# Thermodynamics of Prediction

Susanne Still,[1,*] David A. Sivak,[2,†] Anthony J. Bell,[3] and Gavin E. Crooks[2]

[1]*University of Hawaii at Mānoa, Information and Computer Sciences, Honolulu, Hawaii 96822, USA*
[2]*Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*
[3]*Redwood Center for Theoretical Neuroscience, University of California, Berkeley, California 94720, USA*

(Received 15 March 2012; published 19 September 2012)

A system responding to a stochastic driving signal can be interpreted as computing, by means of its dynamics, an implicit model of the environmental variables. The system's state retains information about past environmental fluctuations, and a fraction of this information is predictive of future ones. The remaining nonpredictive information reflects model complexity that does not improve predictive power, and thus represents the ineffectiveness of the model. We expose the fundamental equivalence between this model inefficiency and thermodynamic inefficiency, measured by dissipation. Our results hold arbitrarily far from thermodynamic equilibrium and are applicable to a wide range of systems, including biomolecular machines. They highlight a profound connection between the effective use of information and efficient thermodynamic operation: any system constructed to keep memory about its environment and to operate with maximal energetic efficiency has to be predictive.

DOI: 10.1103/PhysRevLett.109.120604 PACS numbers: 05.70.Ln, 89.70.−a

All systems perform computations by means of responding to their environment. In particular, living systems compute, on a variety of length and time scales, future expectations based on their prior experience. Most biological computation is fundamentally a nonequilibrium process, because a preponderance of biological machinery in its natural operation is driven far from thermodynamic equilibrium. For example, many molecular machines (such as the microtubule-associated motor kinesin) are driven by adenosine triphosphate (ATP) hydrolysis, which liberates ~500 meV per molecule [1]. This energy is large compared with ambient thermal energy, $1\ k_{\mathrm{B}}T \approx 25$ meV ($k_{\mathrm{B}}$ is Boltzmann's constant and the temperature is $T \sim 300$ K). In general, such large energetic inputs drive the operative degrees of freedom of biological machines away from equilibrium averages.

Recently, significant progress has been made in describing driven systems far from equilibrium [2], perhaps most notably Jarzynski's work relation [3] generalizing Clausius' inequality, the further generalization embodied in fluctuation theorems [4,5], and the extension of these relations to calculating potentials of mean force [6]. These advances have allowed researchers to measure equilibrium quantities, such as free energy changes, by observing how a system reacts to being driven out of equilibrium, e.g. [7,8].

This literature typically assumes that the experiment is known, i.e. that the exact time course of the driving signal is given. However, systems that are embedded in realistic environments, for example a biological macromolecule operating under natural conditions, are exposed to stochastic driving. Here, we therefore study driven systems for which the changes in the driving signal(s) are governed by some probability density $P_X$. This can be any stochastic process, and the results we derive require neither that $P_X$ has specific properties, nor that it is known by the system. We assume that there is no feedback from the system to the driving signal. The dissipation, averaged not only over the system's path through its state space, but also over driving protocols, then quantifies the system's energetic inefficiency.

The dynamics of the system perform a computation by changing the system's state, as a function of the driving signal. As a result, the new system state contains some memory about the driving signal. The system dynamics can be interpreted as computing a model: past environmental influences are mapped onto the current state of the system, which through its correlation with forthcoming environmental fluctuations implicitly contains a prediction of the future.

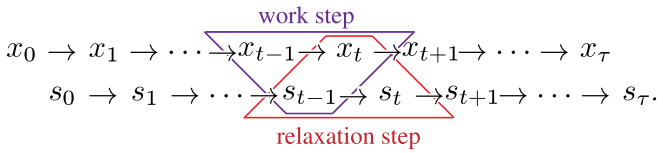
In this Letter, we ask how the quality of this (implicit) model is related to thermodynamic efficiency. But how do we measure the quality of a model? A useful model has to have predictive power (see e.g. [9–12] and references therein), meaning it must capture predictive information [13–16], while not being overly complicated. In other words, the model should contain as little dispensable nonpredictive information as possible.

