Santa Fe Institute Working Paper 2017-07-020
arxiv.org:1707.03962 [physics.gen-ph]

Prediction and Dissipation in Nonequilibrium Molecular Sensors: Conditionally Markovian Channels Driven by Memoryful Environments

Sarah E. Marzen[1, *] and James P. Crutchfield[2, †]

[1] *Physics of Living Systems, Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139*
[2] *Complexity Sciences Center and Department of Physics, University of California at Davis, One Shields Avenue, Davis, CA 95616*

(Dated: July 29, 2018)

Biological sensors must often predict their input while operating under metabolic constraints. However, determining whether or not a particular sensor is evolved or designed to be accurate and efficient is challenging. This arises partly from the functional constraints being at cross purposes and partly since quantifying the prediction performance of even *in silico* sensors can require prohibitively long simulations, especially when highly complex environments drive sensors out of equilibrium. To circumvent these difficulties, we develop new expressions for the prediction accuracy and thermodynamic costs of the broad class of conditionally Markovian sensors subject to complex, correlated (unifilar hidden semi-Markov) environmental inputs in nonequilibrium steady state. Predictive metrics include the instantaneous memory and the total predictable information (the mutual information between present sensor state and input future), while dissipation metrics include power extracted from the environment and the nonpredictive information rate. Success in deriving these formulae relies on identifying the environment's causal states, the input's minimal sufficient statistics for prediction. Using these formulae, we study large random channels and the simplest nontrivial biological sensor model—that of a Hill molecule, characterized by the number of ligands that bind simultaneously, the sensor's cooperativity. We find that the seemingly impoverished Hill molecule can capture an order of magnitude more predictable information than large random channels.

PACS numbers: 02.50.-r 89.70.+c 05.45.Tp 02.50.Ey
Keywords: predictive information rate, information processing, nonequilibrium steady state, thermodynamics

I. INTRODUCTION

To perform functional tasks, synthetic nanoscale machines and their macromolecular cousins simultaneously manipulate energy, information, and matter. They are *information engines*—systems that operate by synergistically balancing the energetics of their physical substrate against required information generation, storage, loss, and transformation to support a given functionality. Classically, information engines were conceived as either potential computers [1]—that is, physical systems that can compute anything given the right program—or as Maxwellian-like demons that use information as a resource to convert disordered energy to useful work [2–11]. Recently, investigations into functional computation [12] embedded in physical systems led to studies of the thermodynamics of various kinds of information processing [13], including the thermodynamic costs of information creation [14], noise suppression [15], error correction and synchronization [6], prediction [16–18], homeostasis [7], learning [19], structure [20], and intelligent control [21]. Due to its broad importance to the survival of biological organisms, here we focus on a specific functional computation in information engines: how sensors predict their environment.

Evolved and designed sensory systems can be tasked with at least two, potentially competing, objectives: accurately predicting inputs [22, 23] and contending with metabolic constraints [24, 25] [26]. Accurate and energy-efficient predictive feature extraction can be used to reap increased rewards from the environment, no matter the particular "reward function" [27–30].

Optimizing such sensors requires a nontrivial matching of sensory statistics and sensor structure [7, 31]. Undaunted, much effort has been invested to find energetically efficient and maximally predictive sensors, often simplifying the challenges by ignoring action policies—how the sensed information is used. Some seek sensor models that maximize a combination of prediction power and (energetic) efficiency; e.g., as in Refs. [17, 31]. Others validate learning rules based on whether or not they maximize the aforementioned objective function [32, 33]. Finally, others compare real biological sensors to *in silico* null models; e.g., as in Ref. [23].

All these efforts require estimating prediction and dissipation metrics of given sensors. Doing so can be surpris-

* semarzen@mit.edu
† chaos@ucdavis.edu

ingly difficult. Consider the total predictable information captured by a sensor [23, 34], which we take to be the mutual information between the present sensor state and the future of the sensory input. There are uncountably infinite possible futures. How, then, are we to estimate the total predictable information from simulations given that we are *always* in the undersampled limit?

One approach is to simulate and employ sophisticated techniques for estimating entropy in the undersampled limit; e.g., as in Ref. [35]. However, even estimating the joint probability distribution of environment and sensor states from simulations can require long simulations, as we find in one example below. Alternatively, one can rewrite prediction and dissipation metrics in terms of generators of the environment and sensor, similar to how the predictive information bottleneck can be recast in terms of forward- and reverse-time causal states [36]. The following pursues this second line of inquiry by solving a Chapman-Kolmogorov equation that yields a partial differential equation for the joint probability distribution of forward-time causal states and channel state. We find new closed-form expressions for total predictable information, power consumption, instantaneous memory, and nonpredictive information rate in nonequilibrium steady-state sensors. The challenge to this analytical approach, of course, is that it requires accurate models of the environment and sensor dynamics.

We address a very general class of environments and sensors: conditionally Markovian channels subject to a realization of a stationary unifilar hidden semi-Markov process [37] with an energy function associated to each environment and channel state combination. This setting is sufficiently general that it allows nonequilibrium sensing. The latter is crucial for accurate sensing [38–42], in that the Boltzmann distribution need not be equivalent to the steady-state distribution over sensor states for a frozen environment. Unlike previous treatments, we do not explicitly demarcate a separate memory or energy source accessed by the sensor [41, 43, 44], but include components as part of the sensor. Our framework allows for nondetailed balance dynamics. A lack of detailed balance in sensor dynamics is a reminder of the presence of such a reservoir. Finally, the environment is assumed to be unaffected by the sensor. That is, either there is no feedback or the environment is infinitely large. After all of these assumptions, we calculate the power consumed from the environment by the sensor.

For illustration, we study both large random channels (randomly-wired channels with a large number of states) and an "optimal" Hill molecule, which is a simple model of a ligand-gated channel; see Fig. 1. For these examples, we assume that the ligand concentration is a realization of a semi-Markov process—a generalization over previous efforts that assumed Markovian [17, 45] or Gaussian [31] processes. This generalization is necessary when, for instance, the Hill molecule represents a nicotinic acetylcholine receptor on a synapse and ligands are acetylcholine molecules since, as a practical matter, neuronal dynamics are often non-Markovian and non-Gaussian [46].

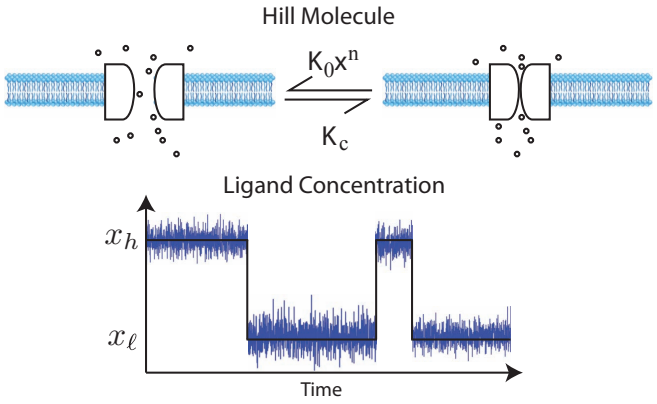

FIG. 1. Hill molecule: (Top) Ion channel in a membrane in the open (left) and closed (right) states in which ions can and cannot travel through. (Below) Ligand concentrations during the open-closed cycle of the molecule.

We find that (i) increases in cooperativity (sharpness of response) of the Hill molecule lead to increases in both prediction power and heat dissipation rate, (ii) a large fraction of the heat dissipation rate comes from inefficient prediction, and (iii) simple gradient-based adaptation rules lead to hysteresis. Furthermore, we find that the total predictable information captured by this seemingly impoverished Hill molecule exceeds the total predictable information captured by large random channels by an order of magnitude, despite the latter's ability to capture potentially useful information about the past of the stimulus. This latter result would be impossible to obtain via interactions with a continuous-time Markovian environment.

II. BACKGROUND

Central to our analysis is an appreciation of causal states (minimal sufficient statistics of prediction or retrodiction), unifilar hidden semi-Markov processes, and conditionally Markovian channels. We review these concepts here, simultaneously introducing relevant notation.

A. Environment

Input symbols x take on any value in the observation alphabet $\mathcal{A}$. We code the *past* of the input time series as $\overleftarrow{x} = \ldots (x_{-2}, \tau_{-2}), (x_{-1}, \tau_{-1}), (x_0, \tau_+)$ and the input's *future* as $\overrightarrow{x} = (x_0, \tau_-), (x_1, \tau_1), (x_2, \tau_2), \ldots$, where τ_i is the total dwell time for symbol x_i. To ensure a unique coding, we stipulate that $x_i \neq x_{i+1}$. Note that symbol x_0 is seen for a total dwell time of $\tau_+ + \tau_- = \tau_0$; that is, the present splits the dwell time τ_0 in two.

As is typical, $\overleftarrow{X}$ is the random variable corresponding to semi-infinite input pasts and $\overrightarrow{X}$ the random variable corresponding to semi-infinite input futures. We now briefly review the definition of causal states, as described in Ref. [47]. Forward-time causal states $\mathcal{S}^+$, the minimal sufficient statistics for prediction, are defined via the following equivalence relation: two semi-infinite pasts, $\overleftarrow{x}$ and $\overleftarrow{x}'$, are considered "predictively" equivalent if:

$$\overleftarrow{x} \sim_{\epsilon^+} \overleftarrow{x}' \Leftrightarrow \Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}) = \Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}') \ .$$

The relation partitions the set of semi-infinite pasts into clusters of pasts. Each cluster is a *forward-time causal state* σ^+, a realization of the random variable $\mathcal{S}^+$. *Reverse-time causal states* $\mathcal{S}^-$, the minimal sufficient statistics for retrodiction, are defined similarly. Two semi-infinite futures, $\overrightarrow{x}$ and $\overrightarrow{x}'$, are considered "retrodictively" equivalent if:

$$\overrightarrow{x} \sim_{\epsilon^-} \overrightarrow{x}' \Leftrightarrow \Pr(\overleftarrow{X}|\overrightarrow{X} = \overrightarrow{x}) = \Pr(\overleftarrow{X}|\overrightarrow{X} = \overrightarrow{x}') \ .$$

This equivalence relation partitions the set of semi-infinite futures into clusters, each cluster being a *reverse-time causal state* σ^-, a realization of the random variable $\mathcal{S}^-$.

