Emergence of local irreversibility in complex interacting systems

Christopher W. Lynn,[1,2] Caroline M. Holmes,[2] William Bialek,[1,2] and David J. Schwab[1]

[1] *Initiative for the Theoretical Sciences, Graduate Center, City University of New York, New York, NY 10016*

[2] *Joseph Henry Laboratories of Physics and Lewis–Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544*

(Dated: March 4, 2022)

Living systems are fundamentally irreversible, breaking detailed balance and establishing an arrow of time. But how does the evident arrow of time for a whole system arise from the interactions among its multiple elements? We show that the local evidence for the arrow of time, which is the entropy production for thermodynamic systems, can be decomposed. First, it can be split into two components: an independent term reflecting the dynamics of individual elements and an interaction term driven by the dependencies among elements. Adapting tools from non–equilibrium physics, we further decompose the interaction term into contributions from pairs of elements, triplets, and higher–order terms. We illustrate our methods on models of cellular sensing and logical computations, as well as on patterns of neural activity in the retina as it responds to visual inputs. We find that neural activity can define the arrow of time even when the visual inputs do not, and that the dominant contribution to this breaking of detailed balance comes from interactions among pairs of neurons.

I. INTRODUCTION

Living systems consume energy in order to maintain order and function. Being away from equilibrium, we expect that their microscopic dynamics violate detailed balance. Macroscopically, their behaviors define an arrow of time. Despite recent progress in non–equilibrium statistical physics [1–3], there remain basic questions about how irreversibility at one scale emerges from collective dynamics at the scale below. To what extent does the irreversibility of a system arise from interactions between elements, rather than the independent dynamics of the elements themselves? Can simple dynamics involving pairs or triplets of elements build upon one another to generate large–scale irreversibility, thereby defining a macroscopic arrow of time, or do complex biological systems depend on higher–order combinatorial interactions?

To answer these questions, we propose a framework for decomposing the local evidence for the arrow of time in systems with many degrees of freedom. We demonstrate that the local irreversibility can be divided into two non–negative components: one that reflects the independent irreversibilities of the individual elements, and another that reflects the irreversibility due to interactions between elements. We then show that the interaction term can be further decomposed into contributions from groups of elements of different sizes, from pairs of elements to triplets to complex higher–order terms. In this way, one can determine not only whether the arrow of time arises from the dependencies between elements, but also the specific scale at which it emerges [4]. This decomposition is similar in spirit to the idea of connected correlations in the decomposition of the entropy itself [5].

We apply our methods to investigate the arrow of time in neural activity. Our visual perception is built out of the patterns of electrical activity of cells in the retina, and evidence for the arrow of time must be found in these patterns. Recent experiments that record the activity of many retinal neurons simultaneously [6, 7] make it possible for us to estimate all the relevant quantities directly, without introducing any model assumptions, in groups of up to five cells. We find that roughly two–thirds of these groups exhibit significant irreversibility, even when the movies shown to the retina are completely reversible. Thus, collective neural activity can define an arrow of time even when the visual inputs do not. Moreover, across distinct stimulus ensembles, we consistently find that the local irreversibility is dominated by the dynamics of neuron pairs. Together, these results demonstrate that neuronal populations can define an arrow of time that (i) emerges primarily from pairwise dynamics and (ii) does not merely reflect the irreversibility of the stimulus.

The paper is organized as follows. In Sec. II, we define the local irreversibility and multipartite dynamics. In Sec. III, we show analytically that the local irreversibility of a multipartite system can be split into two non–negative terms, the first stemming from the independent elements and the second arising from the interactions between elements. In Sec. IV, we compare these independent and interaction irreversibilities in a simple model sensing system. In Sec. V, we show that the irreversibility due to interactions can be further decomposed into a series of contributions from pairs of elements, triplets, and higher–order terms. In Sec. VI, we illustrate this decomposition using a minimal model of logical computations. In Sec. VII, we apply the above methods to investigate the irreversibility of neuronal dynamics in the vertebrate retina. Finally, in Sec. VIII, we provide conclusions and outlook, highlighting directions for future work.

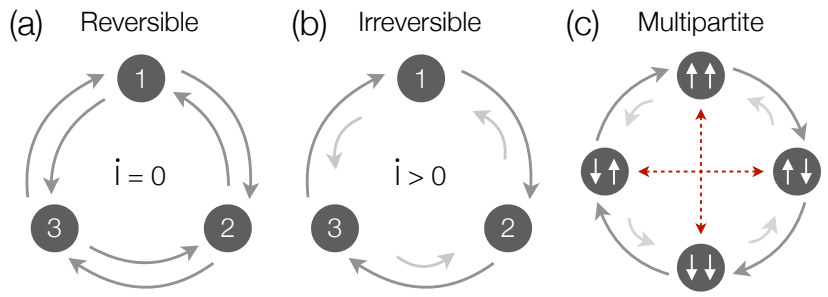

FIG. 1. Irreversibility and multipartite dynamics. (a–b) A simple three–state system, with states x represented as circles and joint transition probabilities $P(x \to x')$ as arrows. (a) In a reversible system, there are no net fluxes of transitions between states, the dynamics obey detailed balance, and there is no evidence for the arrow of time. (b) Irreversible systems exhibit net fluxes between states, thereby breaking detailed balance and establishing an arrow of time. (c) A multipartite system composed of two binary spins. Only one spin is allowed to change state at a time, thus disallowing the transitions indicated by red arrows.

II. LOCAL IRREVERSIBILITY AND MULTIPARTITE DYNAMICS

When a system is reversible, its dynamics obey detailed balance, and there are no net fluxes between states [Fig. 1(a)]. By contrast, for an irreversible system, fluxes from one state to another break detailed balance [Fig. 1(b)]. Critically, such irreversible dynamics establish an arrow of time: Just by observing the evolution of the system, one can distinguish whether time is flowing forward or backward.

To quantify irreversibility, consider a system with *joint* transition probabilities

$$P(x \to x') \equiv \text{Prob}[x_t = x,\, x_{t+1} = x'], \tag{1}$$

where x_t is the state of the system at time t. In words, this is the probability of observing the state x followed by the state x', and should not be confused with the *conditional* transition probabilities $\text{Prob}[x_{t+1} = x' \,|\, x_t = x]$. The evidence that these dynamics carry about the arrow of time is quantified by the relative entropy, or Kullback–Leibler divergence, between the forward– and reverse–time transition probabilities [8],

$$\dot{I} = \sum_{x,x'} P(x \to x') \log \left[\frac{P(x \to x')}{P(x' \to x)} \right], \tag{2}$$

where if we choose base two for the logarithms then the evidence is measured in bits. If a system obeys detailed balance, such that $P(x \to x') = P(x' \to x)$ for all pairs of states x and x', then this local irreversibility vanishes [Fig. 1(a)]. Conversely, any violation of detailed balance, such that $P(x \to x') \neq P(x' \to x)$, leads to an increase in the local irreversibility [Fig. 1(b)].

For Markov systems, the transition probabilities $P(x \to x')$ completely define the dynamics, and so $\dot{I}$ captures all available information about the arrow of time. Notably, if the states x and x' include all of the microscopic degrees of freedom in a system, then, because the laws of physics are Markovian, Eq. (2) defines the physical rate at which the system produces entropy [9]. In general, if we don't observe all the relevant degrees of freedom then the dynamics of the observable states x will be non–Markovian, but $\dot{I}$ still has a precise meaning: it represents the *local* evidence for the arrow of time.

We are interested in systems where the overall state x consists of states $\{x_\text{i}\}$ for many interacting elements, $\text{i} = 1, 2, \ldots, N$. Given sufficient temporal resolution, no two elements will change state at exactly the same time. In this limit, the dynamics are defined by the joint probabilities $P(x_\text{i} \to x'_\text{i},\, x_{-\text{i}})$ of one element i transitioning from x_i to x'_i and the rest of the system remaining in the same state, denoted $x_{-\text{i}}$ [Fig. 1(c)]. Such dynamics, which are referred to as multipartite, exhibit a number of useful properties [10, 11]. Chief among these properties is the fact that the local irreversibility simplifies to a sum over the individual elements:

$$\dot{I} = \sum_{x,x'} P(x \to x') \log \left[\frac{P(x \to x')}{P(x' \to x)} \right] \tag{3}$$

$$= \sum_{x} \sum_{\text{i}=1}^{N} \sum_{x'_\text{i}} P(x_\text{i} \to x'_\text{i},\, x_{-\text{i}}) \log \left[\frac{P(x_\text{i} \to x'_\text{i},\, x_{-\text{i}})}{P(x'_\text{i} \to x_\text{i},\, x_{-\text{i}})} \right] \tag{4}$$

$$= \sum_{\text{i}=1}^{N} \sum_{x_{-\text{i}}} \sum_{x_\text{i}, x'_\text{i}} P(x_\text{i} \to x'_\text{i},\, x_{-\text{i}}) \log \left[\frac{P(x_\text{i} \to x'_\text{i},\, x_{-\text{i}})}{P(x'_\text{i} \to x_\text{i},\, x_{-\text{i}})} \right] \tag{5}$$

$$= \sum_{\text{i}=1}^{N} \dot{I}_\text{i}, \tag{6}$$

where

$$\dot{I}_\text{i} = \sum_{x_{-\text{i}}} \sum_{x_\text{i}, x'_\text{i}} P(x_\text{i} \to x'_\text{i},\, x_{-\text{i}}) \log \left[\frac{P(x_\text{i} \to x'_\text{i},\, x_{-\text{i}})}{P(x'_\text{i} \to x_\text{i},\, x_{-\text{i}})} \right] \tag{7}$$

is the local irreversibility associated with element i.

III. INDEPENDENT AND INTERACTION IRREVERSIBILITY

We are now prepared to investigate the impact of interactions between elements on the irreversibility of a system. To begin, consider a hypothetical system in which the elements do not interact. In this case, the transitions of each element i are completely defined by the marginal transition probabilities

$$P(x_{\mathrm{i}} \to x'_{\mathrm{i}}) = \sum_{x_{-\mathrm{i}}} P(x_{\mathrm{i}} \to x'_{\mathrm{i}},\, x_{-\mathrm{i}}), \tag{8}$$

and thus the *independent* irreversibility of element i is given by

$$\dot{I}_{\mathrm{i}}^{\mathrm{ind}} = \sum_{x_{\mathrm{i}}, x'_{\mathrm{i}}} P(x_{\mathrm{i}} \to x'_{\mathrm{i}}) \log \left[\frac{P(x_{\mathrm{i}} \to x'_{\mathrm{i}})}{P(x'_{\mathrm{i}} \to x_{\mathrm{i}})} \right]. \tag{9}$$

How does this independent irreversibility compare to the true irreversibility in Eq. (7)? To answer this question, we consider the difference $\dot{I}_{\mathrm{i}} - \dot{I}_{\mathrm{i}}^{\mathrm{ind}}$, which reflects the local irreversibility of element i due to interactions with the rest of the system. Notably, we find that this difference—which we refer to as the *interaction* irreversibility $\dot{I}_{\mathrm{i}}^{\mathrm{int}}$ of element i—is itself an average of KL divergences,

$$\dot{I}_{\mathrm{i}}^{\mathrm{int}} = \dot{I}_{\mathrm{i}} - \dot{I}_{\mathrm{i}}^{\mathrm{ind}} = \sum_{x_{-\mathrm{i}}} \sum_{x_{\mathrm{i}}, x'_{\mathrm{i}}} P(x_{\mathrm{i}} \to x'_{\mathrm{i}},\, x_{-\mathrm{i}}) \left(\log \left[\frac{P(x_{\mathrm{i}} \to x'_{\mathrm{i}},\, x_{-\mathrm{i}})}{P(x'_{\mathrm{i}} \to x_{\mathrm{i}},\, x_{-\mathrm{i}})} \right] - \log \left[\frac{P(x_{\mathrm{i}} \to x'_{\mathrm{i}})}{P(x'_{\mathrm{i}} \to x_{\mathrm{i}})} \right] \right) \tag{10}$$

$$= \sum_{x_{-\mathrm{i}}} \sum_{x_{\mathrm{i}}, x'_{\mathrm{i}}} P(x_{\mathrm{i}} \to x'_{\mathrm{i}},\, x_{-\mathrm{i}}) \log \left[\frac{P(x_{-\mathrm{i}} \,|\, x_{\mathrm{i}} \to x'_{\mathrm{i}})}{P(x_{-\mathrm{i}} \,|\, x'_{\mathrm{i}} \to x_{\mathrm{i}})} \right] \tag{11}$$

$$= \sum_{x_{\mathrm{i}}, x'_{\mathrm{i}}} P(x_{\mathrm{i}} \to x'_{\mathrm{i}}) D_{KL} \left[P(x_{-\mathrm{i}} \,|\, x_{\mathrm{i}} \to x'_{\mathrm{i}}) \,||\, P(x_{-\mathrm{i}} \,|\, x'_{\mathrm{i}} \to x_{\mathrm{i}}) \right], \tag{12}$$

where $P(x_{-\mathrm{i}} \,|\, x_{\mathrm{i}} \to x'_{\mathrm{i}})$ is the conditional probability of the state $x_{-\mathrm{i}}$ of the rest of the system given a transition $x_{\mathrm{i}} \to x'_{\mathrm{i}}$ in element i.