Our central contribution is the demonstration of a fundamental equivalence between the instantaneous nonpredictive information carried by the system and the dissipation of energy.

*Problem setup.*—Let $s_t$ denote the state of the system at time $t$, while $x_t$ denotes the driving signal. The dynamics of the system are modeled by discrete time Markovian conditional state-to-state transition probabilities, $p(s_t|s_{t-1}, x_t)$. The external drive is governed by $P_X = p(x_0, \ldots, x_\tau)$. We assume that at time $t = 0$, the system is in thermodynamic equilibrium, in contact with a heat bath with inverse temperature $\beta := 1/k_{\mathrm{B}}T$. A change in the external driving signal $x_0 \to x_1$ forces the system out of equilibrium.

The system responds by changing its state $s_0 \to s_1$, according to the transition probability $p(s_1|s_0, x_1)$. The external signal subsequently changes again $x_1 \to x_2$, and the process is repeated until time $t = \tau$:




$$x_0 \to x_1 \to \cdots \to x_{t-1} \to x_t \to x_{t+1} \to \cdots \to x_\tau$$
$$s_0 \to s_1 \to \cdots \to s_{t-1} \to s_t \to s_{t+1} \to \cdots \to s_\tau.$$

The system remains in thermal contact with the heat bath during the entire protocol $x_0, \ldots, x_\tau$, as in [17]. Work is done during a work step, as the external signal changes from $x_{t-1}$ to $x_t$ [17,18],

$$W[s_{t-1}; x_{t-1} \to x_t] := E(s_{t-1}, x_t) - E(s_{t-1}, x_{t-1}). \quad (1)$$

In response to this change, the system relaxes from $s_{t-1}$ to $s_t$ in a relaxation step. The total work over the course of a driving protocol is $W = \sum_{t=1}^{\tau} W[s_{t-1}; x_{t-1} \to x_t]$. The total change in energy, $\Delta E := E(s_\tau, x_\tau) - E(s_0, x_0) = W + Q$, equals the total work plus the total heat, $Q = \sum_{t=1}^{\tau} [E(s_t, x_t) - E(s_{t-1}, x_t)]$, flowing into the system during the relaxation steps.

For now, we assume that the kernel which describes the dynamics, $p(s_t|s_{t-1}, x_t)$, is fixed. However, the conditional distribution over states after the work step but before the system relaxes, $p(s_{t-1}|x_t)$, changes as a function of time, as does the conditional distribution over states after the relaxation step, $p(s_t|x_t)$. In general, these distributions are not the same, and neither of them is an equilibrium distribution. Under Markovian system dynamics, the probability before a relaxation step simplifies to

$$p(s_{t-1}|x_t) = \langle\langle\langle \cdots \langle p(s_{t-1}|s_{t-2}, x_{t-1})\rangle_{p(s_{t-2}|s_{t-3}, x_{t-2})} \cdots \rangle_{p(s_1|s_0, x_0)}\rangle_{p_{\mathrm{eq}}(s_0|x_0)}\rangle_{p(x_0, \ldots, x_{t-1}|x_t)}, \quad (2)$$

and the distribution after a relaxation step is given by

$$p(s_t|x_t) = \langle p(s_t|s_{t-1}, x_t)\rangle_{p(s_{t-1}|x_t)}. \quad (3)$$

Angled brackets with a subscripted probability $p$ denote an average over $p$.

The equilibrium distribution is the same function, before and after relaxation, $p_{\mathrm{eq}}(s|x_t) := e^{-\beta(E(s,x_t)-F_t)}$, where $s$ refers to the state of the system with energy $E(s, x_t)$, and $F_t := F[x_t]$ denotes equilibrium free energy. The probability of a specific path through the system's state space, given the protocol, is

$$P_{S|X} = p_{\mathrm{eq}}(s_0|x_0) \prod_{t=1}^{\tau} p(s_t|s_{t-1}, x_t), \quad (4)$$

and the joint probability, $P_{S,X} := p(s_0, \ldots, s_\tau, x_0, \ldots, x_\tau)$, is

$$P_{S,X} = p(x_0) p_{\mathrm{eq}}(s_0|x_0) \prod_{t=1}^{\tau} p(x_t|x_0, \ldots, x_{t-1}) p(s_t|s_{t-1}, x_t). \quad (5)$$

In the following, unless otherwise clear from the context, angled brackets without a subscript denote an average over the distribution $P_{S,X}$.