Forward- and reverse-time causal states are useful in the ensuing calculations due to the following Markov chains. First, forward-time causal states are a deterministic function of the input past ($\sigma^+ = \epsilon^+(\overleftarrow{x})$), and reverse-time causal states are a deterministic function of the input future ($\sigma^- = \epsilon^-(\overrightarrow{x})$). Hence, we have the Markov chains $\mathcal{S}^+ \to \overleftarrow{X} \to \overrightarrow{X}$ and $\mathcal{S}^- \to \overrightarrow{X} \to \overleftarrow{X}$; so that, for instance:

$$p(\overrightarrow{x}, \sigma^+|\overleftarrow{x}) = p(\overrightarrow{x}|\overleftarrow{x})p(\sigma^+|\overleftarrow{x}),$$

where we introduce a simplified notation for the probability distributions; e.g., $p(\sigma^+|\overleftarrow{x}) = \Pr(\mathcal{S}^+ = \sigma^+|\overleftarrow{X} = \overleftarrow{x})$. Note that $p(\sigma^+|\overleftarrow{x})$ is singly supported– that is, a Kronecker delta, $\delta_{\sigma^+, \epsilon^+(\overleftarrow{x})}$. However, causal states are minimal sufficient statistics of the past relative to the future and vice versa. And so, $\overleftarrow{X} \to \mathcal{S}^+ \to \overrightarrow{X}$ and $\overrightarrow{X} \to \mathcal{S}^- \to \overleftarrow{X}$ are also valid Markov chains; so that, for instance:

$$p(\overrightarrow{x}, \overleftarrow{x}|\sigma^+) = p(\overrightarrow{x}|\sigma^+)p(\overleftarrow{x}|\sigma^+).$$

Invoking these Markov chains is called *causal shielding*.

For the discussion that follows, we must define hidden Markov models and unifilarity. We first do so in the case of discrete-time, discrete-event processes. A hidden Markov model is equipped with hidden states and labeled transition probabilities that give the probability for emitting a particular symbol and transitioning to a new particular hidden state. (These are edge-emitting hidden Markov models, which are completely general and equivalent to the likely more familiar state-emitting hidden Markov models.) In a unifilar hidden Markov model, the state to which you transition is determined by the state that you're in and the symbol that you emit. These are less general instantiations than nonunifilar hidden Markov models in the following sense: the unifilar hidden Markov model corresponding to most nonunifilar hidden Markov models has uncountably infinite states. However, the process generated by a finite unifilar hidden Markov model has a finite number of causal states.

Now we turn our attention to continuous-time, discrete-event processes. Let's first address the more general case of unifilar hidden semi-Markov input, as in Ref. [37], which explains that the causal state information has three distinct components. Unifilar hidden semi-Markov processes have a causal structure that looks a little like state- and symbol-dependent conveyor belts which transition into one another. Each conveyor belt is labeled by a forward-time hidden state g_+. The forward-time hidden state is analogous to hidden states in unifilar hidden Markov models, except that forward-time hidden states now emit both symbols and dwell times. Dwell times are drawn from $\phi_g(\tau)$; emitted symbols are chosen with probability $p(x|g)$; and $g = \epsilon^+(g', x')$ is the next hidden state given that the current hidden state is g' and the current emitted symbol is x'. In other words, x' is emitted upon transition from g' to g. The last of these properties is called *unifilarity*; here, both the entire model is unifilar and the underlying hidden states exhibit the unifilarity property as well.

Having set up this machinery, forward-time causal states are labeled by $\sigma^+ = (g_+, x_+, \tau_+)$. That is, the forward-time hidden state g_+, current emitted symbol x_+, and time since last symbol τ_+ together comprise the forward-time causal states for unifilar hidden semi-Markov processes. Similarly, the reverse-time causal state $\sigma^- = (g_-, x_-, \tau_-)$ for a unifilar hidden semi-Markov process is a combination of reverse-time hidden state g_-, current emitted symbol x_-, and time to next symbol τ_-. The joint probability distribution $p(\sigma^+, \sigma^-)$ of forward-

time and reverse-time causal states for unifilar hidden semi-Markov processes can be obtained from formulae in Ref. [37].

While Ref. [37]'s hidden semi-Markov processes are completely general, they are restricted in that the unifilar hidden semi-Markov process corresponding to a particular *nonunifilar* hidden semi-Markov process might have unordered uncountably infinite states. That said, we believe that any continuous-time, discrete-event process can be approximated arbitrarily well by a countable unifilar hidden semi-Markov process. (Alas, a proof of this is still lacking.) As such, we do not believe that restricting ourselves to finite unifilar hidden semi-Markov processes is limiting.

In examples, we focus on semi-Markov input. This greatly constrains the forward- and reverse-time causal states. The forward-time causal states are now described by the pair (x_+, τ_+), where x_+ is the input symbol infinitesimally prior to the present and τ_+ is the time since last symbol—i.e., since x_{-1}. The reverse-time causal states are similarly described by the pair (x_-, τ_-), where x_- is the input symbol infinitesimally after the present and τ_- is the time to next symbol—i.e., until x_1. Let $\mathcal{T}_\pm$ be the random variable describing time since (to) last (next) symbol. The dwell time of symbol x has probability density function $\phi_x(\tau)$, and the probability of observing symbol x after x' is $q(x|x')$. By virtue of how we have chosen to encode our input: $q(x|x) = 0$.

Finally, the development to come requires finding the joint density $\rho(\sigma^+, \sigma^-)$ of forward- and reverse-time causal states. Let:

$$\Phi_{x_+}(\tau_+) = \int_{\tau_+}^{\infty} \phi_{x_+}(t)dt, \ \mu_{x_+} = 1/\int_0^{\infty} t\phi_{x_+}(t)dt \ ,$$

where $p(x_+)$ is the probability of observing symbol x_+. As detailed in Ref. [37], for the conditional density we can say:

$$\begin{aligned}\rho(\sigma^-|\sigma^+) &= \rho(g_-, x_-, \tau_-|g_+, x_+, \tau_+) \\ &= \delta_{x_+,x_-}\frac{\phi_{g_+}(\tau_+ + \tau_-)}{\Phi_{g_+}(\tau_+)}p(g_+|g_-, x_-) \ .\end{aligned}$$

For semi-Markov input, this simplifies the density: $\rho(\sigma^+, \sigma^-) = \rho((x_+, \tau_+), (x_-, \tau_-))$. As described in Ref. [37], we have:

$$\rho(x_+, \tau_+) = \mu_{x_+}\Phi_{x_+}(\tau_+)p(x_+) \ . \tag{1}$$

The probability $p(x_+)$ is given by:

$$p(x_+) = \left(\mathrm{diag}(1/\mu_x)\ \mathrm{eig}_1(q)\right)_{x_+} \ ,$$

where $\mathrm{diag}(1/\mu_x)$ is a diagonal matrix and $\mathrm{eig}_1(q)$ is the eigenvector of $q(\cdot|\cdot)$ associated with eigenvalue 1. The conditional density of reverse-time causal states given forward-time causal states is then:

$$\begin{aligned}\rho(\sigma^-|\sigma^+) &= \rho((x_-, \tau_-)|(x_+, \tau_+)) \\ &= \frac{\phi_{x_+}(\tau_+ + \tau_-)}{\Phi_{x_+}(\tau_+)}\delta_{x_+,x_-} \ .\end{aligned} \tag{2}$$

Together, Eqs. (1) and (2) give the joint density $\rho(\sigma^+, \sigma^-) = \rho(\sigma^+)\rho(\sigma^-|\sigma^+)$.

B. Sensory channel

We assume the channel is conditionally Markovian. As such, its dynamics are fully specified by input state-dependent kinetic rates. More precisely, the channel state y, with corresponding random variable Y, can take on any value in $\mathcal{Y}$, and the rate at which channel state y transfers to channel state y' when the input has value x is given by $k_{y\to y'}(x)$. Probability conservation dictates that:

$$k_{y\to y}(x) := -\sum_{y'} k_{y\to y'}(x) \ .$$

Then, the probability $p(y, t)$ of being in channel state y at time t evolves as:

$$\dot{p}(y, t) = \sum_{y'} k_{y'\to y}(x(t))\ p(y', t) \ ,$$

where $x(t)$ is the input symbol at time t.

To simplify notation and ease computation, we write dynamical evolution rules in matrix-vector form. Let $\vec{p}(y, t)$ be the column vector of probabilities that the channel is in a particular state y at time t, and let $M(x)$ be a matrix of rates: $M_{y',y}(x) = k_{y\to y'}(x)$. Then, we have:

$$\dot{\vec{p}}(y, t) = M(x(t))\ \vec{p}(y, t) \ . \tag{3}$$

With this, it is clear that $\vec{p}(y, t)$ can oscillate or decay to a steady state. The Perron-Frobenius theorem guarantees that:

$$p_{eq}(x) := \mathrm{eig}_0(M(x)) \ ,$$

the probability distribution over channel states when ligand concentration is set to x, is unique. This need not be the Boltzmann distribution.

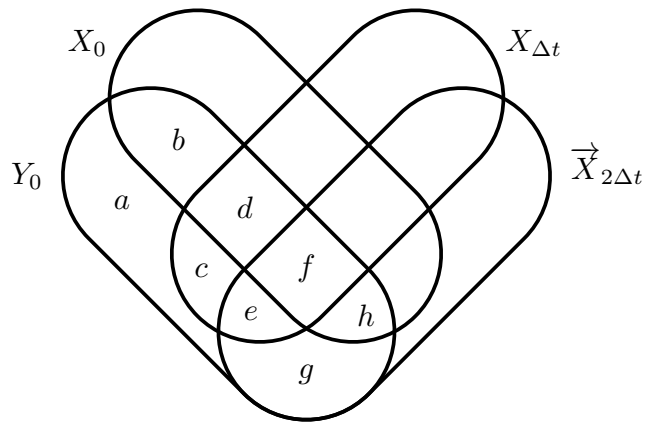

FIG. 2. Sensor information diagram [48] giving the relationship between prediction metrics and nonpredictive information rate. Instantaneous memory is $\mathrm{I}_{\mathrm{mem}} = \mathrm{I}[X_0; Y_0] = b + d + f + h$ and total predictable information $\mathrm{I}_{\mathrm{fut}} = \lim_{\Delta t \to 0} \mathrm{I}[Y_0; X_{0:}] = \lim_{\Delta t \to 0}(b + c + d + e + f + g + h)$, while the lower bound on power consumption, the nonpredictive information rate is $\dot{\mathrm{I}}_{\mathrm{np}} = \lim_{\Delta t \to 0}(b + h - c - e)/\Delta t$. Recall $a = \mathrm{H}[Y_0|X_0, X_{\Delta t}, \overrightarrow{X}_{2\Delta t}]$, $b = \mathrm{I}[Y_0; X_0|X_{\Delta t}, \overrightarrow{X}_{2\Delta t}]$, $f = \mathrm{I}[Y_0; X_0; X_{\Delta t}; \overrightarrow{X}_{2\Delta t}]$, and so on.

III. PREDICTION AND DISSIPATION METRICS

We employ several metrics to quantify the sensor's prediction performance and its energetic efficiency. *Instantaneous memory* $\mathrm{I}_{\mathrm{mem}}$ [16] and *total predictable information* $\mathrm{I}_{\mathrm{fut}}$ [33, 49] characterize prediction power, while the *nonpredictive information rate* $\dot{\mathrm{I}}_{\mathrm{np}}$ [16–18] and temperature-normalized *power consumption* βP monitor dissipation. In nonequilibrium steady state, when detailed balance holds, heat dissipation rate and power consumption are equivalent. Figure 2 uses an information diagram [48, 50] to illustrate the relationships between the various information-theoretic quantities in terms of the elementary information atoms—entropies, conditional entropies, and mutual informations—out of which they are constructed.