Equation (12) immediately tells us that $\dot{I}_{\mathrm{i}}^{\mathrm{int}} \geq 0$, thereby establishing that the presence of interactions can only increase the local irreversibility of a system. Moreover, the interaction irreversibility $\dot{I}_{\mathrm{i}}^{\mathrm{int}}$ of element i admits an insightful information–theoretic interpretation: it is the amount of information that one gains about the state $x_{-\mathrm{i}}$ of the rest of the system by observing the forward–time dynamics of element i rather than the reverse–time dynamics [8]. Thus, if i's forward– and reverse–time dynamics contain the same information about the rest of the system, then interactions with element i do not contribute to the arrow of time ($\dot{I}_{\mathrm{i}}^{\mathrm{int}} = 0$), and all of i's local irreversibility arises from independent dynamics ($\dot{I}_{\mathrm{i}} = \dot{I}_{\mathrm{i}}^{\mathrm{ind}}$). Importantly, we note that Eqs. (10–12) require multipartite dynamics; if multiple elements can change state at once, then the interaction irreversibility $\dot{I}^{\mathrm{int}}$ is ill–defined (see Appendix A).

Together, Eqs. (9–12) establish our first result: that the local irreversibility of a system can be split into two *non–negative* components,

$$\dot{I} = \dot{I}^{\mathrm{ind}} + \dot{I}^{\mathrm{int}}, \tag{13}$$

where $\dot{I}^{\mathrm{ind}} = \sum_{\mathrm{i}=1}^{N} \dot{I}_{\mathrm{i}}^{\mathrm{ind}}$ is the independent irreversibility of the system (reflecting the local irreversibilities of the individual elements) and $\dot{I}^{\mathrm{int}} = \sum_{\mathrm{i}=1}^{N} \dot{I}_{\mathrm{i}}^{\mathrm{int}}$ is the interaction irreversibility (reflecting the local irreversibility due to the dependencies between elements).

IV. DECOMPOSING IRREVERSIBILITY IN A SENSING SYSTEM

To illustrate the decomposition in Eq. (13), we examine a sensing system, wherein a sensing variable y attempts to copy an environmental variable x [Fig. 2(a)]. Such sensing networks have been a topic of significant focus in non–equilibrium statistical mechanics [3, 10, 12–14], revealing the thermodynamic costs of simple computations in living systems [3, 12, 14–17].

Here, we consider an environmental variable x with three states and dynamics defined by

$$P(x' \,|\, x) = \begin{array}{c} \\ \\ \end{array} \begin{array}{c} x' \\ \end{array} \begin{pmatrix} \frac{1}{2}(1-p_x) & p_x & \frac{1}{2}(1-p_x) \\ \frac{1}{2}(1-p_x) & \frac{1}{2}(1-p_x) & p_x \\ p_x & \frac{1}{2}(1-p_x) & \frac{1}{2}(1-p_x) \end{pmatrix} \begin{array}{c} \\ \\ x \end{array}, \tag{14}$$

where p_x is the probability of x increasing from one state to the next [Fig. 2(a), left]. Meanwhile, the dynamics of

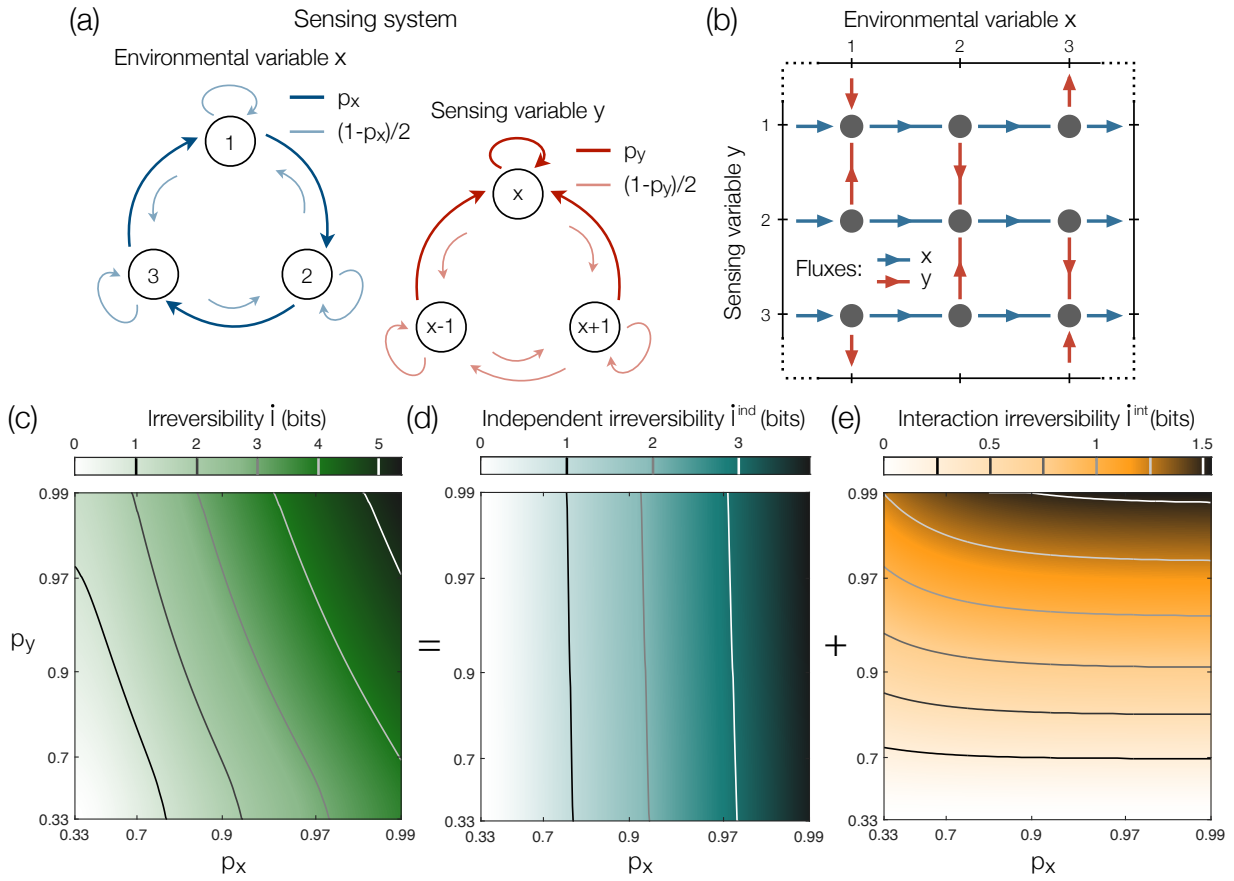

FIG. 2. Independent and interaction irreversibility in a sensing system. (a) Sensing system with an environmental variable x (left) and a sensing variable y (right), each with three states. At each point in time, one of the two variables is updated at random. With probability p_x, the environmental variable x rotates clockwise, and with probability p_y, the sensing variable y copies x. (b) Fluxes between states of the sensing system (for $p_x, p_y > 1/3$) induced by the environmental variable (blue) or the sensing variable (red). (c–d) Irreversibility $\dot{I}$ (c), independent irreversibility $\dot{I}^{\rm ind}$ (d), and interaction irreversibility $\dot{I}^{\rm int}$ (e) as functions of p_x and p_y. While $\dot{I}$ grows with both p_x and p_y, $\dot{I}^{\rm ind}$ only increases with p_x, and $\dot{I}^{\rm int}$ mostly increases with p_y, thereby distinguishing the two sources of irreversibility in the system.

the sensing variable y are given by

$$P(y' \,|\, x) = \begin{array}{c} \\ x \\ \\ \end{array} \overset{\overset{\qquad y'}{\qquad}}{\begin{pmatrix} p_y & \frac{1}{2}(1-p_y) & \frac{1}{2}(1-p_y) \\ \frac{1}{2}(1-p_y) & p_y & \frac{1}{2}(1-p_y) \\ \frac{1}{2}(1-p_y) & \frac{1}{2}(1-p_y) & p_y \end{pmatrix}}, \tag{15}$$

where p_y is the probability that y copies x [Fig. (2(a), right]. Randomly picking one variable to update at a time, one can solve for the joint transition probabilities of the combined system $P(x, y \to x', y')$; for details see Appendix B. Notably, since the dynamics are Markovian, $\dot{I}$ reflects the full (not just local) irreversibility of the system.

If $p_x = p_y = 1/3$, then both variables behave randomly, and the system obeys detailed balance. By contrast, if p_x or $p_y > 1/3$, then the tendencies for x to increase and y to copy x give rise to fluxes between the states of the system [Fig. 2(b)], thereby breaking detailed balance. Indeed, the irreversibility $\dot{I}$ increases with both p_x and p_y [Fig. 2(c)]. The independent irreversibility $\dot{I}^{\rm ind}$, however, only increases with p_x, capturing the quickening dynamics of x [Fig. 2(d)]. Meanwhile, the interaction irreversibility $\dot{I}^{\rm int}$ primarily increases with p_y, capturing the strengthening dependence of y on x [Fig. 2(e)]. We therefore find that the independent irreversibility is generated by the individual motion of the environmental variable, while the interaction irreversibility arises predominantly from the dependence of the sensing variable on the environment. In this way, the decomposition in Eq. (13) reveals the distinct ways that the environmental and sensing variables generate irreversibility.

V. IRREVERSIBILITY DUE TO $k^{\rm th}$–ORDER DYNAMICS

Can we tell whether the arrow of time emerges from the dynamics of two or three elements at a time, or whether we require higher–order information about the system

as a whole? Answering this question requires further decomposing the local irreversibility into contributions from pairs of elements, triplets, and so on. For now, consider the marginal dynamics of pairs of elements i and j; namely, the marginal transition probabilities

$$P(x_{\mathrm{i}} \to x'_{\mathrm{i}},\, x_{\mathrm{j}}) = \sum_{x_{-\{\mathrm{i},\mathrm{j}\}}} P(x_{\mathrm{i}} \to x'_{\mathrm{i}},\, x_{-\mathrm{i}}) \tag{16}$$

$$P(x_{\mathrm{j}} \to x'_{\mathrm{j}},\, x_{\mathrm{i}}) = \sum_{x_{-\{\mathrm{i},\mathrm{j}\}}} P(x_{\mathrm{j}} \to x'_{\mathrm{j}},\, x_{-\mathrm{j}}). \tag{17}$$

Imagine a hypothetical system that matches these marginal dynamics for all pairs i and j, but otherwise contains minimal information about the arrow of time, so that the dynamics are maximally reversible. This minimal irreversibility, which we denote $\dot{I}^{(2)}$, sets a lower bound on the true local irreversibility $\dot{I}$, capturing all of the local irreversibility in pairs of elements and nothing more. In this way, by casting our decomposition as an optimization problem, we are able to directly translate knowledge about a system into a lower bound on its irreversibility. From a practical perspective, the local irreversibility $\dot{I}$ is convex (see Appendix C), and so there exist efficient algorithms for computing global minima. In fact, the equivalent problem of minimizing entropy production has garnered significant attention in non–equilibrium physics [11, 15, 18, 19], dating back to the foundational work of Onsager and Prigogine [20, 21].