*Dissipation out of equilibrium.*—After the conclusion of the protocol, the probability over system states is given by $p(s_\tau|x_\tau)$, in general not an equilibrium distribution. Then, in addition to the equilibrium free energy, $F_\tau$, there is another contribution to the free energy, because the system is not in thermodynamic equilibrium. This additional free energy would be dissipated as heat to the environment if the system were to relax to thermodynamic equilibrium. For Markovian system dynamics, the additional nonequilibrium contribution is proportional [13] to the relative entropy (Kullback-Leibler divergence) between the actual distribution $p(s_\tau|x_\tau)$ at the end of the protocol and the equilibrium distribution,

$$F_\tau^{\mathrm{add}}[p(s_\tau|x_\tau)] = k_B T D_{\mathrm{KL}}[p(s_\tau|x_\tau) \,\|\, p_{\mathrm{eq}}(s_\tau|x_\tau)]. \quad (6)$$

The non-negative Kullback-Leibler (KL) divergence [19] between distributions $p(x)$ and $q(x)$ is defined as

$$D_{\mathrm{KL}}[p(x) \,\|\, q(x)] := \left\langle \ln\left[\frac{p(x)}{q(x)}\right]\right\rangle_{p(x)} \geq 0. \quad (7)$$

The sum of both contributions to the free energy constitutes the overall nonequilibrium free energy, $F_{\mathrm{neq}}[p(s_\tau|x_\tau)] = F_\tau + F_\tau^{\mathrm{add}}[p(s_\tau|x_\tau)]$. Here, nonequilibrium free energy is defined in analogy to the standard equilibrium free energy as a functional of the probability distribution, but applied to any probability distribution [20–22], that is to any $p(s|x)$,

$$F_{\mathrm{neq}}[p(s|x)] := \langle E(s, x)\rangle_{p(s|x)} + k_B T \langle \ln[p(s|x)]\rangle_{p(s|x)}. \quad (8)$$

The average work irretrievably lost over the course of a driving protocol,

$$\langle W_{\mathrm{diss}}\rangle_{P_{S|X}} := \langle W\rangle_{P_{S|X}} - \Delta F_{\mathrm{neq}}, \quad (9)$$

equals the average work performed on the system minus the nonequilibrium free energy change $\Delta F_{\mathrm{neq}} := F_{\mathrm{neq}}[p(s_\tau|x_\tau)] - F_{\mathrm{neq}}[p(s_0|x_0)]$. We can compare this to the average *excess work* for a given protocol, $\langle W_{\mathrm{ex}}\rangle_{P_{S|X}} := \langle W\rangle_{P_{S|X}} - \Delta F$, the total work done on the system in excess of the equilibrium free energy change $\Delta F := F_\tau - F_0$ which would be the work done if the driving signal changed quasistatically (infinitely slowly), and hence the system remained in thermodynamic equilibrium throughout the protocol. This excess work equals the dissipated work only when the protocol includes a final equilibration of the system.

Since the system starts in equilibrium, the total change in nonequilibrium free energy is the equilibrium free energy change plus the above-mentioned additional contribution to the free energy, $\Delta F_{\text{neq}} = \Delta F + F_\tau^{\text{add}}[p(s_\tau|x_\tau)]$. The dissipation is then the excess work minus this additional nonequilibrium contribution,

$$\langle W_{\text{diss}}\rangle_{P_{S|X}} = \langle W_{\text{ex}}\rangle_{P_{S|X}} - F_\tau^{\text{add}}[p(s_\tau|x_\tau)] \leq \langle W_{\text{ex}}\rangle_{P_{S|X}}. \quad (10)$$

Later, we derive a lower bound on the dissipation and excess work averaged over all protocols, denoted by $\langle W_{\text{diss}}\rangle$ and $\langle W_{\text{ex}}\rangle$, respectively.