In defining the sensor, we grouped into a single sensor what might separately be called sensor, memory, and internal energy source(s). This allows us to consider active sensors with memory. However, reading the sensor by other downstream parts of an organism might entail a different grouping. In any case, since we use information-theoretic quantities to define prediction metrics, the Data Processing Inequality employed below guarantees that we establish a (not necessarily tight) upper bound on the instantaneous memory and total predictable information when reading a sensor.

Our selection of metrics is inspired by previous efforts to characterize prediction and dissipation in sensors. Instantaneous memory and nonpredictive information were first defined in Ref. [16]. Reference [18] also focused on nonpredictive information rate and power consumption, but did not calculate total predictable information or a more standard prediction-related metric. Reference [17] used the ratio of nonpredictive information rate to entropy production to characterize learning, though nonpredictive information rate is not necessarily an intuitive metric for learning. Moreover, this ratio can be greater than unity when nonpredictive information rate is negative. Finally, Ref. [31] focused on metrics for prediction, including a natural continuous-time extension of instantaneous predictive information, but not on metrics for dissipation.

A. Prediction metrics

Instantaneous memory and total predictable information are far from being the only metrics useful for characterizing a sensor's prediction capability. One could instead study instantaneous predictable information, $\mathrm{I}[Y_t; X_{t+\Delta t}]$ [16, 23], which requires choosing a relevant Δt. Or, following the echo-state network literature, one might instead calculate "memory capacity" [51] or "prediction capacity" [52]. Or, if a reward function is known [27], then the relevant metric would depend on the reward function and the organism's action policy.

Interestingly, information-theoretic prediction metrics might have more relevance to biology than, say, the so-called prediction capacity. In a discrete-time setting, extending Kelly's classic bet-hedging analysis shows that increases in expected log-growth rate via increases in sensory information is equal to the instantaneous predictable information $\mathrm{I}[Y_0; X_{\Delta t}]$ [53].

The *total predictable information* $\mathrm{I}_{\mathrm{fut}}$, defined as:

$$\mathrm{I}_{\mathrm{fut}} := \mathrm{I}[Y; \overrightarrow{X}] \ ,$$

is an upper bound on this increase in expected log-growth rate. (Note that we dropped time subscripts t due to stationarity.) This is the mutual information between present channel state and the input's future. It is the amount of information that is predictable about the input future from the present channel state. We merely assert that on ontogenetic timescales total predictable information constitutes a reasonable reward function [54, 55].

Even in a discrete-time setting, calculating the total predictable information appears intractable, as there is an uncountable infinity of possible input futures. To cope, we employ the causal shielding relations specified in Sec. II. Due to the feedforward nature of the channel-

input setup—that is, the channel's state does not affect the input—and the Markov chains given earlier—e.g., see Ref. [56]—we have the Markov chain $Y \to \mathcal{S}^+ \to \mathcal{S}^- \to \overleftarrow{X}$ and the Markov chain $Y \to \mathcal{S}^+ \to \overleftarrow{X} \to \mathcal{S}^-$. Two applications of the Data Processing Inequality yield:

$$\mathrm{I}_{\mathrm{fut}} = \mathrm{I}[Y; \mathcal{S}^-] = \mathrm{I}[Y; X_-, \mathcal{T}_-, \mathcal{G}_-] \ ,$$

which decomposes into:

$$\mathrm{I}_{\mathrm{fut}} = \mathrm{I}[Y; X_-] + \mathrm{I}[Y; \mathcal{T}_-, \mathcal{G}_- | X_-] \ .$$

The term $\mathrm{I}[Y; X_-]$ is called the *instantaneous memory* $\mathrm{I}_{\mathrm{mem}}$ [16], since it is the amount of information available from the channel state about the just-seen input symbol. Rewriting we have:

$$\mathrm{I}_{\mathrm{fut}} = \mathrm{I}_{\mathrm{mem}} + \mathrm{I}[Y; \mathcal{T}_-, \mathcal{G}_- | X] \ .$$

Thus, the total predictable information is the sum of instantaneous memory and information that is truly about the future, which here is the combined time-to-next-symbol and reverse-time hidden state. For the special case of semi-Markov input, the reverse-time hidden state is equivalent to the present observed symbol and so:

$$\mathrm{I}_{\mathrm{fut}} = \mathrm{I}_{\mathrm{mem}} + \mathrm{I}[Y; \mathcal{T}_- | X] \ .$$

The difference between the total predictable information and instantaneous memory for semi-Markov input is the information that the present sensor state captures about the time to next observed symbol.

B. Dissipation metrics

Next, we quantify the power extracted from the environment by the sensor system. Note that this does not include the power required to maintain the sensor in nonequilibrium steady state, even at fixed environment, due to the violation of detailed balance. Assuming access to a temperature-normalized "energy function" $\beta E(x, y)$, the temperature-normalized power βP is given by:

$$\beta P = \lim_{\Delta t \to 0} \frac{\langle \beta E(x_{t+\Delta t}, y_t) \rangle - \langle \beta E(x_t, y_t) \rangle}{\Delta t} \ . \qquad (4)$$

In effect, we group any internal energy sources into the sensor state, so that we calculate the work done by the environment on the entire sensor and its internal energy sources. Given an energy function, we do not assume that the distribution over sensor states reached, if the environment is fixed, is identical to the Boltzmann distribution.

If determining an energy function is not possible, we can calculate a lower bound using a continuous-time adaptation of the inequality in Ref. [16]:

$$\dot{\mathrm{I}}_{\mathrm{np}} := \lim_{\Delta t \to 0} \frac{\mathrm{I}[Y_t; X_t] - \mathrm{I}[Y_t; X_{t+\Delta t}]}{\Delta t} \leq \beta P \ , \qquad (5)$$

with an alternate equivalent definition in Ref. [18]. Note that Eq. (5) only holds in nonequilibrium steady state. See App. A.

$\dot{\mathrm{I}}_{\mathrm{np}}$ is called the *nonpredictive information rate* since it loosely corresponds to how much of the instantaneous memory is useless for predicting the next input. Reference [17] viewed $\dot{\mathrm{I}}_{\mathrm{np}}/\beta P$ as a learning efficiency, though see Ref. [45]. We take the view that $\dot{\mathrm{I}}_{\mathrm{np}}$ is a potentially useful lower bound on temperature-normalized power consumption and use $\mathrm{I}_{\mathrm{mem}}$ and $\mathrm{I}_{\mathrm{fut}}$ instead to characterize learning. Any differences between the formulae shown here and in Ref. [16, Eq. (2)] are superficial; we merely adapted the derivation for continuous-time processes. Unfortunately, the nonpredictive information rate is not necessarily a tight lower bound on temperature-normalized power. When the environment is Markovian, there is another lower bound on the heat dissipation rate that is proportional to the instantaneous memory [57], but we focus mainly on non-Markovian environments here.

IV. RESULTS

Given a known environment and sensor, two approaches to evaluate the aforementioned prediction and dissipation metrics present themselves:

- First, simulate the environment and sensor and approximate prediction and dissipation metrics based on observed frequencies;
- Or, second, find (new) closed-form expressions for prediction and dissipation metrics in terms of the environment generators ($\phi_x(\tau)$, $\epsilon^+(g, x)$, $p(x|g)$) and sensor $M(x)$.

The first can lead to prohibitively long simulations for accurate estimates. The following pursues the second. This requires solving a Chapman-Kolmogorov equation that turns into partial differential equations with coupled boundary conditions. The explicit derivations are lengthy and therefore are relegated to the appendices. The appropriate equations there are referenced here, where relevant.

That said, let's briefly expand upon the first approach to highlight its several potential difficulties. To calculate prediction and dissipation metrics other than the total predictable information $\mathrm{I}_{\mathrm{fut}}$, one can estimate $p(x_t, y_t)$ and $p(x_{t+\Delta t}, y_t)$ for some small Δt from

simulation and then use standard formulae ($\mathrm{I}[U;V] = \sum_{u,v} p(u,v)\log\left(p(u,v)/p(u)p(v)\right)$) and Eqs. (6) and (9) to calculate $\mathrm{I}_{\mathrm{mem}}$, $\dot{\mathrm{I}}_{\mathrm{np}}$, and βP. More sophisticated algorithms for calculating prediction metrics would directly approximate $\mathrm{I}[X_t, Y_t]$ [35] instead of "plugging-in" estimates of the corresponding probability distribution. As total simulation time T increases, estimates of the various probability distributions from empirical frequencies become more and more accurate and our corresponding estimates of prediction and dissipation metrics also become more accurate. As discussed in Sec. V, this might require surprisingly long times T to obtain accurate estimates.

Using the first approach to calculate the total predictable information $\mathrm{I}_{\mathrm{fut}}$ from simulations is difficult for two reasons. Both imply that we are always in the undersampled limit. First, we are working with a continuous-time system. As such, input futures $\overrightarrow{x}$ are described by lists of dwell times, and so to calculate $\mathrm{I}_{\mathrm{fut}}$ we are implicitly estimating probability density functions $\rho(\cdot)$ rather than probability distributions $p(\cdot)$. Second, we decided to consider semi-infinite input futures, rather than input futures of some finite time. Even for discrete-time systems, there are an uncountable infinity of possible semi-infinite input futures. The combination of these two challenges is so daunting that we will not calculate $\mathrm{I}_{\mathrm{fut}}$ from simulations.

Notably, from the closed-form expressions of Sec. IV later on we find that $\mathrm{I}_{\mathrm{fut}}$ is close to $\mathrm{I}_{\mathrm{mem}}$ for random channels and the simple Hill molecule. This result, though, will likely not hold for the interesting case of highly predictive sensors in complex environments [34].

A. Calculating prediction metrics

$\mathrm{I}_{\mathrm{fut}}$ and $\mathrm{I}_{\mathrm{mem}}$ can be analytically calculated, though, once $\rho(\sigma^+, y) = \Pr(\mathcal{S}^+ = \sigma^+, Y = y)$ is in hand. This follows, in turn, by manipulating a Chapman-Kolmogorov equation, shown in App. B. Set any ordering on the pairs (g, x); e.g., the ordering $(g_1, x_1), (g_1, x_2), \ldots, (g_{|\mathcal{G}|}, x_{|\mathcal{A}|})$. An expression for $p(y|\sigma^+) = p(y|g, x, \tau)$ is given by a combination of Eqs. (B5) and (B9):

$$p(y|g,x,\tau) = \left(e^{M(x)\tau}\mathrm{eig}_1(\mathbf{C})_{(g,x)}/\mu_g p(g)\right)_y \ ,$$

where $\mathbf{C}$ is a block matrix with entries:

$$\mathbf{C}_{(g,x),(g',x')} = \delta_{g,\epsilon^+(g',x')} p(x'|g') \int_0^\infty \phi_{g'}(t) e^{M(x')t} dt \ .$$

And so, $\mathrm{eig}_1(\mathbf{C})$ is a vector. Normalization forces $\vec{1}^\top \mathrm{eig}_1(\mathbf{C})_{(g,x)} = \mu_g p(g)$.