In general, one can compute the minimum irreversibility $\dot{I}^{(k)}$ consistent with the dynamics of k elements at a time. Since these k^{th}–order dynamics contain all of the information about smaller groups of size $1, 2, \ldots, k-1$, the minimum irreversibilities $\dot{I}^{(k)}$ form a hierarchy of lower bounds that increase toward the true local irreversiblity $\dot{I}$:

$$0 \leq \dot{I}^{(1)} \leq \dot{I}^{(2)} \leq \ldots \leq \dot{I}^{(N-1)} \leq \dot{I}^{(N)} = \dot{I}, \tag{18}$$

where N is the size of the system. There are several things to note about these inequalities. First, for thermodynamic systems, the zeroth–order bound ($\dot{I} \geq 0$) is the second law of thermodynamics, which follows from the fact that $\dot{I}$ is a KL divergence without any knowledge of the system dynamics. Second, as one might suspect, the first–order irreversibility $\dot{I}^{(1)}$—that is, the minimum irreversibility consistent with individual dynamics—is equivalent to the independent irreversibility $\dot{I}^{\mathrm{ind}}$ (see Appendix D). Finally, since the N^{th}–order dynamics contain a full description of the transition probabilities $P(x_{\mathrm{i}} \to x'_{\mathrm{i}},\, x_{-\mathrm{i}})$, we have $\dot{I}^{(N)} = \dot{I}$.

Inspecting the hierarchy in Eq. (18), we see that the local irreversibility due to k^{th}–order dynamics alone can be captured captured by the difference $\dot{I}^{(k)}_{\mathrm{int}} = \dot{I}^{(k)} - \dot{I}^{(k-1)} \geq 0$, which we refer to as the interaction irreversibility of order k. Indeed, combining these contributions from $\dot{I}^{(1)}_{\mathrm{int}} = \dot{I}^{(1)} = \dot{I}^{\mathrm{ind}}$ to $\dot{I}^{(N)}_{\mathrm{int}}$, we arrive at a full decomposition of the local irreversibility:

$$\dot{I} = \underbrace{\dot{I}^{(1)}_{\mathrm{int}}}_{\dot{I}^{\mathrm{ind}}} + \underbrace{\dot{I}^{(2)}_{\mathrm{int}} + \dot{I}^{(3)}_{\mathrm{int}} + \ldots + \dot{I}^{(N)}_{\mathrm{int}}}_{\dot{I}^{\mathrm{int}}}, \tag{19}$$

which is our main contribution. We note that this decomposition is in many ways similar to the decomposition of the entropy itself into connected components [5].

VI. DECOMPOSING IRREVERSIBILITY OF LOGICAL FUNCTIONS

To illustrate how irreversibility arises from dynamics of different orders, we apply the decomposition in Eq. (19) to a class of noisy logical functions. Specifically, we consider binary variables x and y that change state at each time step with probability p_{flip}, and third binary variable z that is the output of a logical function with a probability of error p_{error} [Fig. 3(a); see Appendix E for a full description]. As for the sensing system in Fig. 2, because the dynamics are Markovian the local irreversibility $\dot{I}$ represents the full irreversibility of the system.

We note that binary variables in steady state, such as those considered here, cannot break detailed balance on their own (Appendix F). Thus, for binary steady–state systems, the independent irreversibility vanishes ($\dot{I}^{\mathrm{ind}} = 0$), such that the arrow of time arises entirely from the interactions between the elements, $\dot{I} = \dot{I}^{\mathrm{int}} = \dot{I}^{(2)}_{\mathrm{int}} + \ldots + \dot{I}^{(N)}_{\mathrm{int}}$. Specifically for the logical functions [Fig. 3(a)], there are only two contributions to the irreversibility: that due to pairwise dynamics $\dot{I}^{(2)}_{\mathrm{int}}$ and that due to the full triplet dynamics $\dot{I}^{(3)}_{\mathrm{int}}$.

To begin, consider a simple function where z copies either x or y while ignoring the other input [Fig. 3(b–c)]; these are binary simplifications of the sensing system in Fig. 2. As p_{error} increases—that is, as the accuracy of the function decreases—we find that the irreversibility $\dot{I}$ decreases [Fig. 3(d)]. Indeed, as p_{error} approaches $1/2$, the output z completely decouples from the inputs x and y, and the system becomes reversible ($\dot{I} = 0$). Additionally, the arrow of time vanishes if the inputs x and y are static ($p_{\mathrm{flip}} = 0$) and grows as the inputs become more dynamic [that is, as p_{flip} increases; Fig. 3(d)]. Visualizing the fluxes between states of the system, we see that the tendency for z to copy x (equivalently, y) only induces fluxes in the x–z (or y–z) plane [Fig. 3(b–c)]. Accordingly, for all values of p_{flip} and p_{error}, the irreversibility arises entirely from pairwise dynamics ($\dot{I} = \dot{I}^{(2)}_{\mathrm{int}}$), while triplet dynamics do not contribute to the irreversibility [$\dot{I}^{(3)}_{\mathrm{int}} = 0$; Fig. 3(d)].

For comparison, consider the AND and OR functions [Fig. 4(a–b)]. Just as for the copy functions (Fig. 3), the irreversibilities of AND and OR (which we note are identical) increase both with the accuracy of the system (as p_{error} decreases) and with the speed of dynamics [as

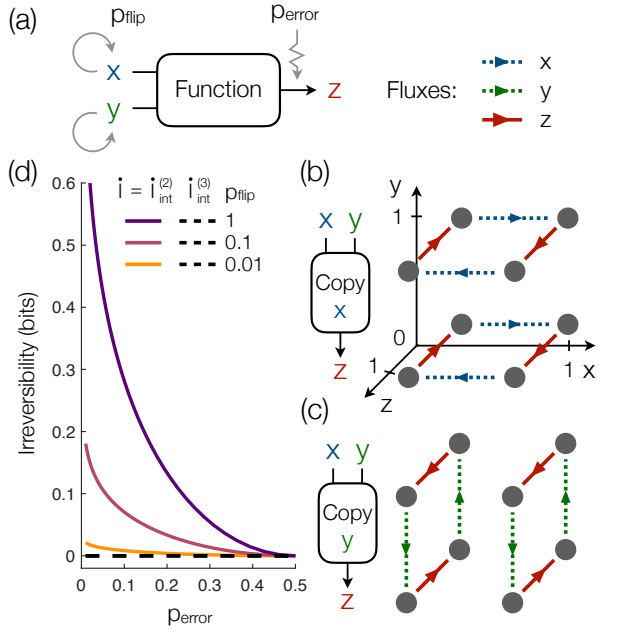

FIG. 3. Decomposing irreversibility in noisy logical functions. (a) System of three binary variables x, y, and z, where z performs a noisy logical function on the inputs x and y. At each point in time, one of the variables is updated at random. With probability p_{flip}, the inputs x and y change value, and with probability p_{error}, the output z fails to perform the specified function (see Appendix E). (b–c) Fluxes between states of the system when z either copies x (b) or copies y (c). (d) Irreversibilities $\dot{I}$ (full), $\dot{I}^{(2)}_{\text{int}}$ (pairwise interaction), and $\dot{I}^{(3)}_{\text{int}}$ (triplet interaction) versus p_{error} for different values of p_{flip}. For both of the copy functions, irreversibility arises entirely from pairwise dynamics ($\dot{I} = \dot{I}^{(2)}_{\text{int}}$), while triplet dynamics do not contribute ($\dot{I}^{(3)}_{\text{int}} = 0$).

p_{flip} increases; Fig. 4(c)]. However, in contrast to the copy functions, the full dynamics of AND and OR cannot be deduced from pairs of variables alone. Thus, the irreversibility arises from a combination of both pairwise and triplet dynamics [Fig. 4(c)]. Finally, for the XOR function, the behavior of the system only becomes apparent when all three variables are observed simultaneously [Fig. 4(d)]. As such, the irreversibility of XOR arises entirely from triplet dynamics ($\dot{I} = \dot{I}^{(3)}_{\text{int}}$), while the pairwise dynamics are completely reversible [$\dot{I}^{(2)}_{\text{int}} = 0$; Fig. 4(e)]. This is consistent with the status of XOR as the prototype of combinatorial interactions.

The results of this section are summarized in Fig. 4(f–g), where we plot the minimum irreversibilities $\dot{I}^{(k)}$ [Fig. 4(f)] and interaction irreversibilities $\dot{I}^{(k)}_{\text{int}}$ [Fig. 4(g)] of the different logical functions, normalized by the full irreversibilities $\dot{I}$. Since the systems all consist of binary steady–state dynamics, the first–order irreversibilities $\dot{I}^{(1)} = \dot{I}^{(1)}_{\text{int}}$ vanish, and therefore the independent dynamics do not define an arrow of time ($\dot{I}^{\text{ind}} = 0$). For the copy functions [Fig. 3(b–c)], irreversibility is driven

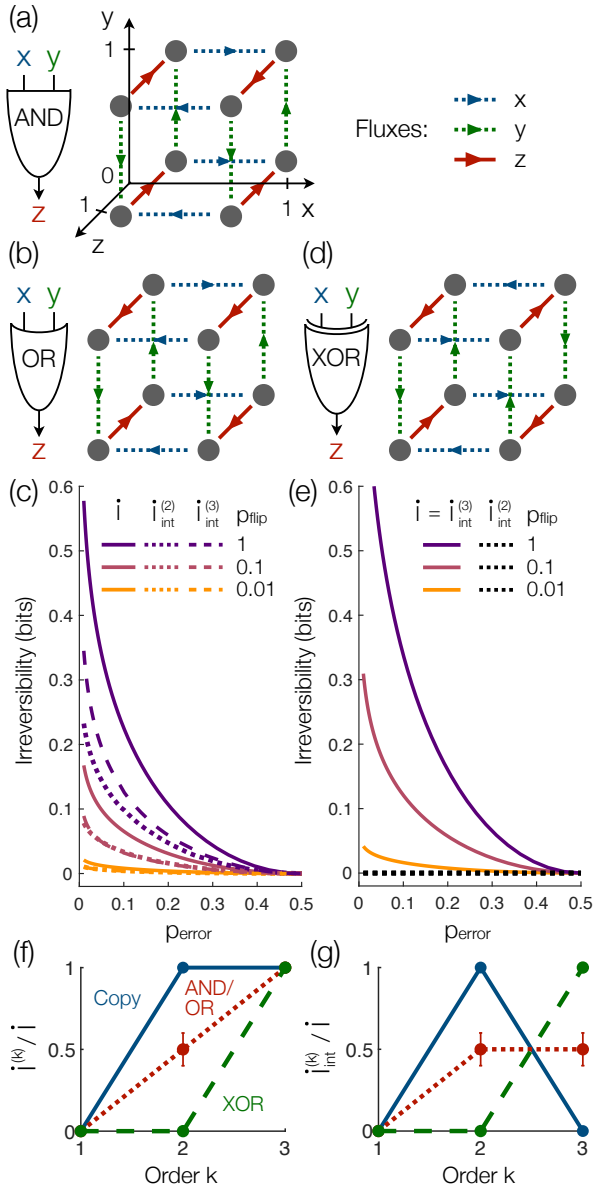

FIG. 4. Decomposing irreversibility in AND, OR, and XOR functions. (a–b) For logical systems defined as in Fig. 3(a), we illustrate the fluxes between states when z executes either AND (a) or OR (b). (c) Irreversibilities $\dot{I}$ (full), $\dot{I}^{(2)}_{\text{int}}$ (pairwise interaction), and $\dot{I}^{(3)}_{\text{int}}$ (triplet interaction) of the AND and OR systems versus p_{error} for different values of p_{flip}. Irreversibility arises from a combination of pairwise ($\dot{I}^{(2)}_{\text{int}} > 0$) and triplet ($\dot{I}^{(3)}_{\text{int}} > 0$) dynamics. (d–e) Fluxes (d) and irreversibilities (e) when z performs XOR. Irreversibility arises entirely from triplet dynamics ($\dot{I} = \dot{I}^{(3)}_{\text{int}}$), while pairwise dynamics are reversible ($\dot{I}^{(2)}_{\text{int}} = 0$). (f–g) Minimum irreversibilities $\dot{I}^{(k)}$ (f) and interaction irreversibilities $\dot{I}^{(k)}_{\text{int}}$ (g), normalized by the full irreversibility $\dot{I}$, as functions of the order k for different logical functions. The errorbars for AND and OR reflect the small variability in $\dot{I}^{(k)}/\dot{I}$ and $\dot{I}^{(k)}_{\text{int}}/\dot{I}$ over the range of different p_{error} and p_{flip} values.