Each of the incremental work steps, $x_t \to x_{t+1}$, is accompanied by a nonequilibrium free energy change given by $\Delta F_{\text{neq}}[x_t \to x_{t+1}] := F_{\text{neq}}[p(s_t|x_{t+1})] - F_{\text{neq}}[p(s_t|x_t)]$, so that the average dissipation during each work step is

$$\langle W_{\text{diss}}[x_t \to x_{t+1}]\rangle := \langle W[s_t; x_t \to x_{t+1}]\rangle_{p(s_t,x_t,x_{t+1})} - \langle \Delta F_{\text{neq}}[x_t \to x_{t+1}]\rangle_{p(x_t,x_{t+1})}. \quad (11)$$

The nonequilibrium free energy change during each relaxation step is

$$\Delta F_{\text{neq}}[x_t; s_{t-1} \to s_t] = F_t^{\text{add}}[p(s_t|x_t)] - F_t^{\text{add}}[p(s_{t-1}|x_t)], \quad (12)$$

which equals the change in KL divergence from the equilibrium distribution.

*Predictive power, memory, and dissipation.*—The system state and the external signal are random variables that share information $I[s_t, x_t] := \langle \ln[p(s_t, x_t)/p(s_t)p(x_t)]\rangle_{p(s_t,x_t)}$, where $p(s_t) = \langle p(s_t|x_t)\rangle_{p(x_t)}$. Mutual information [23] measures the reduction in uncertainty about the outcome of a random variable upon learning the identity of another variable, and it is symmetric: $I[s_t, x_t] = H[s_t] - H[s_t|x_t] = H[x_t] - H[x_t|s_t]$. Uncertainty is quantified by the entropy, $H[s_t] := -\langle \ln p(s_t)\rangle_{p(s_t)}$, and the conditional entropy, $H[s_t|x_t] := -\langle \ln p(s_t|x_t)\rangle_{p(s_t,x_t)}$, respectively. The system transition probability, $p(s_t|s_{t-1}, x_t)$, is assumed to depend on the current signal value $x_t$ and system state $s_{t-1}$. These two dependencies are sufficient to induce correlations between the system's current state and previous signal values. The memory the system keeps about the external signal can then be quantified by the information that the system state $s_t$ retains about a trajectory $\{x_{t-\tau_m}, \ldots, x_t\}$. In general, there are temporal correlations in the input signal, and hence, there can be correlations between $s_t$ and future signal values. That is, some of the memory retained in the system's state is information about the future trajectory $\{x_{t+1}, \ldots, x_{t+\tau_f}\}$. Here we focus on the instantaneous memory, $I_{\text{mem}}(t) := I[s_t, x_t]$, and the instantaneous predictive power [11], $I_{\text{pred}}(t) := I[s_t, x_{t+1}] = H[x_{t+1}] - H[x_{t+1}|s_t]$.

The implicit model computed by the system's dynamics, $p(s_t|s_{t-1}, x_t)$, which map the signal $x_t$ onto state $s_t$, given the current state $s_{t-1}$, contains the probabilistic map $p(x_{t+1}|s_t)$, which represents the prediction of $x_{t+1}$, given the value of $s_t$.

The instantaneous nonpredictive information is defined as the difference between instantaneous memory and predictive power, $I_{\text{mem}}(t) - I_{\text{pred}}(t)$. It represents useless nostalgia and provides a measure for the ineffectiveness of the model.

Averaging the nonequilibrium free energy over protocols allows us to write

$$\beta\langle F_{\text{neq}}[p(s|x)]\rangle_{p(x)} = \beta\langle E(s, x)\rangle_{p(s,x)} - H[s|x]. \quad (13)$$

Combining this with Eqs. (1) and (11) [24], we arrive at our first result: the instantaneous nonpredictive information is proportional to the average work dissipated while the signal changes from $x_t$ to $x_{t+1}$,

$$\beta\langle W_{\text{diss}}[x_t \to x_{t+1}]\rangle = I_{\text{mem}}(t) - I_{\text{pred}}(t). \quad (14)$$

In summary, the unwarranted retention of past information is fundamentally equivalent to energetic inefficiency.