We then find the joint probability density as $\rho(y, \sigma^+) = p(y|\sigma^+)\rho(\sigma^+)$, which enables computation of all prediction metrics. Instantaneous memory is given by $\mathrm{I}_{\mathrm{mem}} = \mathrm{I}[X;Y]$, whereas $\mathrm{I}_{\mathrm{fut}} = \mathrm{I}[Y;\mathcal{S}^-]$. All the relevant probabilities—namely, $p(x,y)$ and $\rho(\sigma^-, y)$—are obtained from the previously derived density $\rho(\sigma^+, y)$. For instance, to calculate $\rho(\sigma^-, y)$, we employ the Markov chain $Y \to \mathcal{S}^+ \to \mathcal{S}^-$ to find:

$$\rho(\sigma^-, y) = \sum_{\sigma^+} \rho(\sigma^-|\sigma^+) p(y|\sigma^+) \rho(\sigma^+) \ .$$

And, to calculate $p(x,y)$, we recall that $\sigma^- = (g_-, x, \tau_-)$, so we only need marginalize the joint density $\rho((g_-, x, \tau_-), y)$.

B. Calculating dissipation metrics

Additionally, dissipation metrics can be calculated once:

$$\frac{\delta p(x,y)}{\delta t} = \lim_{\Delta t \to 0} \Big(\Pr(X_{t+\Delta t} = x, Y_t = y) - \Pr(X_t = x, Y_t = y) \Big) / \Delta t$$

is obtained. (Note that this is not the total time derivative dp/dt.) An expression for $\delta p/\delta t$ in terms of input and channel properties is given in Eq. (B11):

$$\begin{aligned}\frac{\delta p}{\delta t} = &\sum_{g',x' \neq x} \int d\tau' \ p(x|\epsilon^+(g',x')) p(x'|g') \phi_{g'}(\tau') \\ &\times \left(e^{M(x')\tau'} \mathrm{eig}_1(\mathbf{C})_{(g',x')}\right)_y \\ &- \sum_{g'} \int d\tau' \ p(x|g') \phi_{g'}(\tau') \left(e^{M(x)\tau'} \mathrm{eig}_1(\mathbf{C})_{(g',x)}\right)_y ,\end{aligned}$$

where normalization again requires $\vec{1}^\top \mathrm{eig}_1(\mathbf{C})_{(g,x)} = \mu_x p(g)$. Then, from earlier, we find that:

$$\begin{aligned}\dot{\mathrm{I}}_{\mathrm{np}} &= \lim_{\Delta t \to 0} \frac{\mathrm{H}[Y_t, X_{t+\Delta t}] - \mathrm{H}[Y_t, X_t]}{\Delta t} \\ &= \lim_{\Delta t \to 0} \Big(\sum_{x,y} \left(p(x,y) + \frac{\delta p}{\delta t} \Delta t \right) \log \frac{1}{p(x,y) + \frac{\delta p}{\delta t}\Delta t} \\ &\qquad - \sum_{x,y} p(x,y) \log \frac{1}{p(x,y)} \Big) / \Delta t \\ &= -\sum_{x,y} \frac{\delta p(x,y)}{\delta t} \log p(x,y) \ . \end{aligned} \tag{6}$$

If there is an energy function $E(x, y)$, then it is straightforward to show that power can be obtained from:

$$\beta P = \sum_{x,y} \frac{\delta p(x,y)}{\delta t} E(x,y). \tag{7}$$

For passive sensors, despite lacking direct access to an energy function, we find βP by calculating the steady-state distribution over channel states with fixed input:

$$\begin{aligned} \vec{p}_{eq}(y|x) &= \mathrm{eig}_0(M(x)) \\ &= \frac{e^{-\beta E(x,y)}}{Z_\beta(x)} \ , \end{aligned}$$

where the partition function is $Z_\beta(x) := \sum_y e^{-\beta E(x,y)}$. Hence:

$$\beta E(x,y) = \log \frac{1}{p_{eq}(y|x)} - \log Z_\beta(x) \ . \tag{8}$$

Recalling Eq. (4) and invoking stationarity—that $\Pr(X_t = x) = \Pr(X_{t+\Delta t} = x)$—yields:

$$\begin{aligned} \beta P &= \lim_{\Delta t \to 0} \Big(\left\langle \log \frac{1}{p_{eq}(x,y)} \right\rangle_{\Pr(X_{t+\Delta t}=x, Y_t=y)} \\ &\qquad - \left\langle \log \frac{1}{p_{eq}(x,y)} \right\rangle_{\Pr(X_t=x, Y_t=y)} \Big) / \Delta t \\ &= \sum_{x,y} \frac{\delta p(x,y)}{\delta t} \log \frac{1}{p_{eq}(y|x)} \ . \end{aligned} \tag{9}$$

The distributions $\Pr(X_{t+\Delta t} = x, Y_t = y)$ and $\Pr(X_t = x, Y_t = y)$ can be obtained from $M(x)$. In other words, for passive sensors, we can calculate βP directly from the kinetic rates $k_{y \to y'}(x)$ and input generator $(\phi_g(\tau),\ \epsilon^+(g,x),\ p(x|g))$ alone.

V. EXAMPLES

Using this theory we now study four sensor-environment examples. The first two—a two-state sensor channel in a binary environment—serves to both check our formulae and quantify the gains in computational efficiency that result from using these formulae. The third studies the performance of large random sensor channels at lossy prediction. The fourth example is a Hill molecule sensor in a fluctuating environment. In this, we wish to find Hill molecules that "optimally" balance the need to predict sensory input with constraints on heat dissipation. Thereafter we compare to prior results on optimized biochemical sensors.

A. Two-state sensors in Markovian environments

When the environment is Markovian, we can check our formula for $p(x, y)$ in two ways. First, we can analytically calculate (using similar ideas) a seemingly different expression for $p(x, y)$. Though an analytical match between the formula given here and the formulae shown below is not obvious, there is an exact numerical match for all tested randomly chosen sensor channels. Next, we simulate the system using the temporal Gillespie algorithm [58], and this results in a nearly exact numerical match with simulations.

The Markovian environment that we consider has two observed symbols with dwell time densities of $\phi_A(t) = 4e^{-4t}$ and $\phi_B(t) = 5e^{-5t}$. This implies that the kinetic rate at which one moves from A to B is $k_{A \to B} = 4$ and that at which one moves from B to A is $k_{B \to A} = 5$.

For a Markovian environment, only the present environmental symbol is needed to predict future environmental states. In the language above, the present environmental symbol is the causal state. As a result, to calculate $p(x, y)$, we need only find the stationary distribution of a master equation with kinetic rates given by:

$$k_{(x,y)\to(x',y')} = \begin{cases} 0 & x \neq x',\ y \neq y' \\ k^{(x)}_{y \to y'} & x = x',\ y \neq y' \\ k_{x \to x'} & x \neq x',\ y = y' \\ -k^{(x)}_{y \to y} - k_{x \to x} & x = x',\ y = y' \end{cases} . \tag{10}$$

We use these kinetic rates to analytically characterize the system's steady state probability $p(x, y)$ as:

$$\frac{dp(x,y)}{dt} = \sum_{x',y'} k_{(x',y')\to(x,y)} p(x',y') \ . \tag{11}$$

That is, the steady state $p(x, y)$ is an eigenvector of eigenvalue 0 of the kinetic-rate matrix $k_{(x,y')\to(x,y)}$. Using this technique, we get an exact match with Eqs. (B5) and (B9) for all randomly chosen two-state channels with the Markovian environment listed above. Ideally, we would find an analytic match in the case of Markovian environments between the two analytic methods. However, we could not find an obvious mapping. The exact numerical match is good evidence that it exists, though.

Finally, we compare theory to simulation using the temporal Gillespie algorithm [58]. The standard Gillespie algorithm [59] assumes constant environments in which the waiting time density to the next transition between sensor states is a simple exponential. When the environment fluctuates, the waiting time density changes from

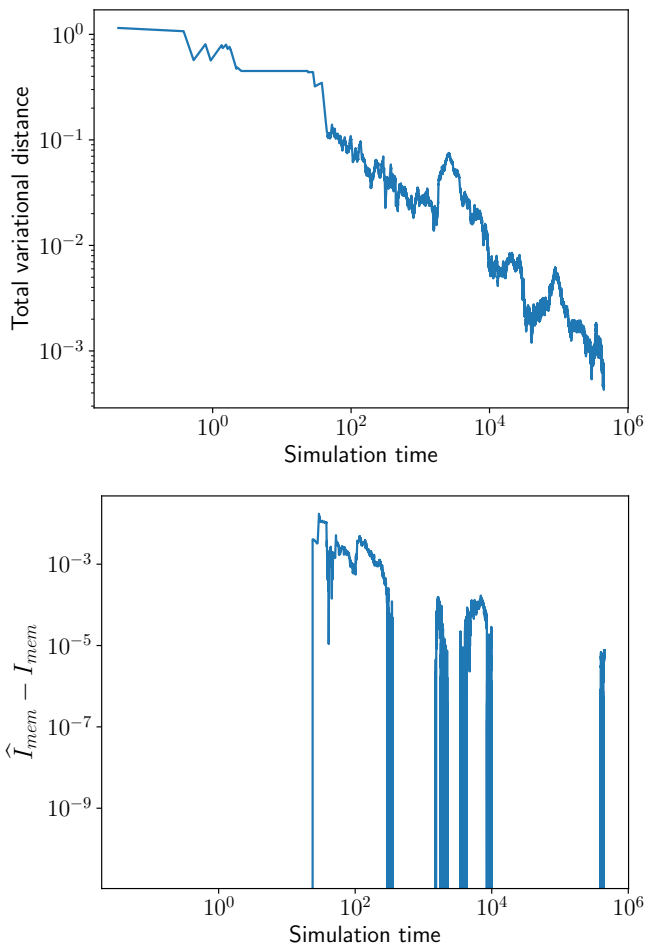

FIG. 3. Two-state sensor in Markov environment—theory versus simulation: (Top) Comparing total variational distance $\sum_{x,y} |p(x,y) - p_{sim}(x,y)|$ between analytic estimate $p(x,y)$ from App. B and simulation estimate $p_{sim}(x,y)$ as a function of simulation time. (Bottom) The corresponding difference $\widehat{I}_{mem} - I_{mem}$ between estimates of the instantaneous memory also decreases as simulation time increases. A random two-state sensor is chosen, such that $k^{(A)}_{B\to A} = 0.711$, $k^{(A)}_{A\to B} = 0.341$, $k^{(B)}_{B\to A} = 0.928$, and $k^{(B)}_{A\to B} = 0.101$. As time increases, error between the true $p(x,y)$ and $p_{sim}(x,y)$ decreases.

a simple exponential to the exponential of an integral of the time-varying kinetic rate out of the current sensor state. In the present case, this time-varying kinetic rate is piecewise-constant. We first generate exponentially distributed random numbers that correspond to the dwell times in the environment and, then, simulate transitions between sensor states using this temporal Gillespie algorithm. The convergence to agreement with analytics is shown in Fig. 3. Error is less than 10^{-3}.