entirely by second–order dynamics; for the AND and OR functions [Fig. 4(a–b)], the arrow of time arises from a combination of second– and third–order dynamics; and for the XOR function [Fig. 4(d)], irreversibility is driven entirely by third–order dynamics [see Fig. 4(g)]. In this way, the decomposition in Eq. (19) can be used to reveal the scale at which irreversibility arises in interacting systems.

VII. DECOMPOSING IRREVERSIBILITY IN NEURONAL POPULATIONS

Using the framework developed above, we are ultimately interested in understanding how irreversibility emerges in biological systems. Here, we study electrical activity in groups of neurons at the output of the retina. These ganglion cells provide all the data that the brain has about the visual world, and hence their state provides the ingredients out of which visual perceptions are synthesized, including our perception of the arrow of time. Importantly, information about visual stimuli is encoded not just in the firing of individual neurons, but also in the web of dependencies between neurons [7, 22, 23]. It remains unknown, however, whether groups of neurons exhibit fluxes between collective states—thereby breaking detailed balance—and if so, whether such irreversibility arises from pairs of neurons or from complicated higher–order dynamics.

Here we analyze experiments on the salamander retina [Fig. 5(a)], where it is possible to record form many neurons simultaneously as they respond to complex visual stimuli [6]. These experiments explored three very different kinds of visual inputs: natural movies [Fig. 5(b)], a single horizontal bar whose vertical motion is equivalent to a Brownian particle on a spring [Fig. 5(c)], and the Brownian bar with precise repetitions of the same trajectory. Although this was not the goal of the original experiments, we note that the natural movies violate time–reversal invariance, being easily recognized when played forward vs. backward, while the Brownian bar is an equilibrium system and obeys detailed balance. Appendix G gives a fuller description of the experimental setup and procedures from Ref [6].

A. Broken detailed balance in neuronal dynamics

The problems of detecting and quantifying irreversibility in data have garnered significant attention in the statistical mechanics of living systems [3, 19, 24–27]. To detect irreversibility, one must simply search for violations of detailed balance; namely, fluxes between the states of a system [3, 24, 25]. To quantify the irreversibility of a system, however, one must estimate or bound $\dot{I}$ from time–series measurements [19, 25–27]. Here, in addition to estimating the local irreversibility $\dot{I}$, we further wish

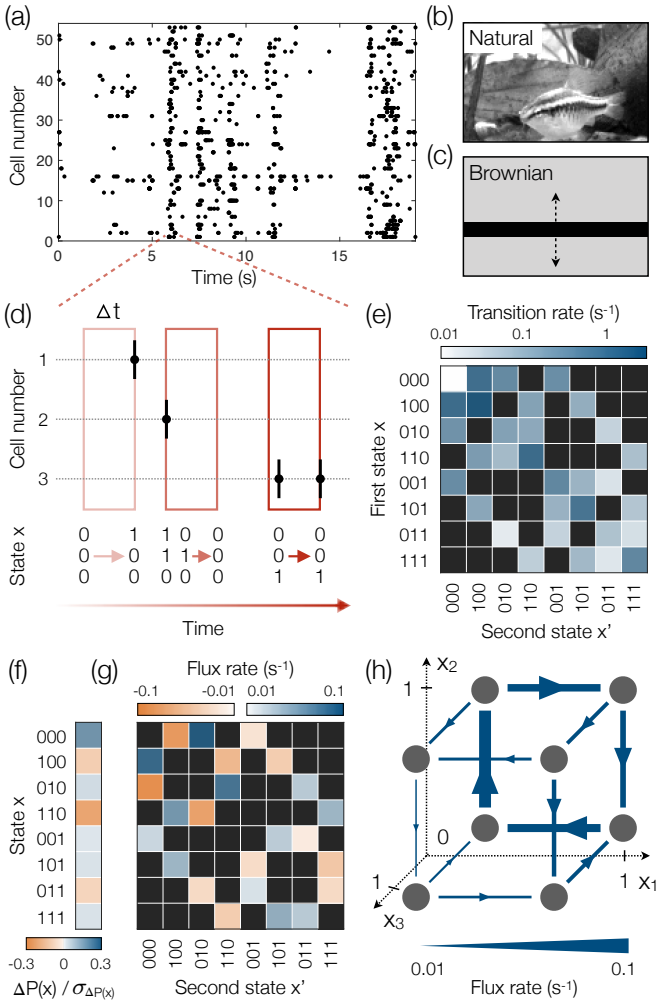

FIG. 5. Broken detailed balance in a group of neurons. (a) Dots mark the times of action potentials ("spikes") from 53 neurons in the salamander retina responding to a visual stimulus (see Appendix G for experimental details). (b–c) The same 53 neurons are exposed to three different stimuli: a natural movie of fish swimming (b), a horizontal bar whose movement is defined by a Brownian particle on a spring (c), and the same Brownian bar in panel (c) but with one trajectory repeated multiple times. (d) To record multipartite transitions [see Fig. 1(c)], we slide a window of width $\Delta t = 20\,\text{ms}$ along the time series. A neuron transitions to active when a spike enters the front of the window (left), or inactive when a spike exits the back of the window (center). Self–transitions can occur when a cell spikes twice within the same window (right). (e–h) A random group of three neurons responding to a natural movie. (e) Transition rates $P(x \to x')$ (normalized to units of transitions per second), where black entries indicate transitions that are disallowed under multipartite dynamics. (f) Changes in state probabilities $\Delta P(x)$ are small relative to their standard deviations $\sigma_{\Delta P(x)}$, indicating that the system is in steady state. (g–h) Flux rates $P(x \to x') - P(x' \to x)$ (normalized to transitions per second) illustrated as a matrix (g) and as a directed network of fluxes (h). The presence of fluxes demonstrates that the neurons break detailed balance, defining an arrow of time.

to decompose $\dot{I}$ into contributions from dynamics of various orders [as in Eq. (19)]. In order to do so—that is, in order to compute the minimum irreversibilities $I^{(k)}$ consistent with k^{th}–order dynamics—we must begin by estimating the transition probabilities $P(x_{\text{i}} \to x'_{\text{i}}, x_{-\text{i}})$ themselves.

We consider a neuron i active ($x_{\text{i}} = 1$) if it generates an action potential ("spike") at least once within a time window of width $\Delta t = 20\,\text{ms}$, or inactive ($x_{\text{i}} = 0$) if it is silent. In this way, the collective state of N neurons is a binary vector $x = \{x_1, x_2, \ldots, x_N\}$. As we slide the window along the time series, it is almost always the case that only one cell i changes state at a time, either by having a spike enter the front of the window [Fig. 5(d), left window] or exit the back of the window [Fig. 5(d), center window]. We remark that self–transitions can occur when a cell spikes twice within the same window [Fig. 5(d), right window], but note that the all–silent state $\{0, \ldots, 0\}$ cannot have a self-transition. In the rare instances when two spikes enter or exit the window at exactly the same time (within the experimental resolution of 0.1 ms), we break ties by adding small random noise to the spike times, thus yielding multipartite dynamics wherein only one cell changes state at a time.

For example, in Fig. 5(e) we illustrate the transition rates between the states of $N = 3$ neurons responding to a natural movie [Fig. 5(b)]. We note that the transition rates are proportional to the joint transition probabilities: $2rP(x \to x')$, where r is the spike rate, and the factor of two stems from the fact that there are two transitions for every spike (one when the spike enters the front of the sliding window and one when it exits the back). Notably, the changes in state probabilities $\Delta P(x) = \sum_{x'} P(x' \to x) - P(x \to x')$ are small relative to errors [Fig. 5(f)], indicating that the group of neurons is in a stochastic steady state. Additionally, we find that the cells exhibit fluxes between states [Fig. 5(g–h)], thereby breaking detailed balance. In combination, these results establish that the group of neurons operates at a non–equilibrium steady state.

B. Local irreversibility depends on stimulus

We are now prepared to estimate the collective irreversibility of groups of neurons. We note that neurons—indeed, biological systems generally—can have long–range temporal dependencies. Thus, in contrast to the Markov systems examined in previous sections (Figs. 2–4), here $\dot{I}$ reflects the local (rather than total) irreversibility of the system. As with other information–theoretic quantities, estimating the local irreversibility from data is challenging, and prone to systematic errors due to finite data. As described in Appendix H, we find that these can be controlled using the strategy of Ref. [22] if we restrict our attention to groups of no more than $N = 5$ cells.

After correcting for finite–data effects, out of 100 random 5–cell groups, across the different stimuli we find that 60–68% exhibit significant local irreversibility $\dot{I}$, thereby defining an arrow of time. Surprisingly, despite the fact that the Brownian bar is completely reversible, neuronal dynamics are more irreversible when responding to this stimulus than the natural movie [Fig. 6(a)]. Moreover, the local irreversibility is even larger when the same Brownian trajectory is repeated multiple times [Fig. 6(a)], suggesting that a repeated input can induce a stronger arrow of time in the neuronal responses. We confirm that these differences in local irreversibility hold even after accounting for variations in the overall rate of spiking across the stimulus ensembles [Fig. 6(b)]. Additionally, the same ordering of stimuli holds for all group sizes from $N = 2$ to $N = 5$ cells [Fig. 6(c–d)]. These results demonstrate that the arrow of time in neuronal activity does not simply reflect the irreversibility of the stimulus. Instead, neuronal dynamics can define an arrow of time even when the stimulus does not.

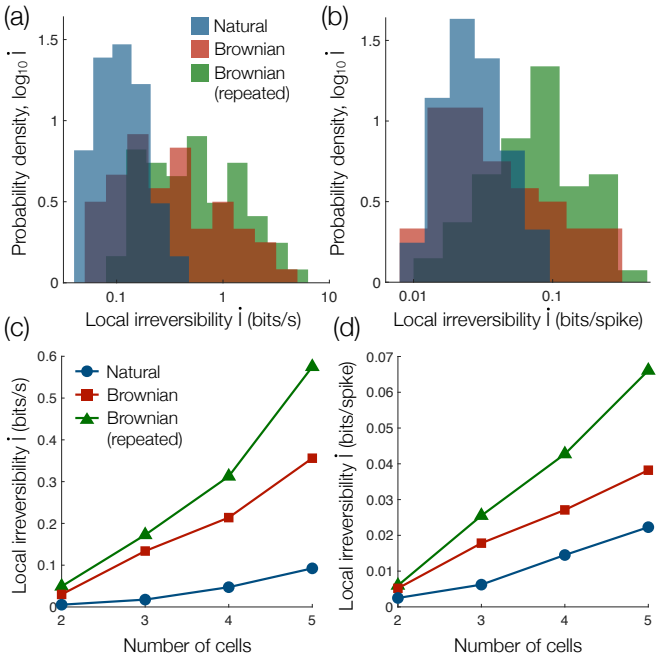

FIG. 6. Stimulus–dependence of local irreversibility. (a) Distributions of local irreversibilities $\dot{I}$ (normalized to bits per second) for 5–cell groups responding to a natural movie (blue), a Brownian bar (red), and a repeated Brownian bar (green). (b) The same as panel (a), but normalized to bits per spike to control for variations in spike rates across stimuli. In panels (a–b), out of 100 random 5–cell groups, we only include those with significant local irreversibility (see Appendix H). (c–d) Local irreversibilities per second (c) and per spike (d) for different stimuli as functions of the number of cells in a group. Data points are averaged over 100 random groups.