*Lower bound on total dissipation.*—We now relate the total average dissipated work during the entire protocol, averaged over all protocols, $\langle W_{\text{diss}}\rangle$, to the total nostalgia, $I_{\text{mem}} - I_{\text{pred}}$, given by the difference between the total instantaneous memory, $I_{\text{mem}} := \sum_{t=0}^{\tau-1} I_{\text{mem}}(t)$, and the total instantaneous predictive power, $I_{\text{pred}} := \sum_{t=0}^{\tau-1} I_{\text{pred}}(t)$. To that end, we need to sum Eq. (11) over all time steps. This sum includes an average over the sum of changes in nonequilibrium free energy, which can be expressed as

$$\left\langle \sum_{t=0}^{\tau-1} \Delta F_{\text{neq}}[x_t \to x_{t+1}] \right\rangle = \langle \Delta F_{\text{neq}} - \Delta F_{\text{neq}}^{\text{relax}}\rangle, \quad (15)$$

in terms of the average total nonequilibrium free energy change, $\langle \Delta F_{\text{neq}}\rangle$, and the average sum of nonequilibrium free energy changes during relaxation steps,

$$\langle \Delta F_{\text{neq}}^{\text{relax}}\rangle := \left\langle \sum_{t=0}^{\tau-1} \Delta F_{\text{neq}}[x_t; s_t \to s_{t+1}] \right\rangle \leq 0. \quad (16)$$

This quantity is nonpositive because, on average, during relaxation steps the system evolves toward equilibrium. The total dissipation then becomes, using Eq. (14),

$$\beta\langle W_{\text{diss}}\rangle = I_{\text{mem}} - I_{\text{pred}} - \beta\langle \Delta F_{\text{neq}}^{\text{relax}}\rangle. \quad (17)$$

The total nostalgia therefore provides a lower bound on the total average dissipation and also, due to Eq. (10), on the total average excess work,

$$I_{\text{mem}} - I_{\text{pred}} \leq \beta\langle W_{\text{diss}}\rangle \leq \beta\langle W_{\text{ex}}\rangle. \quad (18)$$

We can use this result to refine Landauer's principle [25], which states that any erasure of information must be balanced by an increase in entropy elsewhere. The information erased during a protocol, such as the reset protocol of Landauer [25], is the entropy change $\mathcal{I}_e := H[s_0|x_0] - H[s_\tau|x_\tau]$. Note that the information erased here is not

mutual information about the driving signal, but rather information that could have potentially been extracted from the system by some measurement process. Landauer pointed out that the erasure of information requires heat to flow out of the system, which obeys [using the first and second laws of thermodynamics, and Eqs. (8) and (9)]

$$-\beta\langle Q\rangle = \mathcal{I}_e + \beta\langle W_{\text{diss}}\rangle \geq \mathcal{I}_e. \tag{19}$$

Substituting our result from Eq. (17) into Eq. (19) yields the new relation

$$-\beta\langle Q\rangle = \mathcal{I}_e + I_{\text{mem}} - I_{\text{pred}} - \beta\langle \Delta F_{\text{neq}}^{\text{relax}}\rangle. \tag{20}$$

Thus [using Eq. (16)] we obtain a refinement of Landauer's principle,

$$-\beta\langle Q\rangle \geq \mathcal{I}_e + I_{\text{mem}} - I_{\text{pred}}, \tag{21}$$

where the bound is augmented by the total nostalgia. The system dynamics of a computing device that retains memory therefore must be maximally predictive to approach Landauer's limit.