B. Two-state sensors driven by semi-Markov environments

The advantages of the analytic approach championed here are not clear when considering a Markovian environment. There, the more typical analytic approach is faster and more accurate as it requires no integrals.

To illustrate the advantages, we now consider a random two-state sensor channel driven by a semi-Markov environment in which the dwell-time density for symbol A is $\phi_A(t) = 16te^{-4t}$ and the dwell-time density for symbol B is $\phi_B(t) = 25te^{-5t}$. Figure 4 compares the aforementioned temporal Gillespie algorithm to the formulae developed above. In particular, it shows that as time increases, the simulation becomes more accurate, as expected.

The temporal Gillespie simulation takes a surprising amount of time relative to the formulae developed here. The bottleneck comes not in simulating the sensor's response to the environment, but rather, in simulating the environment itself. It takes roughly eight hours to simulate the environment for the total time shown.

C. Random multistate sensors driven by semi-Markov environments

Now consider lossy prediction—in which not all predictive information is captured—by large (multistate) random sensor channels in a semi-Markov environment. Since random channels can achieve coding-theoretic limits [60], random sensors are a class worth exploring. On the one hand, in principle they potentially can store information about an environment's past that is useful for prediction. On the other, they cannot store all the information about the past necessary to predict the future as well as possible [37]. As such, large random sensors are *lossy* predictors.

We ask two questions. First, how well do random sensors perform relative to the optimal possible performance? One can calculate bounds on lossy prediction [36] using a combination of causal-state machinery and the generalized Blahut-Arimoto algorithm. Second, are larger random sensors better or worse at lossy prediction? Potentially, larger sensor channels have more capacity to store information about the environment's past [52], but whether or not they deploy this capacity well is still undetermined.

We subjected random sensors of varying number of states $N \in \{3, 6, 10, 20, 30, 40, 50, 60, 100, 300, 1000\}$ to a semi-Markov environment in which the dwell-time density of symbol A is $\phi_A(\tau) = 16\tau e^{-4\tau}$ and the dwell-time density of symbol B is $\phi_B(\tau) = 25\tau e^{-5\tau}$. The kinetic rate

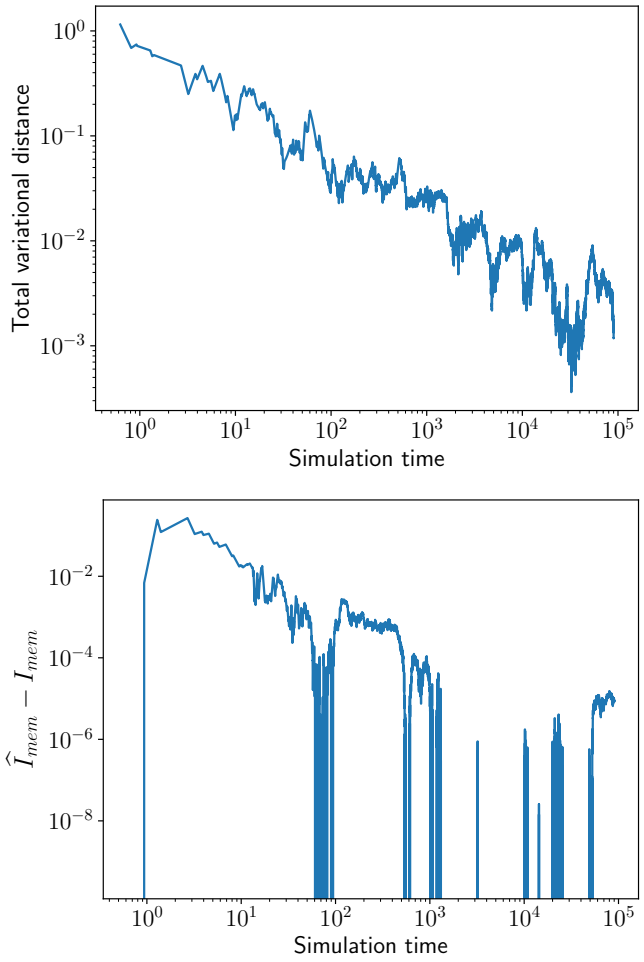

FIG. 4. Two-state sensor in a semi-Markov environment—theory versus simulation: (Top) Total variational distance $\sum_{x,y} |p(x,y) - p_{sim}(x,y)|$ between true $p(x,y)$ and simulation estimate $p_{sim}(x,y)$ as a function of simulation time. As time increases, error between the true $p(x,y)$ and $p_{sim}(x,y)$ decreases. (Bottom) The corresponding difference $\widehat{I}_{mem} - I_{mem}$ between estimates of the instantaneous memory also decreases as simulation time increases. Same kinetic rates as in Fig. 3.

$k^{(x)}_{y \to y'}$ between one sensor state y and another y', conditioned on observing a particular environmental symbol x, is drawn uniformly from the unit interval.

Figure 5 (Top) shows that random sensors are not good lossy predictors in that their total predictable information $\mathrm{I}[Y; \overrightarrow{X}]$ falls three orders of magnitude below the maximal achievable total predictable information for their coding rate $\mathrm{I}[Y; \mathcal{S}^+]$. In using this bound on $\mathrm{I}[Y; \overrightarrow{X}]$, we assumed that one recodes the dynamical system's state so as to eliminate extraneous information about the past using the procedure outlined in the proof of Theorem 1 of Ref. [36]. This achieves a coding rate of $\mathrm{I}[Y; \mathcal{S}^+]$ rather than one of $\mathrm{I}[Y; \overleftarrow{X}]$.

Perhaps more interestingly, Fig. 5 (Top) also shows that as sensor size grows, large random sensors become worse lossy predictors, in that they store information about the environment's forward-time causal states that is not the "right" information to remember for predicting environment futures. This result is only potentially true in a non-Markovian environment, as in a continuous-time Markovian environment every bit stored about the forward-time causal states is useful for understanding environmental futures. Typically, both prediction and memory capacity of random sensors increase with increasing channel size [52]. And so, the lack of efficacy of random sensors is somewhat surprising.

In addition, random sensors at fixed size tend to have similar achievable coding rates $\mathrm{I}[Y; \mathcal{S}^+]$ and total predictable informations $\mathrm{I}[Y; \overrightarrow{X}]$, despite differences between random instantiation. Likely, this results from central limit theorem-like behavior. Though, proving as much in even simpler settings can be difficult [61].

In a semi-Markov environment, the total predictable information I_{fut} is the shared information between our sensor's representation and both the present symbol and the time to next symbol, and the dynamical system (sensor) obtains information about the time to next symbol from the time since last symbol. In particular, the time since last symbol can be used to better predict the time to next symbol, as the time between successive symbols is nonexponential [37]. Figure 5 (Bottom) shows that random channels essentially are sensitive only to information about the current environmental symbol, which is predictive of immediately following environmental symbols.

If large random dynamical systems had been "good enough", then evolutionary search to find predictive sensors would not be needed. Naturally, such a result flies in the face of current thinking [23]. Though many more examples need be considered, the present results for this class demonstrate that one cannot reliably design good lossy predictors simply by generating large randomly wired sensors. This hints at the challenge faced in the evolution of biological sensors.

Note that these results hold no matter the energy function. To better understand their thermodynamic efficiency, we now assume that they are passive sensors, so that the steady-state probability distribution in a fixed environment is equivalent to the Boltzmann distribution. We then calculate both nonpredictive information rate $\dot{\mathrm{I}}_{\mathrm{np}}$ and power consumption βP. While we average the dissipation metric over many channels of the same size, we track the standard deviations in Fig. 6. The mean nonpredictive information rate varies nonmonotonically with channel size, peaking at around $N = 30$. The heat dissipation rate (equivalent to power consumption in nonequilibrium steady state) decreases monotonically with channel size. Therefore, larger channels have

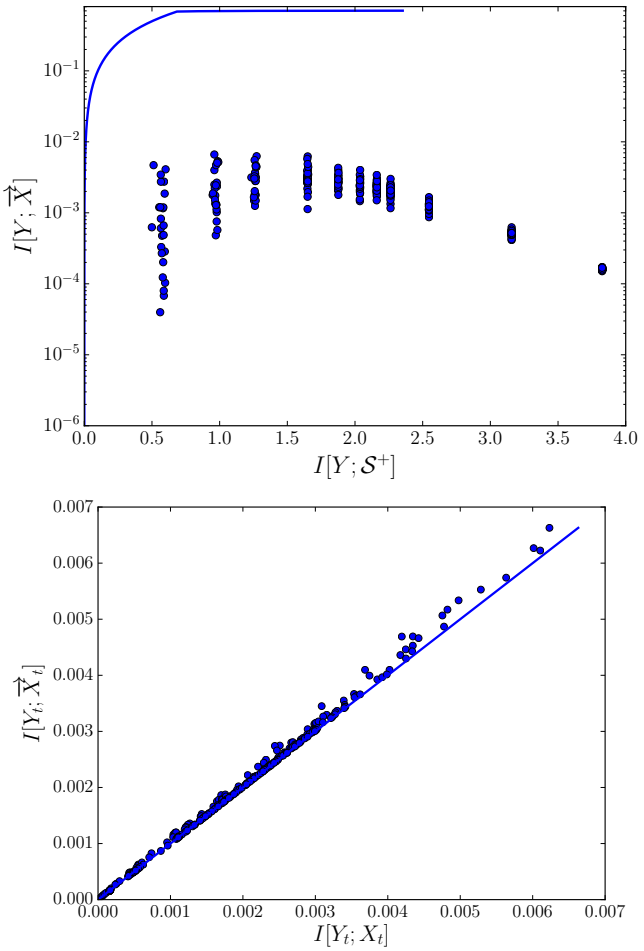

FIG. 5. (Top) Random sensors are not good lossy predictors of semi-Markov stimuli: Blue solid curve is the theoretical bound on total predictable information $\mathrm{I}[Y;\overleftarrow{X}]$ for a given constraint on mutual information $\mathrm{I}[Y;\mathcal{S}^+]$ between past and representation, calculated using the algorithm described in Ref. [36]. Blue dots represent realizations of prediction performance and compression for randomly generated channels with states sets of size $N \in \{3, 6, 10, 20, 30, 40, 50, 60, 100, 300, 1000\}$ from left to right, with 25 random channels per size. (Bottom) Most prediction power comes from correctly storing information about the present environmental symbol: Each dot indicates the combination of instantaneous memory $\mathrm{I}[Y_t;X_t]$ and total predictable information $\mathrm{I}[Y_t;\overleftarrow{X}]$ of a random channel. The solid line corresponds to their equality.