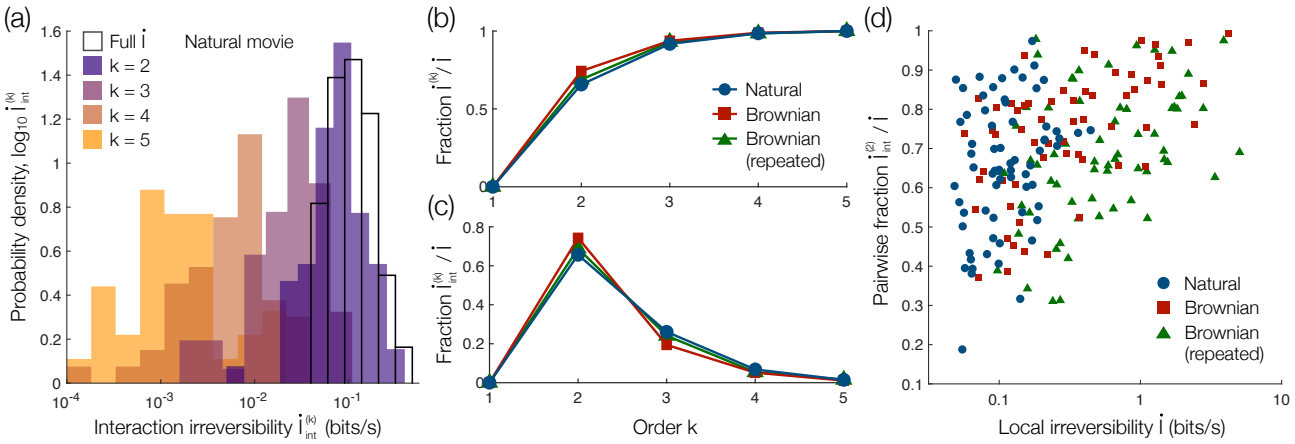

FIG. 7. Decomposing local irreversibility in neuronal activity. (a) Distributions of interaction irreversibilities $\dot{I}_{\text{int}}^{(k)}$ of different orders k (normalized to bits per second) for 5–cell groups responding to a natural movie. (b–c) Minimum irreversibilities $\dot{I}^{(k)}$ (b) and interaction irreversibilities $\dot{I}_{\text{int}}^{(k)}$ (c), normalized by the true local irreversibilities $\dot{I}$, as functions of the order k averaged over 5–cell groups. (d) Across all stimuli, the fraction of pairwise irreversibility $\dot{I}_{\text{int}}^{(2)}/\dot{I}$ increases significantly with the local irreversibility $\dot{I}$ for 5–cell groups (Spearman coefficient $r = 0.26$, $p = 0.03$ for natural movie; $r = 0.61$, $p < 0.01$ for Brownian bar; and $r = 0.39$, $p < 0.01$ for repeated Brownian bar). In all panels, out of 100 random groups, we only include those with significant local irreversibility for each stimulus.

C. Local irreversibility arises from low–order dynamics

To implement the decomposition of local irreversibility from Eq. (19), we need numerical methods to construct the probability distributions that minimize $\dot{I}$ while matching the observed k^{th}–order dynamics. We provide one such method for binary systems in Appendix I.

For groups of $N = 5$ cells responding to the natural movie, we find that pairwise dynamics account for much more of the local irreversibility than higher–order dynamics [Fig.7(a)]. In fact, across all stimuli, pairwise dynamics generate 66–74% of the local irreversibility [Fig. 7(b)], more than 3^{rd}–, 4^{th}–, and 5^{th}–order dynamics combined [Fig. 7(c)]. Moreover, the fraction of the irreversibility captured by pairwise dynamics increases significantly with the local irreversibility itself [Fig. 7(d)], demonstrating that groups of neurons that operate further from detailed balance do so in an even more pairwise fashion. Perhaps most notably, despite the fact that the magnitude of the local irreversibility varies significantly from one stimulus to another (Fig. 6), we find that the proportions of irreversibility captured by different types of dynamics remain consistent across stimuli [Fig. 7(b–c)].

In combination, the results of this section indicate that the arrow of time in retinal neurons (i) varies depending on the specific stimulus (Fig. 6), yet (ii) does not simply reflect the irreversibility of the stimulus, and (iii) consistently arises from the same combination of low–order dynamics, driven primarily by pairs of neurons (Fig. 7).

VIII. CONCLUSIONS

Irreversible dynamics support a wide range of biological functions, yet it remains unclear how macroscopic irreversibility arises from the microscopic dynamics of individual components. In this study, we propose a framework to uncover the scale at which irreversibility emerges in complex interacting systems. To do so, we develop analytic and numerical techniques for decomposing the information–theoretic evidence for the arrow of time into contributions from individual elements, pairs, and higher–order dynamics. We illustrate our methods on the examples of irreversible dynamics in models for sensing systems (Fig. 2) and logical functions (Figs. 3 and 4). Moving to real data, we find that the irreversibility of retinal neurons varies from one stimulus to another, but consistently arises from pairwise dynamics (Figs. 5–7).

These results suggest several new directions. For example, given that the irreversibility of retinal neurons does not simply reflect that of the stimulus, it is natural to wonder which stimulus properties are, in fact, responsible for inducing irreversibility in groups of cells. Additionally, in the process of decomposing the local irreversibility, one must compute a hierarchy of minimally irreversible models consistent with observed k^{th}–order dynamics. Just as maximum entropy models have been successful in describing distributions over states at single moments in time [23, 28–30], might these minimum irreversibility models provide insights into the dynamical flow of living systems from one state to another? More generally, we remark that the proposed framework is non–invasive, applying to any system with time–series data. Thus, the methods can be used to examine irreversible dynamics in a wide range of other biological sys-

tems, from molecular and cellular networks [3, 12, 14, 31–33], to large–scale recordings in the brain [25, 34], to entire populations of animals and humans [28, 30, 35].

ACKNOWLEDGMENTS

We thank SE Palmer for helpful discussions and for guiding us through the data of Ref. [7]. This work was supported in part by the National Science Foundation, through the Center for the Physics of Biological Function (PHY–1734030) and a Graduate Research Fellowship (CMH); by the National Institutes of Health through the BRAIN initiative (R01EB026943); by the James S McDonnell Foundation through a Postdoctoral Fellowship Award (CWL); by the Simons Foundation; and by a Sloan Research Fellowship (DJS).

CITATION DIVERSITY STATEMENT

Recent work in several fields of science [36–40], and physics in particular [41], has identified citation bias negatively impacting women and minorities. Here we sought to proactively consider choosing references that reflect the diversity of the field in thought, form of contribution, gender, and other factors. Excluding (including) self–citations to the current authors, our references contain 28% (27%) women lead authors and 31% (34%) women senior authors.

Appendix A: Multipartite dynamics required to decompose irreversibility

In Sec. III, we show that the local irreversibility $\dot{I}$ of a multipartite system can be decomposed into two non–negative components [Eq. (13)]: the independent irreversibility $\dot{I}^{\text{ind}}$ and the interaction irreversibility $\dot{I}^{\text{int}}$. Here we show that multipartite dynamics are necessary for this decomposition. Specifically, we establish that if multiple elements are allowed to change state at the same time, then the decomposition in Eq. (13) can break down and the interaction irreversibility can become ill–defined.

Consider, for example, a system with two identical elements x and y, such that $x = y$ at all moments in time. Given the joint transition probabilities of one of the variables $P(x \to x')$, the dynamics of the combined system are given by

$$P(x, y \to x', y') = P(x \to x')\delta_{x,y}\delta_{x',y'}. \tag{A1}$$

For such a system, the irreversibility is given by

$$\dot{I} = \sum_{x,x',y,y'} P(x, y \to x', y') \log \frac{P(x, y \to x', y')}{P(x', y' \to x, y)} \tag{A2}$$

$$= \sum_{x,x'} P(x \to x') \log \frac{P(x \to x')}{P(x' \to x)}. \tag{A3}$$

To compute the independent irreversibility $\dot{I}^{\text{ind}}$, we note that the marginal dynamics of y are identical to that of x:

$$P(y \to y') = \sum_{x,x'} P(x, y \to x', y') \tag{A4}$$

$$= \sum_{x,x'} P(x \to x')\delta_{x,y}\delta_{x',y'} \tag{A5}$$

$$= P(x \to x'). \tag{A6}$$

Thus, the independent irreversibility is given by

$$\dot{I}^{\text{ind}} = \dot{I}^{\text{ind}}_x + \dot{I}^{\text{ind}}_y \tag{A7}$$

$$= \sum_{x,x'} P(x \to x') \log \frac{P(x \to x')}{P(x' \to x)} \tag{A8}$$

$$+ \sum_{y,y'} P(y \to y') \log \frac{P(y \to y')}{P(y' \to y)} \tag{A9}$$

$$= 2 \sum_{x,x'} P(x \to x') \log \frac{P(x \to x')}{P(x' \to x)} \tag{A10}$$

$$= 2\dot{I}. \tag{A11}$$

Since $\dot{I}^{\text{int}} = \dot{I} - \dot{I}^{\text{ind}} = -\dot{I}$, we find that the interaction irreversibility is negative, thus violating the decomposition of the local irreversibility into non–negative terms.

Appendix B: Solving the sensing system

Consider a sensing system composed of an environmental variable x and a sensing variable y, each with three states. The environmental variable x increases with probability p_x, and the sensing variable y copies x with probability p_y, yielding the dynamics in Eqs. (14–15). Randomly choosing one variable to update at each point in time, the dynamics of the combined system are defined by the conditional transition probabilities

$$P(x', y' \,|\, x, y) = \frac{1}{2}\big(P(x' \,|\, x)\delta_{y,y'} + P(y' \,|\, x)\delta_{x,x'}\big). \tag{B1}$$

Using the stationary condition $\pi(x, y) = \sum_{x',y'} P(x, y \,|\, x', y')\pi(x', y')$, one can solve for the stationary distribution:

$$\pi(x, y) \propto \begin{pmatrix} 2 + 7p_y + p_x(1 + 3p_y + 3p_x) \\ 5 - 2p_y + p_x(4 - 6p_y + 3p_x) \\ 6 - 5p_y + p_x(1 + 3p_y + 3p_x) \\ 6 - 5p_y + p_x(1 + 3p_y + 3p_x) \\ 2 + 7p_y + p_x(1 + 3p_y + 3p_x) \\ 5 - 2p_y + p_x(4 - 6p_y + 3p_x) \\ 5 - 2p_y + p_x(4 - 6p_y + 3p_x) \\ 6 - 5p_y + p_x(1 + 3p_y + 3p_x) \\ 2 + 7p_y + p_x(1 + 3p_y + 3p_x) \end{pmatrix} \begin{matrix} (1,1) \\ (1,2) \\ (1,3) \\ (2,1) \\ (2,2) \\ (2,3) \\ (3,1) \\ (3,2) \\ (3,3) \end{matrix}. \tag{B2}$$

Combining Eqs. (B1) and (B2), we arrive at the joint transition probabilities $P(x, y \to x', y') = P(x', y' \,|\, x, y)\pi(x, y)$, which are used to perform the calculations in Sec. IV.