*Discussion.*—The dynamics, $p(s_t|s_{t-1}, x_t)$, have been assumed fixed for any given system. However, biological systems are typically adapted to their environment. One can then ask if there is a simple principle underlying the process producing this adaptation. If such a principle exists, then it may reflect the importance of energetic efficiency, because of the resulting competitive advantage for reproducing organisms. While other criteria, such as robustness and sensitivity, play a role, energetic efficiency is of fundamental relevance. This is exemplified by biological molecules that harness environmental fluctuations to accomplish energetically-costly downstream tasks. The more efficiently such a molecule can operate, the more it can accomplish through coupling to endergonic chemical reactions or mechanical actions. For example, with more efficient coupling to the environment, the molecular motor kinesin can carry larger cargos. Likewise, with greater efficiency cytochrome $c$ oxidase complex, an enzyme that pumps protons across a membrane, can create stronger electrochemical gradients. Evidence for the importance of energetic efficiency is furthermore found in biomolecular machines that approach 100% efficiency when driven in a natural fashion: the stall torque for the $F_1$-ATPase [26] and the stall force for myosin V [27] are near the maximal values possible given the free energy liberated by ATP hydrolysis and the sizes of their respective rotations and steps.

These and many other biological functions require some correspondence between the environment and the systems that implement them. Therefore the memory of their instantiating systems must be nonzero. We have shown that any such system with nonzero memory must conduct predictive inference, at least implicitly, to approach maximal energetic efficiency.

A substantial amount of research has sought to relate emerging biological functions and behaviors to efficient energy usage. Examples range from animal behavior (e.g. [28]) to single neurons, where recently researchers have proposed that the minimization of energy expenditure, subject to constraints given by the desired function, may be "a unifying principle governing neuronal biophysics" [29]. On the other hand, there is much research on optimal information processing in neurons. For a recent review, see e.g. [30], which proposes that the extraction of predictive information in biological signal processing may constitute, or at least lead to, a general principle. By directly relating memory and predictive power to dissipation out of equilibrium, the results we have derived here indicate that these two important paradigms are deeply connected.

While it is perhaps intuitive that neurons and organisms should have to implement predictive inference to function well, our results have the striking implication that on all scales energetic efficiency calls for predictive inference. This includes the smallest biological units, such as molecular machines.

Our results also specify the required kind of predictive inference: maximization of predictive power at a desired level of system memory, as in [11]. This connects with work on optimal predictive inference algorithms discussed in [12,14,31], and references therein. We envision implementing these algorithms in fast and efficient hardware. The results we have derived here could then be used to choose the energetically most efficient implementation among the many possible choices.

*Conclusion.*—We argued that dissipation far from thermodynamic equilibrium is given by average work minus nonequilibrium free energy change. We also argued that the nonpredictive part of a system's memory provides a natural measure for the inefficiency of a system's implicit model of its environment.

We showed that instantaneous nonpredictive information is proportional to the energy dissipated when an external driving signal changes by an incremental amount, thereby doing work on the system. This result demonstrates the intimate connection between prediction and energetic efficiency. Summed over the entire protocol, the total nonpredictive information provides a lower bound on the total dissipation.

These results imply that any system which is built to have nonzero memory has to be predictive in order to allow for minimal possible dissipation, i.e. to operate at maximal energetic efficiency. Our results furthermore allow for a refinement of Landauer's principle, applied to systems away from thermodynamic equilibrium.

We have provided a connection between nonequilibrium thermodynamics and learning theory, by making precise how two important aspects of life are fundamentally related: making a predictive model of the environment and using available energy efficiently.

S. S. thanks William Bialek and Robert S. Shaw for many inspiring discussions, and Matteo Marsili for helpful technical comments. S. S. and A. J. B. are thankful for conversations with James P. Crutchfield and Michael R. DeWeese. D. A. S. and G. E. C. were funded by the Office of Basic Energy Sciences of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. A. J. B. was funded by NSF Grants No. SBE 0542013 and SMA 1041755 to the Temporal Dynamics of Learning Center, an NSF Science of Learning Center.

*sstill@hawaii.edu

†Present address: Center for Systems and Synthetic Biology, University of California, San Francisco, California 94158, USA.

[1] H. Lodish, A. Berk, S. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, *Molecular Cell Biology* (Freeman, San Francisco, 2000).