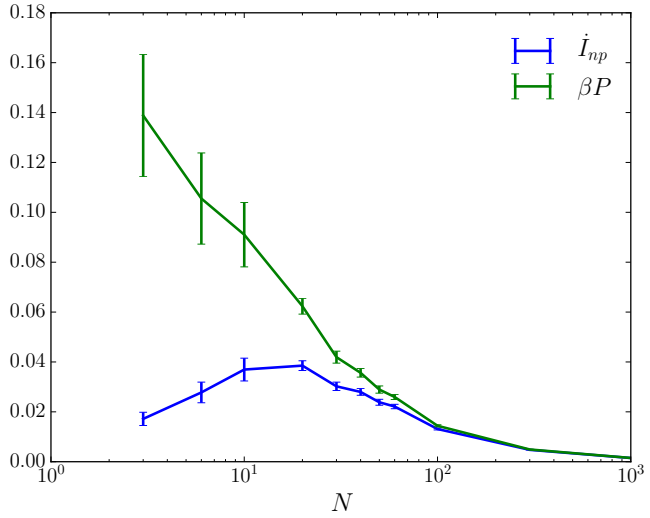

FIG. 6. Dissipation metrics for random channels vary with channel size: Mean nonpredictive information rate $\dot{\mathrm{I}}_{\mathrm{np}}$ and mean power consumption βP (equivalent to heat dissipation rate in nonequilibrium steady state) averaged over 25 random channels at each of several sizes: $N \in \{3, 6, 10, 20, 30, 40, 50, 60, 100, 300, 1000\}$. Error bars give standard deviations across these 25 realized random channels. Higher sensory capacity—larger ratio of nonpredictive information rate to power consumption $\dot{\mathrm{I}}_{\mathrm{np}}/\beta P$—is attained with larger random channels, even though this yields the lower prediction metrics shown in Fig. 5 (Top).

smaller entropy production rates, even though they are emphatically not at equilibrium, as shown by Fig. 5 (Top). The sensory capacity—the ratio of the nonpredictive information rate or learning rate to entropy production rate—increases monotonically with channel size. This qualitatively agrees with Ref. [17] which claimed that adding memory to a sensor increases its sensory capacity. Similar to Ref. [45], then, we find that maximizing sensory capacity and maximizing prediction metrics are at odds. Maximizing sensory capacity yields large random channels, while maximizing prediction metrics from Fig. 5 yields small random channels.

D. Hill molecule interacting with a semi-Markov environment

The Hill molecule—a common fixture in theoretical biochemistry—is the simplest mechanistic model of binding cooperativity [62]. Recall Fig. 1. A Hill molecule can be in one of two states, open or closed. When open, n ligand molecules are bound; when closed, no ligand molecules are bound. Hence, the Hill molecule state carries information about the number of bound ligand molecules. In other words, Hill molecules sense their environment's ligand concentration. Though they seem impoverished in their simplicity, the following concludes that they are sensitive to an order-of-magnitude more total predictable information in their environment than the large random sensors just studied.

Let us outline a simple dynamical model of the Hill molecule's operation. The transition rate from closed C

to open O given a ligand concentration x is:

$$k_{C\to O} = k_O x^n \ , \tag{12}$$

while the rate from open O to closed C is:

$$k_{O\to C} = k_C \ . \tag{13}$$

Given fixed ligand concentration, the steady-state distribution is familiar:

$$\begin{aligned}\Pr_{\text{eq}}(Y = O|X = x) &= \frac{k_O x^n}{k_C + k_O x^n} \\ &= \frac{x^n}{(k_C/k_O) + x^n} \ .\end{aligned}$$

Thus, the rates are determined by the number n of ligands that bind simultaneously—this is called the *cooperativity*. Although the mechanistic model makes sense only when n is a nonnegative integer, this model is often used when n is any nonnegative real number; increases in n can still be thought of as increases in cooperativity. Equations (12) and (13) constitute a complete characterization of its channel properties.

Note that these dynamics obey detailed balance, so no power is required to sustain operation out of equilibrium in a fixed environment. This is a passive sensor for which the power extracted from the environment is identical to the heat dissipation rate, according to the first law of thermodynamics. It may seem *a priori* that sensor designers should not care how much energy a sensor must draw from its environment. However, the power extracted from the environment is dissipated as heat, and increased heat dissipation can require additional biological machinery. With this in mind, it would seem that biology should favor sensors with smaller power consumption.

Increasing the cooperativity n increases the steepness of the molecule's sigmoidal "binding curve"—the probability of being "on" as a function of concentration. In other words, the sensor becomes more switch-like and less a proportionately-responding transducer of the input. If the concentration is greater than $(k_C/k_O)^{1/n}$, the switch is essentially "on" if n is high. A more switch-like sensor is useful if the optimal phenotype depends only upon the condition "ligand concentration greater than X". A less switch-like, smoother-responding sensor helps if the optimal phenotype depends on ligand concentration in a more graded manner.

The concentration scale is set by $(k_C/k_O)^{1/n}$, while the time scale is set by $1/k_C$; as such, we set both to $k_O = k_C = 1$ without loss of generality. We imagine that the ligand concentration alternates between high and low values: x_l and x_h, respectively. When there is less ligand

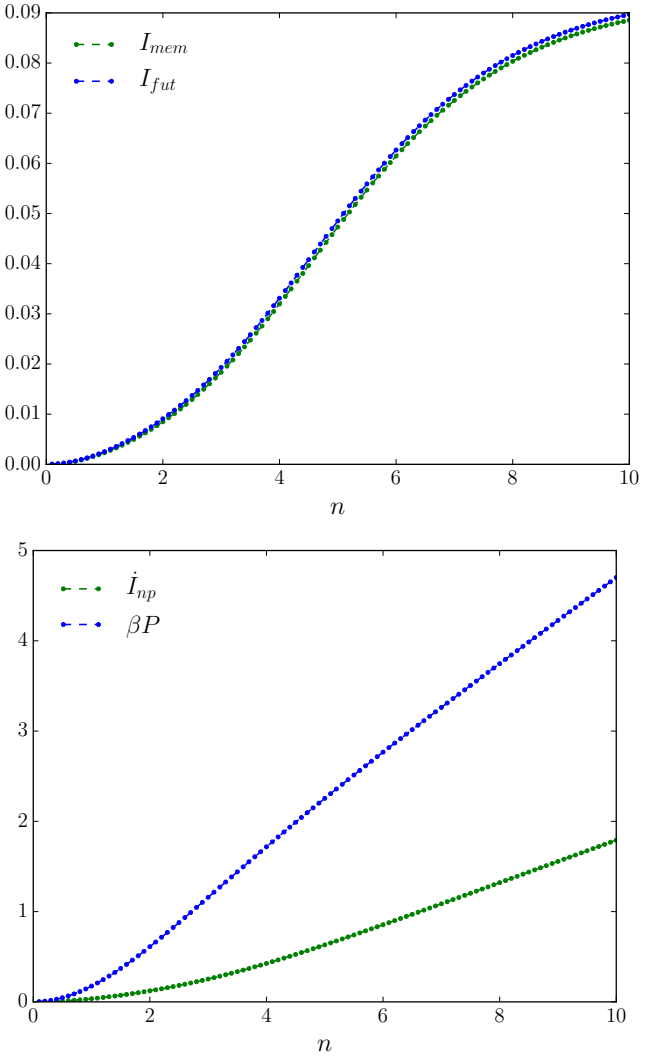

FIG. 7. (Top) Hill molecule prediction-related metrics: Instantaneous memory I_{mem} and total prediction power I_{fut}, as functions of n in nats. (Bottom) Hill molecule temperature-normalized dissipation-related metrics: Nonpredictive information rate $\dot{I}_{np}$ and temperature-normalized power βP, both in nats per unit time. Ligand concentrations take one of two values, $x_l = 0.5$ and $x_h = 2.0$, in units of $(k_C/k_O)^{1/n}$. Dwell-time densities are parametrized as $\phi_x(t) = \lambda(x)^2 t e^{-\lambda(x)t}$ with $\lambda(x_l) = 5$ and $\lambda(x_h) = 4$ in units of $1/k_C$. Hill molecule parameters are set to $k_O = k_C = 1$, with varying cooperativity n.

$(x \approx x_l)$, the Hill molecule reverts to and stays in the closed state. When there is ligand $(x \approx x_h > x_l)$, it reverts to and stays in the open state. With no particular application in mind, we again imagine that the dwell-time densities take the form $\phi_x(\tau) = \lambda(x)^2 \tau e^{-\lambda(x)\tau}$ with $\lambda(x_l) = 5$ and $\lambda(x_h) = 4$.

We now deploy the earlier formulae to study the prediction capabilities and dissipation tendencies of a Hill molecule subject to semi-Markov input. Previous stud-

ies of biological sensors found that increases in cooperativity accompanied increases in channel capacity [63–65] and sensor accuracy [39] in equilibrium systems. Others studied the thermodynamics of prediction of cooperative biological sensors [17, 31], but did not use the more general class of semi-Markov input which, we argued, is more typical of real environmental stimuli. They also did not calculate the total predictable information.

In the absence of a reward function, we assert that energetic rewards are proportional to I_{fut} [54, 55]. Altogether, this argues that $\alpha \, \mathrm{I}_{\mathrm{fut}} - \beta P$ is a reasonable objective function for sensor design, similar to that of Ref. [42]. However, unlike there, we hold the environment fixed and do not consider the scenario in which the environment tries to adversely impact the total predictable information captured by the sensor. The proportionality constant α is a conversion factor between information and energy units. It is ultimately set by the type of environment in which the sensor finds itself. It determines the energetic reward for prediction. All said, a sensor that maximizes this objective is predictive of its input and does not dissipate much heat.

An example—$x_l = 0.5$, $x_h = 2.0$, $k_O = k_C = 1.0$, and $n = 2$—illustrates that roughly 99% of $\mathrm{I}_{\mathrm{fut}}$ is devoted to capturing instantaneous memory $\mathrm{I}_{\mathrm{mem}}$ and roughly 25% of βP is devoted to $\dot{\mathrm{I}}_{\mathrm{np}}$. That is, the inefficiency in choosing what information to store about the present input contributes greatly to energetic inefficiency. These results hold qualitatively even when the dwell-time densities are log-normal; i.e., heavier-tailed.

Figure 7 shows that increased cooperativity—that is, increases in n—lead to increases in prediction performance, qualitatively in line with Refs. [42, 65]. Additionally, the larger the cooperativity, the higher the fraction $\mathrm{I}_{\mathrm{mem}} \, / \, \mathrm{I}_{\mathrm{fut}}$. Larger cooperativity, however, leads to roughly linear increases in the power consumption and nonpredictive information rate. Whereas, increases in prediction power take a more sigmoidal shape. The correlation between prediction power and power consumption are qualitatively in line with the results of Ref. [38].