Appendix C: Convexity of local irreversibility

In order to compute $\dot{I}^{(k)}$, one must minimize the local irreversibility $\dot{I}$ subject to constraints on the k^{th}-order dynamics of the system. Here, we show that the local irreversibility is convex with respect to the transition probabilities $P(x \to x')$, and thus can be minimized using efficient techniques.

The gradient of the local irreversibility [Eq. (2)] is given by

$$\frac{\partial \dot{I}}{\partial P(x \to x')} = \log \frac{P(x \to x')}{P(x' \to x)} - \frac{P(x' \to x)}{P(x \to x')} + 1, \quad \text{(C1)}$$

where for simplicity $\log(\cdot)$ is natural logarithm. Since Eq. (C1) only depends on $P(x \to x')$ and $P(x' \to x)$, we see that the Hessian of $\dot{I}$ takes the block diagonal form

$$H = \begin{pmatrix} \ddots & 0 & 0 \\ 0 & H(x,x') & 0 \\ 0 & 0 & \ddots \end{pmatrix}, \quad \text{(C2)}$$

where

$$H(x,x') = \begin{pmatrix} \frac{\partial^2 \dot{I}}{\partial P(x\to x')^2} & \frac{\partial^2 \dot{I}}{\partial P(x\to x')\partial P(x'\to x)} \\ \frac{\partial^2 \dot{I}}{\partial P(x\to x')\partial P(x'\to x)} & \frac{\partial^2 \dot{I}}{\partial P(x'\to x)^2} \end{pmatrix} \quad \text{(C3)}$$

$$= \begin{pmatrix} \frac{P(x\to x')+P(x'\to x)}{P(x\to x')^2} & -\frac{P(x\to x')+P(x'\to x)}{P(x\to x')P(x'\to x)} \\ -\frac{P(x\to x')+P(x'\to x)}{P(x\to x')P(x'\to x)} & \frac{P(x\to x')+P(x'\to x)}{P(x'\to x)^2} \end{pmatrix} \quad \text{(C4)}$$

is the 2×2 Hessian for the pair of states (x, x'). The eigenvalues of $H(x,x')$ are $\lambda_1 = \frac{(P(x\to x')+P(x'\to x))(P(x\to x')^2+P(x'\to x)^2)}{P(x\to x')^2 P(x'\to x)^2}$ and $\lambda_2 = 0$. Since $\lambda_1, \lambda_2 \geq 0$, and since the eigenvalues of H are simply the eigenvalues of the different blocks $H(x,x')$ combined, we have established that H is positive semidefinite, and therefore that the local irreversibility $\dot{I}$ is convex.

Appendix D: Equivalence between independent and first–order irreversibilities

Here we establish that the independent irreversibility $\dot{I}^{\text{ind}}$ is equivalent to the first–order minimum irreversibility $\dot{I}^{(1)}$. To do so, consider a hypothetical system $Q(x_{\text{i}} \to x'_{\text{i}}, x_{-\text{i}})$ that is consistent with the observed first–order dynamics $P(x_{\text{i}} \to x'_{\text{i}}) = \sum_{x_{-\text{i}}} P(x_{\text{i}} \to x'_{\text{i}}, x_{-\text{i}})$. Since $\dot{I}^{\text{ind}}(Q) = \dot{I}^{\text{ind}}(P)$, we have

$$\dot{I}(Q) = \dot{I}^{\text{ind}}(Q) + \dot{I}^{\text{int}}(Q) \quad \text{(D1)}$$

$$= \dot{I}^{\text{ind}}(P) + \dot{I}^{\text{int}}(Q) \quad \text{(D2)}$$

$$\geq \dot{I}^{\text{ind}}(P), \quad \text{(D3)}$$

where the inequality follows from that fact that $\dot{I}^{\text{int}}(Q) \geq 0$. Thus, the independent irreversibility $\dot{I}^{\text{ind}}(P)$ is a lower bound on the local irreversibility $\dot{I}(Q)$ of any hypothetical system Q consistent with the observed first–order dynamics. Since the first–order irreversibility $\dot{I}^{(1)}$ is just the minimum of $\dot{I}(Q)$ among all such systems Q, we have found that $\dot{I}^{(1)} \geq \dot{I}^{\text{ind}}$.

In order to establish that $\dot{I}^{(1)} = \dot{I}^{\text{ind}}$, all that remains is to identify a hypothetical system Q that achieves the lower bound in Eqs. (D1–D3). Specifically, we seek a system Q that is consistent with the observed first–order dynamics, yet has interaction irreversibility $\dot{I}^{\text{int}}(Q) = 0$. Consider, for example, a system Q in which the dynamics of each element i are independent from the rest of the system, such that $Q(x_{\text{i}} \to x'_{\text{i}}, x_{-\text{i}}) = Q(x_{\text{i}} \to x'_{\text{i}}) = P(x_{\text{i}} \to x'_{\text{i}})$ for all $x_{-\text{i}}$. Using Eqs. (10–12), one can verify that such a system has zero interaction irreversibility, thereby saturating the lower bound in Eqs. (D1–D3). We have therefore shown that $\dot{I}^{(1)}$ (the minimum local irreversibility consistent with first–order dynamics) is equivalent to the independent irreversibility $\dot{I}^{\text{ind}}$.

Appendix E: Noisy logical functions

In Sec. VI, we examine a system of three binary variables: two inputs x and y that flip with probability p_{flip}, and an output variable z that performs a logical function on x and y, but with error rate p_{error} [see Fig. 3(a)]. Specifically, the dynamics of the input variables are defined by the conditional transition probabilities

$$P(x' \,|\, x) = P(y' \,|\, y) = \begin{pmatrix} 1-p_{\text{flip}} & p_{\text{flip}} \\ p_{\text{flip}} & 1-p_{\text{flip}} \end{pmatrix}, \quad \text{(E1)}$$

and the dynamics of the output variable are defined by

$$P(z' \,|\, x, y) = \begin{cases} 1-p_{\text{error}}, & z' = f(x,y) \\ p_{\text{error}}, & z' \neq f(x,y) \end{cases}, \quad \text{(E2)}$$

where $f(x,y)$ is the logical function performed by z. Randomly picking one variable to update at each point in time, the conditional transition probabilities for the entire system are given by

$$P(x',y',z' \,|\, x,y,z) = \frac{1}{3}\big(P(x' \,|\, x)\delta_{y,y'}\delta_{z,z'} + P(y' \,|\, y)\delta_{x,x'}\delta_{z,z'} + P(z' \,|\, x,y)\delta_{x,x'}\delta_{y,y'}\big). \quad \text{(E3)}$$

Using Eq. (E3), one can solve for the stationary distribution $\pi(x,y,z)$ and then compute the joint transition probabilities $P(x,y,z \to x',y',z') = P(x',y',z' \,|\, x,y,z)\pi(x,y,z)$.

Appendix F: Independent irreversibility vanishes for binary steady–state systems

For binary steady–state systems, such as the logical functions in Sec. VI and the neurons in Sec. VII, the

independent irreversibility $\dot{I}^{\rm ind}$ is zero. To see this, note that the marginal dynamics of any binary steady–state variable are defined by the conditional transition probabilities

$$P(x_{\rm i}' \,|\, x_{\rm i}) = \begin{pmatrix} 1-p_{\rm i} & p_{\rm i} \\ q_{\rm i} & 1-q_{\rm i} \end{pmatrix}, \tag{F1}$$

where $0 \leq p_{\rm i}, q_{\rm i} \leq 1$ are the probabilities of i switching between its two states. The marginal steady–state distribution for i is $\pi(x_{\rm i}) = \frac{1}{p_{\rm i}+q_{\rm i}}(q_{\rm i}, p_{\rm i})^T$, and thus the marginal joint transition probabilities are given by

$$P(x_{\rm i} \to x_{\rm i}') = P(x_{\rm i}' \,|\, x_{\rm i})\pi(x_{\rm i}) \tag{F2}$$

$$= \frac{1}{p_{\rm i}+q_{\rm i}} \begin{pmatrix} (1-p_{\rm i})q_{\rm i} & p_{\rm i}q_{\rm i} \\ q_{\rm i}p_{\rm i} & (1-q_{\rm i})p_{\rm i} \end{pmatrix}. \tag{F3}$$

Since the above transition probabilities are symmetric, the marginal dynamics of each element i obey detailed balance (such that $\dot{I}_{\rm i}^{\rm ind} = 0$). Thus, we find that the independent irreversibility of the entire system $\dot{I}^{\rm ind} = \sum_{\rm i=1}^{N} \dot{I}_{\rm i}^{\rm ind}$ is zero. We emphasize that this only holds for the *local* irreversibility; if we consider non–Markov effects in strings of 3, 4, or more points in time, then binary steady–state variables can break time-reversal symmetry and define an arrow of time.

Appendix G: Neuronal recordings

The neuronal data examined in Sec. VII were recorded from larval tiger salamander retina, which were dissected, perfused with Ringer's solution, and pressed onto dense arrays of 252 electrodes with 30–μm spacing, as described in Ref. [7]. In the experiments, which lasted 4–6 hours, movies were projected onto the photoreceptor layer of the retina via an objective lens, and voltages were recorded at 10 kHz. Spikes were sorted conservatively (as described in Ref. [6]), yielding 53 reliable cells from which groups were randomly selected for analysis.

The stimuli were presented on a 360×600 display, with pixels of size $3.81\,\mu$m on the retina and a frame rate of 60 frames per second. All stimuli were normalized to the same average light intensity. The natural movie depicted a fish swimming in a tank, repeated 102 times. The moving bar was 11 pixels wide and black on a gray background, with trajectories displayed 62 times. The trajectory of the bar's vertical position was generated by a stochastic process equivalent to a Brownian particle on a spring attached to the center of the display. Specifically, the vertical position x_t and velocity v_t of the bar were updated at each time t according to the equations of motion:

$$x_{t+\tau} = x_t + v_t\tau, \tag{G1}$$

and

$$v_{t+\tau} = (1-\Gamma\tau)v_t - \omega^2 x_t\tau + \xi_t\sqrt{D\tau}, \tag{G2}$$

where $\tau = 1/60\,$s is the time step (which matches the frame rate of the visual display), $\omega = 3\pi\,{\rm s}^{-1}$ is the natural frequency, $\Gamma = 20\,{\rm s}^{-1}$ parameterizes the damping (chosen such that the dynamics are slightly overdamped), and $D = 2.7\times10^6\,{\rm pixel}^2/{\rm s}^3$ is chosen to allow reasonable range of motion. For the repeated bar stimulus, the same trajectory was repeated 62 times.

Appendix H: Correcting for finite data

For time–series data, such as the neuronal spiking examined in Sec. VII, in order to estimate quantities of interest—such as transition rates, flux rates, and changes in state probabilities (Fig. 5); local irreversibilities $\dot{I}$ (Fig. 6); and interaction irreversibilities $\dot{I}_{\rm int}^{(k)}$ (Fig. 7)—one must correct for finite–data effects [7, 22, 23]. To do so, for a given stimulus and group of neurons, we begin with a list of the observed transitions $\{x(t) \to x(t+1)\}$. For a given set of data fractions f, we subsample the transitions (without replacement) in a hierarchical fashion, such that each subsample of transitions is a subset of the larger subsamples. For the neuronal data in Sec. VII, we find that data fractions $f = \{1, 0.9, 0.8, 0.7, 0.6, 0.5\}$ are sufficient.