[2] C. Jarzynski, Annu. Rev. Condens. Matter Phys. **2**, 329 (2011).

[3] C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997).

[4] D. J. Evans, E. G. D. Cohen, and G. P. Morriss, Phys. Rev. Lett. **71**, 2401 (1993).

[5] G. E. Crooks, Phys. Rev. E **60**, 2721 (1999).

[6] G. Hummer and A. Szabo, Proc. Natl. Acad. Sci. U.S.A. **98**, 3658 (2001).

[7] J. T. Liphardt, S. Dumont, S. B. Smith, I. Tinoco, and C. Bustamante, Science **296**, 1832 (2002).

[8] D. Collin, F. Ritort, C. Jarzynski, S. B. Smith, I. Tinoco, and C. Bustamante, Nature (London) **437**, 231 (2005).

[9] H. Jeffreys, *Theory of Probability* (Oxford University, New York, 1998), 3rd ed.

[10] W. Bialek, in *Physics of Bio-Molecules and Cells*, Proceedings of the Les Houches Summer School, Session LXXV (Springer-Verlag, Berlin, 2001), p. 485–577.

[11] S. Still, Europhys. Lett. **85**, 28005 (2009).

[12] S. Still, J. P. Crutchfield, and C. Ellison, Chaos **20**, 037111 (2010).

[13] R. Shaw, *The Dripping Faucet as a Model Chaotic System* (Aerial, Santa Cruz, CA, 1984).

[14] J. P. Crutchfield and K. Young, Phys. Rev. Lett. **63**, 105 (1989).

[15] W. Bialek, I. Nemenman, and N. Tishby, Neural Comput. **13**, 2409 (2001).

[16] J. P. Crutchfield and D. P. Feldman, Chaos **13**, 25 (2003).

[17] G. E. Crooks, J. Stat. Phys. **90**, 1481 (1998).

[18] A. Imparato, L. Peliti, G. Pesce, G. Rusciano, and A. Sasso, Phys. Rev. E **76**, 050101 (2007).

[19] S. Kullback and R. A. Leibler, Ann. Math. Stat. **22**, 79 (1951).

[20] B. Gaveau and L. S. Schulman, Phys. Lett. A **229**, 347 (1997).

[21] G. E. Crooks, Phys. Rev. E **75**, 041119 (2007).

[22] D. A. Sivak and G. E. Crooks, Phys. Rev. Lett. **108**, 150601 (2012).

[23] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).

[24] $\beta\langle W_{\mathrm{diss}}[x_t \rightarrow x_{t+1}]\rangle = \beta(\langle E(s_t, x_{t+1})\rangle_{p(s_t,x_{t+1})} - \langle E(s_t, x_t)\rangle_{p(s_t,x_t)}) - \beta(\langle F_{\mathrm{neq}}[p(s_t|x_{t+1})]\rangle_{p(x_{t+1})} + \langle F_{\mathrm{neq}}[p(s_t|x_t)]\rangle_{p(x_t)}) = H[s_t|x_{t+1}] - H[s_t|x_t] = I_{\mathrm{mem}}(t) - I_{\mathrm{pred}}(t)$.

[25] R. Landauer, IBM J. Res. Dev. **5**, 183 (1961).

[26] K. Kinosita, R. Yasuda, H. Noji, and K. Adachi, Phil. Trans. R. Soc. B **355**, 473 (2000).

[27] G. Cappello, P. Pierobon, C. Symonds, L. Busoni, J. C. M. Gebhardt, M. Rief, and J. Prost, Proc. Natl. Acad. Sci. U.S.A. **104**, 15328 (2007).

[28] R. M. Alexander, *Optima for Animals* (Edward Arnold, London, 1981).

[29] A. Hasenstaub, S. Otte, E. Callaway, and T. J. Sejnowski, Proc. Natl. Acad. Sci. U.S.A. **107**, 12329 (2010).

[30] W. Bialek, *Biophysics: Searching for Principles* (Princeton University, Princeton, NJ, to be published).

[31] S. Still and D. Precup, Theory Biosci. **131**, 139 (2012).