This suggests a preference for intermediate values of cooperativity (e.g., $n \approx 5$) over larger values of cooperativity (e.g., $n \geq 10$). This is qualitatively similar to the results of Refs. [63, 64] in that physical constraints force optimal information transmission at intermediate levels of cooperativity. Notice that instantaneous memory and total predictable information sit at around 0.05 bits for Hill factors $n \approx 4$. This is an order of magnitude higher than the instantaneous memory and total predictable information of the large random sensors above.

The sensor's objective function $\alpha I_{fut} - \beta P$ is optimized by the cooperativity $\widehat{n} = \arg\max_n (\alpha I_{fut} - \beta P)$. As both I_{fut} and βP increase monotonically with n, there

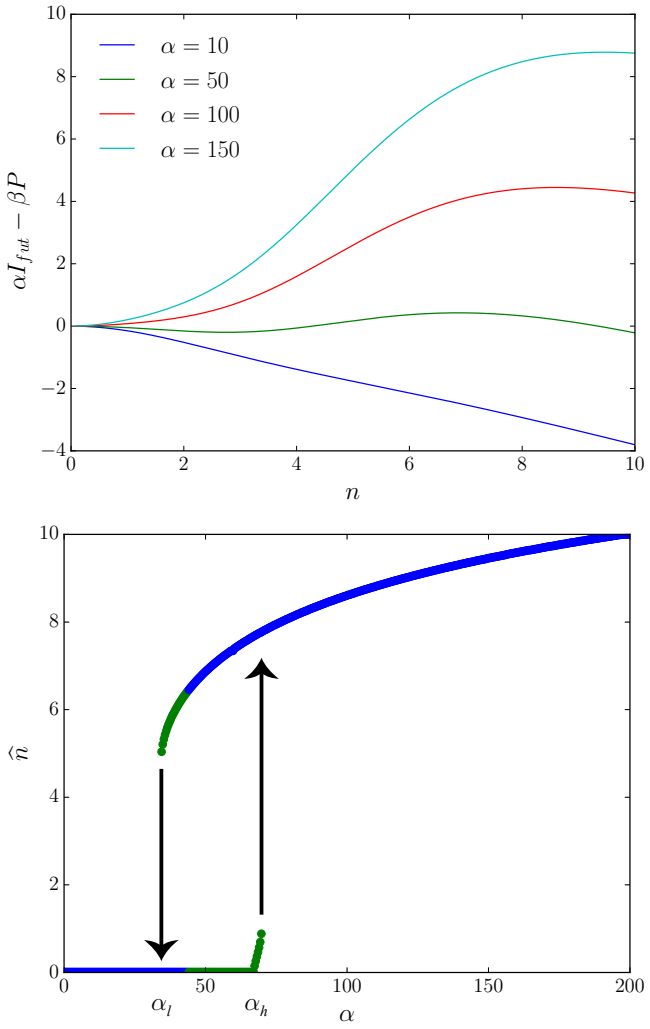

FIG. 8. (Top) "Lagrangian" $\alpha I_{fut} + \beta P$ for several α—that represent the energy rewards of I_{fut}—as a function of cooperativity n. Notice that a particular α singles out a particular optimal $\widehat{n}$. (Bottom) Optimal cooperativity $\widehat{n}$ as a function of conversion factor α. Circles (solid blue) are the values of the global maximum $\widehat{n}$ and circles (solid green) are local maxima. Arrows indicate the directionality in the hysteresis loop: $\widehat{n}$ jumps up at the upper value α_h, if starting at low α, and jumps down at the lower value α_l, if starting at high α.

is generically only one such cooperativity $\widehat{n}$. At lower α, though, there are two local maxima of the function of n given by $\alpha I_{fut} - \beta P$, as shown at Fig. 8 (Top). Let's pursue the consequences of this regime dependence.

There are rules for how sensor biochemical parameters adapt to the present environment. If adaptation rules for cooperativity of a Hill molecule increase the total energy budget by gradient descent then, for a range of α, we expect cooperativity to be in either of the two local maxima of $\alpha \, \mathrm{I}_{\mathrm{fut}} - \beta P$ just noted. Let's assume a separation of timescales—namely, that cooperativity adapts much more slowly than the longest environmental timescale

and that the sensor does not adapt quickly enough to always find the global maximum in cooperativity space.

Figure 8 (Bottom) shows the optimal cooperativity $\widehat{n}$ as a function of conversion factor α. As expected from the presence of two local maxima at lower α, there is a discontinuity (*supercritical bifurcation* [66]) in the function of α given by $\widehat{n} = \arg\max_n(\alpha I_{fut} - \beta P)$. Thus, initially, if the energetic reward α for prediction increases, the cooperativity discontinuously jumps to a higher $\widehat{n}$ at a critical α_h. From there, if one decreases α, optimal cooperativity slowly decreases, but stays high well below α_h, suddenly decreasing to zero cooperativity at the lower value α_l. Thus, there is a substantial hysteresis loop built into the optimal trade-off between energy and sensitivity.

What could be the functional benefit of this hysteresis? Recall that in switching circuits hysteresis is essential to adding stability to a switch's response. Hysteresis stops "race" conditions in which the switch oscillates wildly just as the threshold is passed, amplifying any noise in the control and internal dynamics. In the Hill molecule, hysteresis is helpful if a memory of past environmental conditions (α) provides insight into future conditions (future α). For example, the environment might shift α suddenly to being low, but there is a replenishment mechanism for the available energy that will soon increase α again. Thus, we see that robustness to environmental noise emerges naturally if a Hill molecule sensor adapts to and anticipates changing external conditions.

VI. CONCLUSION

We provided closed-form expressions for instantaneous memory, total predictable information, nonpredictive information rate, and power extracted from the environment for a conditionally Markovian channel subject to the broad class of very complex environments described by unifilar hidden semi-Markov input. The information-theoretic quantities among that list are best motivated by Refs. [16, 53].

We showed that determining these closed-form expressions requires knowledge of the environment's causal states. In some cases, one has generated the environment, and so a reasonable model for its dynamics is known. In other cases, the environment in question is well-studied and a reasonable model for its dynamics is known. Causal states can be straightforwardly derived from these models. In other cases, though, reasonable dynamical models might need to be inferred from data. This is quite a difficult problem on which much work has been done, e.g. Refs. [67–69].

It is not necessary to use the closed-form expressions in Sec. IV to calculate prediction and dissipation metrics. Instead, one may numerically estimate them as follows:

1. Generate a realization of the environment, and
2. Simulate transitions between states using the temporal Gillespie algorithm [58].

This approach has its challenges, though. Generating realizations of non-Markovian environments, as considered here, is straightforward but quite compute intensive, if one desires accurate simulations. (The computational bottleneck arises in determining the waiting time density whose average is some randomly chosen number.) Simulating transitions between Hill molecule states given an environmental realization is straightforward, too. However, surprisingly long simulation times were required to match analytic results for even two-state channels to two significant-digit accuracy in $p(x, y)$. These challenges speak to the difficulty of studying the second two examples—large random channels and the Hill molecule—through simulation.

This highlights the benefits of Sec. IV's closed-form expressions. Their accuracy, though, is limited by that of the numerical algorithms for calculating integrals and finding eigenvectors. Available routines exhibit several weaknesses in accurately evaluating the requisite integrals when, for instance, $\phi_x(\tau)$ is heavy-tailed. For the dwell-time densities considered here, there were no such issues.

All in all, the ease with which the sensor metrics were numerically estimated masks the difficulty of deriving the underlying closed-form expressions in the first place. Appendices B and C give those derivations. That said, leveraging this one-time calculation effort, we provided universal estimators of prediction and dissipation metrics for conditionally Markovian channels. The universality claimed here arises from the fact that unifilar hidden semi-Markov processes are very general memoryful processes that capture highly complex environmental behaviors.

One practical consequence going forward is that analyzing sensor prediction and dissipation no longer requires simulating arbitrarily long trajectories. Instead, one can now validate or invalidate predictive learning rules and sensor designs using the universal estimators. This will also greatly accelerate searching through parameter space for "optimal" (predictive and energy-efficient) sensors. In addition, given that the theories of random dynamical systems and of input-dependent dynamical systems are still under development [70], we believe the formulae presented here will eventually lead in those domains to a precise generalization of time-scale matching for nonlinear systems [31]. Finally, we hope to extend these new calculational techniques to nonsta-

tionary environments and thereafter to analyze sensory adaptation in more familiar scenarios, as in Refs. [43, 44].

ACKNOWLEDGMENTS

We thank N. Ay, A. Bell, W. Bialek, S. Dedeo, C. Hillar, I. Nemenman, and S. Still for useful conversations and the Santa Fe Institute and the Telluride Science Research Center for their hospitality during visits, where JPC is an External Faculty and a General Member, respectively. This material is based upon work supported by, or in part by, the John Templeton Foundation grant 52095, the Foundational Questions Institute grant FQXi-RFP-1609, the U. S. Army Research Laboratory and the U. S. Army Research Office under contract W911NF-13-1-0390. S.E.M. was funded by an MIT Physics of Living Systems Fellowship.

Appendix A: Revisiting the Thermodynamics of Prediction

For completeness, we review the derivation of Eq. (5). Let x_t represent the input at time t, y_t represent the sensor state at time t, and $E(x, y)$ the system's energy function. We assume constant temperature. The system's temperature-normalized nonequilibrium free energy F_{neq} is given by:

$$\beta F_{neq}[p(y|x)] = \beta \langle E(x, y) \rangle_{p(y|x)} - \mathrm{H}[Y|X = x] \ . \tag{A1}$$

Even if this is not a valid expression for nonequilibrium free energy, the validity of Ref. [16]'s derivation only rests on this expression being a Lyapunov function for the dynamics. Intuitively, this corresponds to an assumption that the system reduces its nonequilibrium free energy when the sensor thermalizes to its attached thermal bath. (Accordingly, the β in the above expression refers to the temperature of the sensor when the environment is fixed, indirectly circumventing the difficulty with defining a nonequilibrium temperature [71].) If so, then:

$$\beta F_{neq}[p(y_t|x_{t+\Delta t})] \geq \beta F_{neq}[p(y_{t+\Delta t}|x_{t+\Delta t})] \ ,$$

giving:

$$\begin{aligned}
0 &\leq \beta F_{neq}[p(y_t|x_{t+\Delta t})] - \beta F_{neq}[p(y_{t+\Delta t}|x_{t+\Delta t})] \\
&\leq \left(\beta \langle E(x_{t+\Delta t}, y_t) \rangle_{p(y_t|x_{t+\Delta t})} - \mathrm{H}[Y_t|X_{t+\Delta t} = x_{t+\Delta t}]\right) - \left(\beta \langle E(x_{t+\Delta t}, y_{t+\Delta t}) \rangle_{p(y_{t+\Delta t}|x_{t+\Delta t})} - \mathrm{H}[Y_{t+\Delta t}|X_{t+\Delta t} = x_{t+\Delta t}]\right) \\
&\leq \beta \lim_{\Delta t \to 0} \left(\frac{\langle E(x_{t+\Delta t}, y_t) \rangle_{p(y_t|x_{t+\Delta t})}}{\Delta t} - \frac{\langle E(x_{t+\Delta t}, y_{t+\Delta t}) \rangle_{p(y_{t+\Delta t}|x_{t+\Delta t})}}{\Delta t} \right) \\
&\quad + \lim_{\Delta t \to 0} \left(\frac{\mathrm{H}[Y_{t+\Delta t}|X_{t+\Delta t} = x_{t+\Delta t}]}{\Delta t} - \frac{\mathrm{H}[Y_t|X_{t+\Delta t} = x_{t+\Delta t}]}{\Delta t} \right) \ .
\end{aligned}$$

