decoding the language of the bee. *Science* **185**, 663–668 (1974).
27. M. Giurfa, M. G. de Brito Sanchez, Black lives matter: Revisiting Charles Henry Turner's experiments on honey bee color vision. *Curr. Biol.* **30**, R1235–R1239 (2020).
28. J. L. Gould, *Ethology: The Mechanisms and Evolution of Behavior* (W. W. Norton, New York, 1982).
29. S. Hecht, S. Shlaer, M. H. Pirenne, Energy, quanta and vision. *J. Gen. Physiol.* **25**, 819–840 (1942).
30. W. Bialek, *Biophysics: Searching for Principles* (Princeton University Press, Princeton, NJ, 2012).
31. J. L. Lawson, G. E. Uhlenbeck, *Threshold Signals* (MIT Radiation Laboratory Series, McGraw–Hill, New York, 1950), vol. 24.
32. A. H. Sturtevant, The linear arrangement of six sex–linked factors in *Drosophila*, as shown by their mode of association. *J. Exp. Zool.* **14**, 43–59 (1913).
33. A. L. Hodgkin, A. F. Huxley, A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* **117**, 500–544 (1952).
34. P. Fatt, B. Katz, Spontaneous subthreshold activity at motor nerve endings. *J. Physiol.* **117**, 109–128 (1952).
35. G. Zweig, "Origins of the quark model" in *Proceedings of the Fourth International Conference on Baryon Resonances*, N. Isgur, Ed. (University of Toronto, 1980), pp. 439–479.
36. G. Zweig, Memories of Murray and the quark model. *Int. J. Mod. Phys. A* **25**, 3863–3877 (2010).
37. G. Zweig, Concrete quarks: The beginning of the end. *EPJ Web Conf.* **71**, 00146 (2014).
38. B. W. Brunton, M. M. Botvinick, C. D. Brody, Rats and humans can optimally accumulate evidence for decision-making. *Science* **340**, 95–98 (2013).

39. M. Potters, J.-P. Bouchaid, *A First Course in Random Matrix Theory for Physicists, Engineers and Data Scientists* (Cambridge University Press, Cambridge, 2020).
40. H. C. Berg, D. A. Brown, Chemotaxis in *Escherichia coli* analysed by three-dimensional tracking. *Nature* **239**, 500–504 (1972).
41. C. J. Thompson, *Mathematical Statistical Mechanics* (Princeton University Press, Princeton, NJ, 1972).
42. E. T. Jaynes, Information theory and statistical mechanics. *Phys. Rev.* **106**, 620–630 (1957).
43. P. D. Dixit *et al.*, Perspective: Maximum caliber is a general variational principle for dynamical systems. *J. Chem. Phys.* **148**, 010901 (2018).
44. W. Bialek, R. Ranganathan, Rediscovering the power of pairwise interactions. *arXiv* [Preprint] (2007). https://doi.org/10.48550/arXiv.0712.4397.
45. M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, T. Hwa, Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 67–72 (2009).
46. D. S. Marks *et al.*, Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6**, e28766 (2011).
47. E. Schneidman, M. J. Berry 2nd, R. Segev, W. Bialek, Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **440**, 1007–1012 (2006).
48. J. Shlens *et al.*, The structure of large-scale synchronized firing in primate retina. *J. Neurosci.* **29**, 5022–5031 (2009).
49. E. Granot-Atedgi, G. Tkačik, R. Segev, E. Schneidman, Stimulus-dependent maximum entropy models of neural population codes. *PLOS Comput. Biol.* **9**, e1002922 (2013).
50. G. Tkačik *et al.*, Searching for collective behavior in a large network of sensory neurons. *PLOS Comput. Biol.* **10**, e1003408 (2014).
51. L. Meshulam, J. L. Gauthier, C. D. Brody, D. W. Tank, W. Bialek, Collective behavior of place and non-place neurons in the hippocampal network. *Neuron* **96**, 1178–1191.e4 (2017).
52. W. Bialek *et al.*, Statistical mechanics for natural flocks of birds. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 4786–4791 (2012).
53. W. Bialek *et al.*, Social interactions dominate speed control in poising natural flocks near criticality. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 7212–7217 (2014).
54. T. Mora, S. Deny, O. Marre, Dynamical criticality in the collective activity of a population of retinal neurons. *Phys. Rev. Lett.* **114**, 078105 (2015).
55. A. Cavagna *et al.*, Dynamical maximum entropy approach to flocking. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **89**, 042707 (2014).
56. T. Mora *et al.*, Local equilibrium in bird flocks. *Nat. Phys.* **12**, 1153–1157 (2016).
57. F. Ferretti, V. Chardes, T. Mora, A. M. Walczak, I. Giardina, Building general Langevin models from discrete data sets. *Phys. Rev. X* **10**, 031018 (2020).
58. J. Burg, "*Maximum entropy spectral analysis*," PhD dissertation, Stanford University, Stanford, CA (1975).
59. D. Ruelle, Statistical mechanics of a one-dimensional lattice gas. *Commun. Math. Phys.* **9**, 267–278 (1968).
60. F. J. Dyson, Existence of a phase-transition in a one-dimensional Ising ferromagnet. *Commun. Math. Phys.* **12**, 91–107 (1969).
61. G. Yuval, P. W. Anderson, Exact results for the Kondo problem: One-body theory and extension to finite temperature. *Phys Rev B* **1**, 1522–1528 (1970).
62. G. Yuval, P. W. Anderson, D. R. Hamman, Exact results for the Kondo problem. II. Scaling theory, qualitatively correct solution, and some new results on one-dimensional classical statistical models. *Phys Rev B* **1**, 4464–4473 (1970).
63. V. Alba, G. J. Berman, W. Bialek, J. W. Shaevitz, Exploring a strongly non-Markovian animal behavior. *arXiv* [Preprint] (2020).
64. W. Bialek, I. Nemenman, N. Tishby, Predictability, complexity, and learning. *Neural Comput.* **13**, 2409–2463 (2001).
65. D. Ruelle, Deterministic chaos: The science and the fiction. *Proc. R. Soc. Lond. A Math. Phys. Sci.* **427**, 241–248 (1990).
66. W. Bialek, R. R. de Ruyter van Steveninck, N. Tishby, Efficient representation as a design principle for neural coding and computation. *arXiv* [Preprint] (2007). https://doi.org/10.48550/arXiv.2012.15681.
67. G. Tkačik *et al.*, Thermodynamics and signatures of criticality in a network of neurons. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 11508–11513 (2015).