For each data fraction f, we estimate the quantity of interest. For example, for the local irreversibility, we use the estimate

$$\dot{I} = \sum_{x,x'} \tilde{P}(x \to x') \log \frac{\tilde{P}(x \to x')}{\tilde{P}(x' \to x)}, \tag{H1}$$

where

$$\tilde{P}(x \to x') = \frac{N(x \to x') + 1}{\sum_{y,y'} \left(N(y \to y') + 1\right)} \tag{H2}$$

are the maximum likelihood probabilities with one pseudocount for each transition, and $N(x \to x')$ is the number of times that the transition $x \to x'$ was observed in the data. We include pseudocounts to avoid infinities in Eq. H1, but we confirm that the naïve estimator without pseudocounts yields the same results. After estimating the quantity of interest for all fractions f, we then extrapolate to the infinite–data limit using a linear fit with respect to the inverse data fraction $1/f$ [Fig. 8(a)]. Repeating this process 100 times, we arrive at both an average and standard deviation for the infinite–data estimates of the desired quantity [Fig. 8(b)].

To check that the above procedure gives accurate estimates for the local irreversibility $\dot{I}$, we note that randomizing the timing of spikes should destroy the arrow of time. Thus, for time–randomized data, the estimated local irreversibility should vanish in the infinite–data limit. Consider the 100 groups of $N = 5$ neurons analyzed in Figs. 6 and 7. Among these groups, after correcting for finite–data effects, we find that 60–68% exhibit significant local irreversibility $\dot{I}$, depending on the stimulus [Fig. 9(a)]. By contrast, after randomizing the

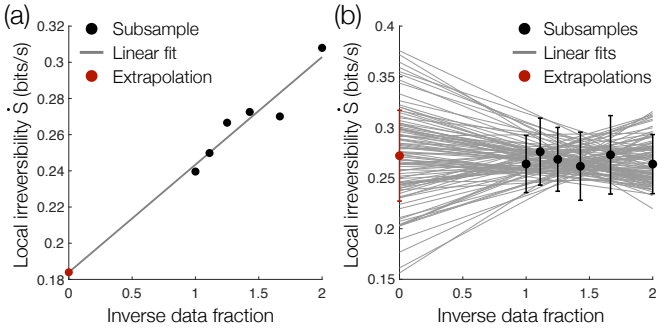

FIG. 8. Correcting for finite–data effects on local irreversibility. (a) Estimated local irreversibility $\dot{I}$ (black markers) versus inverse data fraction for one group of $N = 5$ neurons responding to a natural movie. Grey line indicates a linear fit, and red marker indicates the extrapolation to infinite data. (b) Estimated local irreversibility (black markers), linear fits (grey lines), and extrapolation to infinite data (red marker) after repeating the process in panel (a) 100 times for the same group of neurons. Data points and error bars reflect averages and standard deviations over the 100 repetitions.

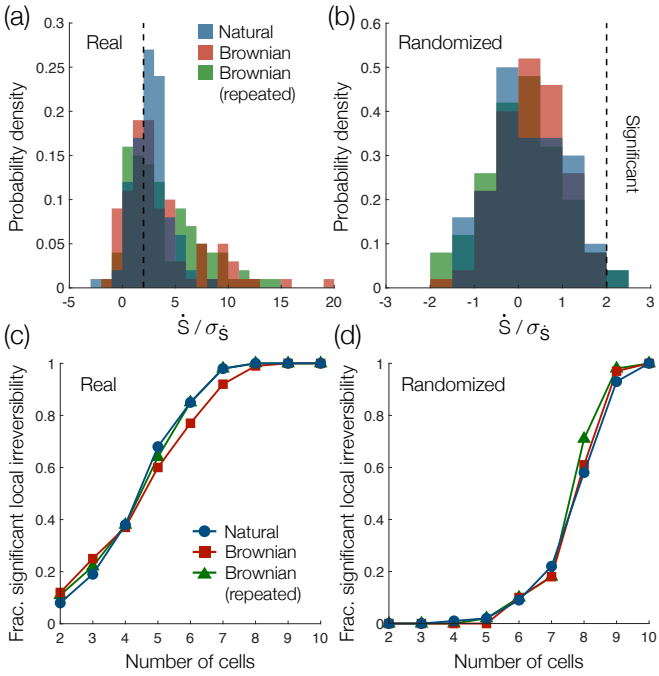

FIG. 9. Estimated local irreversibilities of real and time–shuffled data. (a–b) Distributions of local irreversibility estimates $\dot{I}$, normalized by standard deviations $\sigma_{\dot{I}}$, for real time series (a) and after randomizing spike times (b). Distributions are over the same 100 groups of $N = 5$ cells analyzed in Figs. 6 and 7, and the dashed line indicates the threshold for significance. (c–d) Fraction of cell groups with significant local irreversibility as a function of group size N for real time series (c) and after randomizing spike times (d). For each size, fractions are computed over 100 random groups.

spike times, the local irreversibility estimates are centered around zero, with only 0–2% of groups exhibiting significant local irreversibility [Fig. 9(b)]. Examining different group sizes, we find that the percentage of groups with significant local irreversibility increases from ∼10% for $N = 2$ cells to ∼100% for $N \geq 8$ cells [Fig. 9(c)]. Importantly, after randomizing spike times, groups of $N \leq 5$ cells are almost always locally reversible, as desired [Fig. 9(d)]. However, even for time–randomized data, we find that some groups of $N \geq 6$ cells exhibit significant local irreversibility [Fig. 9(d)], demonstrating finite–data effects cannot be adequately accounted for. We therefore conclude that $N = 5$ is the largest number of cells for which we can consistently estimate local irreversibility $\dot{I}$ in our dataset.

Appendix I: Minimizing local irreversibility

Computing the k^{th}–order minimum irreversibility $\dot{I}^{(k)}$ requires finding a hypothetical system $Q(x_{\text{i}} \to x'_{\text{i}}, x_{-\text{i}})$ that matches the observed k^{th}–order marginal dynamics, but otherwise has minimum local irreversibility $\dot{I}(Q)$. We remark that the local irreversibility $\dot{I}$ is convex (see Appendix C), and thus computing $\dot{I}^{(k)}$ is a constrained convex minimization problem for which there exist efficient optimization methods. From a practical perspective, there are only two main hurdles to overcome: (i) adapting an existing convex minimization technique for our problem, and (ii) writing down the constraints on the k^{th}–order dynamics. Here, we address these challenges for binary systems.

1. Frank–Wolfe algorithm

To minimize the local irreversibility $\dot{I}$ given a set of constraints, we employ the Frank–Wolfe algorithm, which efficiently converges to a local (and therefore global) minimum. Specifically, we initialize Q using any dynamics that match the observed constraints (for example, one can begin with the observed dynamics $Q = P$). We then iterate the following steps:

1. First, we compute the gradient of the local irreversibility:

$$\frac{\partial \dot{I}}{\partial Q(x_{\text{i}} \to x'_{\text{i}}, x_{-\text{i}})} = \log \frac{Q(x_{\text{i}} \to x'_{\text{i}}, x_{-\text{i}})}{Q(x'_{\text{i}} \to x_{\text{i}}, x_{-\text{i}})} - \frac{Q(x'_{\text{i}} \to x_{\text{i}}, x_{-\text{i}})}{Q(x_{\text{i}} \to x'_{\text{i}}, x_{-\text{i}})} + 1, \tag{I1}$$

   where $\log(\cdot)$ represents the natural logarithm for simplicity.

2. Second, we solve for the dynamics Q^* that obey the desired constraints while minimizing the inner

product with the gradient:

$$\sum_{\mathrm{i}=1}^{N} \sum_{x_{-\mathrm{i}}} \sum_{x_{\mathrm{i}}, x_{\mathrm{i}}'} Q(x_{\mathrm{i}} \to x_{\mathrm{i}}',\, x_{-\mathrm{i}}) \frac{\partial \dot{I}}{\partial Q(x_{\mathrm{i}} \to x_{\mathrm{i}}',\, x_{-\mathrm{i}})}. \quad \text{(I2)}$$

We note that this constrained linear minimization problem is a linear program, and thus can be efficiently solved using standard techniques (e.g., the `linprog` function in MATLAB).

3. Finally, we take a step toward Q^*, such that $Q \leftarrow Q + \alpha_t(Q^* - Q)$, where $\alpha_t = \alpha_0/t$ is the step size, which decreases with the number of iterations t.

An implementation of the above algorithm is available at `github.com/ChrisWLynn/Decompose_irreversibility`.

2. Constraining k^{th}–order dynamics

We seek to constrain the k^{th}–order dynamics of a binary, multipartite system with joint transition probabilities $P(x_{\mathrm{i}} \to x_{\mathrm{i}}',\, x_{-\mathrm{i}})$. For each element i, consider a group of $k-1$ of the remaining elements $K \subseteq \{1, \ldots, \mathrm{i}-1, \mathrm{i}+1, \ldots, N\}$. Let x_K denote the states of the elements in K, and $x_{-\{\mathrm{i},K\}}$ the states of the elements not in i nor K. The marginal dynamics of i with the elements in K held fixed are then given by

$$P(x_{\mathrm{i}} \to x_{\mathrm{i}}',\, x_K) = \sum_{x_{-\{\mathrm{i},K\}}} P(x_{\mathrm{i}} \to x_{\mathrm{i}}',\, x_{-\mathrm{i}}). \quad \text{(I3)}$$

Constraining the k^{th}–order dynamics amounts to constraining the marginal probabilities $P(x_{\mathrm{i}} \to x_{\mathrm{i}}',\, x_K)$ for all elements i and all groups of the remaining elements K of size $k-1$. For example, if K is empty, then we arrive at the independent (first–order) dynamics $P(x_{\mathrm{i}} \to x_{\mathrm{i}}')$. If K consists of one element j, then we have the pairwise (second–order) dynamics $P(x_{\mathrm{i}} \to x_{\mathrm{i}}',\, x_{\mathrm{j}})$ discussed in Sec. V. We remark, however, that these marginal probabilities are not all independent, and therefore the set of constraints is overdetermined.

To write down independent constraints that fully define the k^{th}–order dynamics, it helps to consider an analogy with Ising systems. Consider a binary system with state probabilities $P(x)$. It is known that the k^{th}–order marginal probabilities $P(x_K) = \sum_{x_{-K}} P(x)$ are completely defined by the correlations between groups of elements up to size k: $\langle x_{\mathrm{i}} \rangle$, $\langle x_{\mathrm{i}} x_{\mathrm{j}} \rangle$, …, $\left\langle \prod_{\mathrm{i} \in K} x_{\mathrm{i}} \right\rangle$, where $\langle \cdot \rangle$ represents an average over $P(x)$. Moreover, these correlations are independent, thus forming a basis for the k^{th}–order probabilities $P(x_K)$.

Here, we wish to constrain the k^{th}–order transition probabilities $P(x_{\mathrm{i}} \to x_{\mathrm{i}}',\, x_K)$ for all elements i and all groups of the remaining elements K of size $k-1$. For a given transition $x_{\mathrm{i}} \to x_{\mathrm{i}}'$, we denote the correlation between a set of the remaining elements K by

$$\left\langle \prod_{\mathrm{j} \in K} x_{\mathrm{j}} \right\rangle_{x_{\mathrm{i}} \to x_{\mathrm{i}}'} = \sum_{x_{-\mathrm{i}}} \prod_{\mathrm{j} \in K} x_{\mathrm{j}} P(x_{\mathrm{i}} \to x_{\mathrm{i}}',\, x_{-\mathrm{i}}). \quad \text{(I4)}$$

If K is empty, then we simply arrive at the independent transition probabilities:

$$\langle 1 \rangle_{x_{\mathrm{i}} \to x_{\mathrm{i}}'} = \sum_{x_{-\mathrm{i}}} P(x_{\mathrm{i}} \to x_{\mathrm{i}}',\, x_{-\mathrm{i}}) = P(x_{\mathrm{i}} \to x_{\mathrm{i}}'). \quad \text{(I5)}$$

By analogy with Ising systems, for each transition $x_{\mathrm{i}} \to x_{\mathrm{i}}'$, the k^{th}–order marginal probabilities $P(x_{\mathrm{i}} \to x_{\mathrm{i}}',\, x_K)$ can be defined by the correlations between groups of elements (not including i) from the empty set up to size $k-1$: $\langle 1 \rangle_{x_{\mathrm{i}} \to x_{\mathrm{i}}'}$, $\langle x_{\mathrm{j}} \rangle_{x_{\mathrm{i}} \to x_{\mathrm{i}}'}$, …, $\left\langle \prod_{\mathrm{j} \in K} x_{\mathrm{j}} \right\rangle_{x_{\mathrm{i}} \to x_{\mathrm{i}}'}$. We can then constrain the k^{th}–order dynamics of the entire system by computing the above correlations for each of the $2N$ transitions $x_{\mathrm{i}} \to x_{\mathrm{i}}'$ (not including self–transitions). We remark that we do not need to constrain self–transitions $x \to x$ because they do not contribute to the local irreversibility [Eq. 2]. Code for constraining the k^{th}–order dynamics of binary, multipartite systems is available at `github.com/ChrisWLynn/Decompose_irreversibility`.

Appendix J: Groups of neurons operate at steady state

In Figure 5, we see that a group of $N = 3$ neurons operates at a non–equilibrium steady state. Here, we demonstrate that steady–state dynamics are not specific just to this group, but are instead a general feature of all groups of neurons analyzed in this paper.

To determine if a system operates at steady state, one must examine whether its state probabilities are stationary in time. The change in the probability $P(x)$ of a state x during one time step is given by $\Delta P(x) = \sum_{x'} P(x' \to$

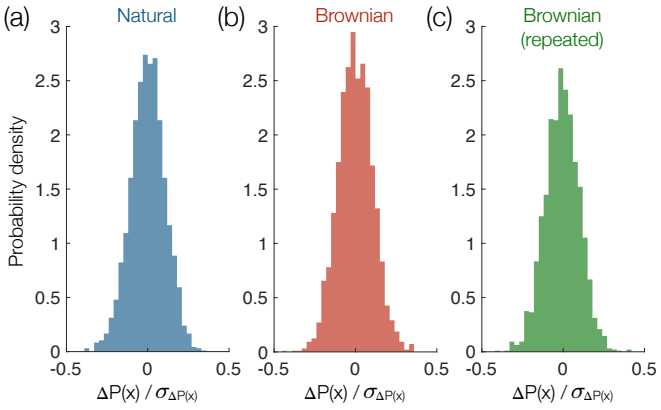

FIG. 10. Neurons operate at stochastic steady states. (a–c) Distributions of changes in state probabilities $\Delta P(x) = \sum_{x'} P(x' \to x) - P(x \to x')$, normalized by the standard deviation $\sigma_{\Delta P(x)}$, for groups of $N = 5$ cells responding to a natural movie (a), moving bar stimulus (b), and a repeated moving bar stimulus (c). Distributions are over the $2^N = 32$ different states for the 100 random groups analyzed in Figs. 6 and 7.

$x) - P(x \to x')$. In steady state this should be zero, but more precisely we expect that it will be a random number with a variance set by the errors in sampling the underlying distributions. In Fig. 10, we plot the distributions of $\Delta P(x)$, normalized by the relevant standard deviation $\sigma_{\Delta P(x)}$, for the groups of $N = 5$ cells analyzed in Figs. 6 and 7. We note that these quantities are estimated using the same finite–data correction techniques described in Appendix H. Across all stimuli, we find that the changes in state probabilities $\Delta P(x)$ for all cell groups are small relative to errors; that is, for all stimuli, all groups of neurons appear to operate at steady state.

Appendix K: Data and code availability

The data and code used to perform the analyses in this paper are openly available at `github.com/ChrisWLynn/Decompose_irreversibility`.

[1] Juan MR Parrondo, Jordan M Horowitz, and Takahiro Sagawa, "Thermodynamics of information," Nat. Phys. **11**, 131–139 (2015).

[2] Luca Peliti and Simone Pigolotti, *Stochastic Thermodynamics: An Introduction* (Princeton University Press, 2021).

[3] F S Gnesotto, Federica Mura, Jannes Gladrow, and C P Broedersz, "Broken detailed balance and non-equilibrium dynamics in living systems: A review," Rep. Prog. Phys. **81**, 066601 (2018).

[4] Christopher W Lynn, Caroline M Holmes, William Bialek, and David J Schwab, "Decomposing the local arrow of time in interacting systems," Preprint: arxiv.org/abs/2112.14721.

[5] E Schneidman, S Still, M J Berry II, and W Bialek, "Network information and connected correlations," Phys. Rev. Lett. **91**, 238701 (2003).

[6] O Marre, D Amodei, N Deshmukh, K Sadeghi, F Soo, TE Holy, and M J Berry II, "Mapping a complete neural population in the retina," J. Neurosci. **32**, 14859–14873 (2012).

[7] S E Palmer, O Marre, M J Berry II, and W Bialek, "Predictive information in a sensory population," Proc. Natl. Acad. Sci. **112**, 6908–6913 (2015).

[8] T M Cover and J A Thomas, *Elements of Information Theory* (John Wiley & Sons, 2012).

[9] Udo Seifert, "Entropy production along a stochastic trajectory and an integral fluctuation theorem," Phys. Rev. Lett. **95**, 040602 (2005).

[10] J M Horowitz and M Esposito, "Thermodynamics with continuous information flow," Phys. Rev. X **4**, 031015 (2014).

[11] D H Wolpert, "Minimal entropy production rate of interacting systems," New J. Phys. **22**, 113013 (2020).

[12] Ganhui Lan, Pablo Sartori, Silke Neumann, Victor Sourjik, and Yuhai Tu, "The energy–speed–accuracy tradeoff in sensory adaptation," Nat. Phys. **8**, 422 (2012).

[13] AC Barato, D Hartich, and U Seifert, "Information-theoretic versus thermodynamic entropy production in autonomous sensory networks," Phys. Rev. E **87**, 042104 (2013).

[14] Vudtiwat Ngampruetikorn, David J Schwab, and Greg J Stephens, "Energy consumption and cooperation for optimal sensing," Nat. Commun. **11**, 1–8 (2020).

[15] Susanne Still, "Thermodynamic cost and benefit of memory," Phys. Rev. Lett. **124**, 050601 (2020).

[16] Susanne Still, David A Sivak, Anthony J Bell, and Gavin E Crooks, "Thermodynamics of prediction," Phys. Rev. Lett. **109**, 120604 (2012).

[17] Sarah E Marzen and James P Crutchfield, "Optimized bacteria are environmental prediction engines," Phys. Rev. E **98**, 012408 (2018).

[18] J Schnakenberg, "Network theory of microscopic and macroscopic behavior of master equation systems," Rev. Mod. Phys. **48**, 571 (1976).

[19] DJ Skinner and J Dunkel, "Improved bounds on entropy production in living systems," Proc. Natl. Acad. Sci. **118** (2021).

[20] L Onsager, "Reciprocal relations in irreversible processes. I." Phys. Rev. **37**, 405 (1931).

[21] I Prigogine, "Modération et transformation irréversible des systèmes ouverts," Acad. R. Belg. Bull. Cl. Sci. **31**, 600–606 (1945).

[22] S P Strong, R Koberle, R R de Ruyter van Steveninck, and W Bialek, "Entropy and information in neural spike trains," Phys. Rev. Lett. **80**, 197 (1998).

[23] E Schneidman, M J Berry II, R Segev, and W Bialek, "Weak pairwise correlations imply strongly correlated network states in a neural population," Nature **440**, 1007–1012 (2006).

[24] C Battle, CP Broedersz, N Fakhri, VF Geyer, J Howard, CF Schmidt, and FC MacKintosh, "Broken detailed balance at mesoscopic scales in active biological systems," Science **352**, 604–607 (2016).

[25] Christopher W Lynn, E J Cornblath, L Papadopoulos, M A Bertolero, and Danielle S Bassett, "Broken detailed balance and entropy production in the human brain," Proc. Natl. Acad. Sci. **118** (2021).

[26] Ignacio A Martínez, Gili Bisker, Jordan M Horowitz, and Juan MR Parrondo, "Inferring broken detailed balance in the absence of observable currents," Nat. Commun. **10**, 1–10 (2019).

[27] Junang Li, Jordan M Horowitz, Todd R Gingrich, and Nikta Fakhri, "Quantifying dissipation using fluctuating currents," Nat. Commun. **10**, 1–9 (2019).

[28] W Bialek, A Cavagna, I Giardina, T Mora, E Silvestri, M Viale, and A M Walczak, "Statistical mechanics for natural flocks of birds," Proc. Natl. Acd. Sci. **109**, 4786–4791 (2012).

[29] L Meshulam, J L Gauthier, C D Brody, D W Tank, and W Bialek, "Collective behavior of place and non-place neurons in the hippocampal network," Neuron **96**, 1178–1191 (2017).

[30] Christopher W Lynn, Lia Papadopoulos, Daniel D Lee, and Danielle S Bassett, "Surges of collective human activity emerge from simple pairwise correlations," Phys. Rev. X **9**, 011022 (2019).

[31] Björn Stuhrmann, Marina Soares e Silva, Martin Depken, Fred C MacKintosh, and Gijsje H Koenderink, "Nonequilibrium fluctuations of a remodeling in vitro cytoskeleton," Phys. Rev. E **86**, 020901(R) (2012).

[32] Marina Soares Soares e Silva, Martin Depken, Björn Stuhrmann, Marijn Korsten, Fred C MacKintosh, and Gijsje H Koenderink, "Active multistage coarsening of actin networks driven by myosin motors," Proc. Natl. Acad. Sci. **108**, 9408–9413 (2011).

[33] Nikta Fakhri, Alok D Wessel, Charlotte Willms, Matteo Pasquali, Dieter R Klopfenstein, Frederick C MacKintosh, and Christoph F Schmidt, "High-resolution mapping of intracellular fluctuations using carbon nanotubes," Science **344**, 1031–1035 (2014).

[34] Christopher W Lynn and Danielle S Bassett, "The physics of brain network structure, function and control," Nat. Rev. Phys. **1**, 318 (2019).

[35] A Cavagna, I Giardina, F Ginelli, T Mora, D Piovani, R Tavarone, and A M Walczak, "Dynamical maximum entropy approach to flocking," Phys. Rev. E **89**, 042707 (2014).

[36] Sara McLaughlin Mitchell, Samantha Lange, and Holly Brus, "Gendered citation patterns in international relations journals," Int. Stud. Perspect. **14**, 485–492 (2013).

[37] Michelle L Dion, Jane Lawrence Sumner, and Sara McLaughlin Mitchell, "Gendered citation patterns across political science and social science methodology fields," Polit. Anal. **26**, 312–327 (2018).

[38] Neven Caplar, Sandro Tacchella, and Simon Birrer, "Quantitative evaluation of gender bias in astronomical publications from citation counts," Nat. Astron. **1**, 1–5 (2017).

[39] Jordan D Dworkin, Kristin A Linn, Erin G Teich, Perry Zurn, Russell T Shinohara, and Danielle S Bassett, "The extent and drivers of gender imbalance in neuroscience reference lists," Nat Neurosci. **23**, 918–926 (2020).

[40] Maxwell A Bertolero, Jordan D Dworkin, Sophia U David, Claudia López Lloreda, Pragya Srivastava, Jennifer Stiso, Dale Zhou, Kafui Dzirasa, Damien A Fair, Antonia N Kaczkurkin, *et al.*, "Racial and ethnic imbalance in neuroscience reference lists and intersections with gender," bioRxiv (2020).

[41] Erin G Teich, Jason Z Kim, Christopher W Lynn, Samantha C Simon, Andrei A Klishin, Karol P Szymula, Pragya Srivastava, Lee C Bassett, Perry Zurn, Jordan D Dworkin, *et al.*, "Citation inequity and gendered citation practices in contemporary physics," arXiv preprint arXiv:2112.09047 (2021).