Finally, we average over possible environmental realizations, equivalent in nonequilibrium steady states (NESSs) to averages over time, to find:

$$0 \leq \beta \lim_{\Delta t \to 0} \frac{\langle E(x_{t+\Delta t}, y_t) \rangle - \langle E(x_{t+\Delta t}, y_{t+\Delta t}) \rangle}{\Delta t} + \lim_{\Delta t \to 0} \frac{\mathrm{H}[Y_{t+\Delta t}|X_{t+\Delta t}] - \mathrm{H}[Y_t|X_{t+\Delta t}]}{\Delta t} \ .$$

The former term could be called $-\dot{Q}$—the negative of the sensor's heat dissipation rate—and the latter $\dot{\mathrm{I}}_{\mathrm{lost}}$—the rate of lost information. And so:

$$0 \leq -\beta \dot{Q} + \dot{\mathrm{I}}_{\mathrm{lost}} \ . \tag{A2}$$

This is valid even outside of NESSs. In a NESS, though, we can invoke stationarity, concluding that:

$$\dot{Q} = -P \ , \tag{A3}$$

where P is the power extracted by the sensor from the environment. Furthermore, $\dot{\mathrm{I}}_{\mathrm{lost}}$ reduces to the negative of the nonpredictive information rate, since $\mathrm{H}[Y_{t+\Delta t}|X_{t+\Delta t}] = \mathrm{H}[Y_t|X_t]$, giving:

$$\dot{\mathrm{I}}_{\mathrm{lost}} = \lim_{\Delta t \to 0} \frac{\mathrm{H}[Y_t|X_t] - \mathrm{H}[Y_t|X_{t+\Delta t}]}{\Delta t} \ .$$

Calling on a standard information theory identity [72]—$\mathrm{H}[U|V] = \mathrm{H}[U] - \mathrm{I}[U;V]$—leads to:

$$\dot{\mathrm{I}}_{\mathrm{lost}} = \lim_{\Delta t \to 0} \frac{\mathrm{I}[Y_t;X_{t+\Delta t}] - \mathrm{I}[Y_t;X_t]}{\Delta t} \ .$$

We recognize this as the continuous-time version of the nonpredictive information rate $\dot{\mathrm{I}}_{\mathrm{np}}$; also called the learning rate. Hence, the nonpredictive information rate is the increase in unpredictability of sensor state Y_t given a slightly-delayed environmental state:

$$\dot{\mathrm{I}}_{\mathrm{np}} = \lim_{\Delta t \to 0} \frac{\mathrm{H}[Y_t|X_{t+\Delta t}] - \mathrm{H}[Y_t|X_t]}{\Delta t} \ . \tag{A4}$$

In a NESS, then:

$$\dot{\mathrm{I}}_{\mathrm{lost}} = -\dot{\mathrm{I}}_{\mathrm{np}} \ . \tag{A5}$$

Outside of NESSs, these terms are augmented by the time-derivative $d\,\mathrm{H}[Y_t|X_t]/dt$ of the conditional entropy. This leads to the addition of an Landauer-erasure information [73] when integrated. One of Ref. [16]'s main results follows directly from Eqs. (A2), (A3), and (A5):

$$\dot{\mathrm{I}}_{\mathrm{np}} = \lim_{\Delta t \to 0} \frac{\mathrm{I}[Y_t;X_t] - \mathrm{I}[Y_t;X_{t+\Delta t}]}{\Delta t} \leq \beta P \ . \tag{A6}$$

In contrast to Ref. [16]'s implication, this is true only in a NESS and Eq. (A2) should be used otherwise.

Differences in presentation between the derivation here and that of Ref. [16] come from the difference between discrete- and continuous-time formulations. To make this clear, we present a continuous-time formulation of the same result, following Ref. [18]. We start from $\beta F_{neq}[p(y_{t'}|x_t)]$ being a Lyapunov function in t':

$$\begin{aligned}
0 &\geq \beta \frac{\partial F_{neq}[p(y_{t'}|x_t)]}{\partial t'} \\
&= \beta \frac{\partial}{\partial t'} \left\langle E(x_t, y_{t'}) \right\rangle_{p(y_{t'}|x_t)} \Big|_{t'=t} - \frac{\partial}{\partial t'} \mathrm{H}[Y_{t'}|X_t = x_t]|_{t'=t} \\
&= \beta \left\langle \frac{\partial E(x_t, y_t)}{\partial y_t} \dot{y}_t \right\rangle_{p(y_{t'}|x_t)} - \frac{\partial}{\partial t'} \mathrm{H}[Y_{t'}|X_t = x_t]|_{t'=t} \ .
\end{aligned}$$

Next, as before, we average over protocols (or, equivalently in NESS, over time) to find:

$$0 \geq \beta \left\langle \frac{\partial E(x_t, y_t)}{\partial y_t} \dot{y}_t \right\rangle_{p(x_t, y_t)} - \frac{\partial}{\partial t'} \mathrm{H}[Y_{t'}|X_t]|_{t'=t}$$

We then recognize $\beta \left\langle \frac{\partial E(x_t, y_t)}{\partial y_t} \dot{y}_t \right\rangle_{p(x_t, y_t)}$ as the temperature-normalized rate of heat dissipation $\beta \dot{Q}$, so that:

$$\beta \dot{Q} \leq \frac{\partial}{\partial t'} \mathrm{H}[Y_{t'}|X_t]|_{t'=t} \ .$$

The quantity on the righthand side is simply the rate $\dot{\mathrm{I}}_{lost}$ of information loss, defined earlier. In NESS, $d\langle E \rangle/dt$ and $d\,\mathrm{H}[Y_t|X_t]/dt$ vanish. As a result, $\beta \dot{Q} + \beta P = 0$ and:

$$\frac{\partial}{\partial t'} \mathrm{H}[Y_{t'}|X_t]|_{t'=t} = -\frac{\partial}{\partial t'} \mathrm{H}[Y_t|X_{t'}]|_{t'=t} \ ,$$

giving:

$$\beta P \geq \frac{\partial}{\partial t'} \mathrm{H}[Y_t|X_{t'}]|_{t'=t} \ . \tag{A7}$$

We recognize this as the continuous-time formulation of Eq. (A4). Again invoking stationarity, $d\,\mathrm{H}[X_t]/dt$ vanishes and so:

$$\beta P \geq -\frac{\partial}{\partial t'} \mathrm{I}[X_{t'}; Y_t]|_{t'=t} \ , \tag{A8}$$

the continuous-time formulation of Eq. (A6). We have, in Eqs. (A4), (A6), (A7), and (A8), four equivalent definitions for the nonpredictive information rate in the NESS limit.

Appendix B: Closed-form Expressions for Unifilar Hidden Semi-Markov Environments

To find $\rho(\sigma^+, y)$, we start with the following:

$$\begin{aligned}
&\Pr(\mathcal{S}^+_{t+\Delta t} = (g, x, \tau), Y_{t+\Delta t} = y) \\
&\quad = \sum_{g', x, ', \tau', y'} \Pr(\mathcal{S}^+_{t+\Delta t} = (g, x, \tau), Y_{t+\Delta t} = y | \mathcal{S}^+_t = (g', x', \tau'), Y_t = y') \Pr(\mathcal{S}^+_t = (g', x', \tau'), Y_t = y') \ .
\end{aligned} \tag{B1}$$

We decompose the transition probability using the lack of feedback as:

$$\begin{aligned}
&\Pr(\mathcal{S}^+_{t+\Delta t} = (g, x, \tau), Y_{t+\Delta t} = y | \mathcal{S}^+_t = (g', x', \tau'), Y_t = y') \\
&\quad = \Pr(\mathcal{S}^+_{t+\Delta t} = (g, x, \tau) | \mathcal{S}^+_t = (g', x', \tau')) \Pr(Y_{t+\Delta t} = y | \mathcal{S}^+_t = (g', x', \tau'), Y_t = y') \ .
\end{aligned}$$

From the setup, we have:

$$\Pr(Y_{t+\Delta t} = y | \mathcal{S}^+_t = (g', x', \tau'), Y_t = y') = \begin{cases} k_{y' \to y}(x')\Delta t & y \neq y' \\ 1 - k_{y' \to y'}(x')\Delta t & y = y' \end{cases} \ ,$$

with corrections of $O(\Delta t^2)$.

Now split this into two cases. As long as $\tau > \Delta t$, so that $x = x'$, we have:

$$\Pr(\mathcal{S}^+_{t+\Delta t} = (g, x, \tau) | \mathcal{S}^+_t = (g', x', \tau')) = \frac{\Phi_g(\tau)}{\Phi_g(\tau')} \delta(\tau - (\tau' + \Delta t)) \delta_{x,x'} \delta_{g,g'} \ .$$

Then, Eq. (B1) reduces to:

$$\begin{aligned}
&\Pr(\mathcal{S}^+_{t+\Delta t} = (g,x,\tau), Y_{t+\Delta t} = y) \\
&\qquad = \sum_{y'} \Pr(\mathcal{S}^+_{t+\Delta t} = (g,x,\tau) | \mathcal{S}^+_t = (g,x,\tau-\Delta t)) \Pr(Y_{t+\Delta t} = y | \mathcal{S}^+_t = (g,x,\tau-\Delta t), Y_t = y') \\
&\qquad\quad \times \Pr(\mathcal{S}^+_t = (g,x,\tau-\Delta t), Y_t = y') \\
&\qquad = \sum_{y' \neq y} \Pr(\mathcal{S}^+_{t+\Delta t} = (g,x,\tau) | \mathcal{S}^+_t = (g,x,\tau-\Delta t)) \Pr(Y_{t+\Delta t} = y | \mathcal{S}^+_t = (g,x,\tau-\Delta t), Y_t = y') \\
&\qquad\quad \times \Pr(\mathcal{S}^+_t = (g,x,\tau-\Delta t), Y_t = y') \\
&\qquad\quad + \Pr(\mathcal{S}^+_{t+\Delta t} = (g,x,\tau) | \mathcal{S}^+_t = (g,x,\tau-\Delta t)) \Pr(Y_{t+\Delta t} = y | \mathcal{S}^+_t = (g,x,\tau-\Delta t), Y_t = y) \\
&\qquad\quad \times \Pr(\mathcal{S}^+_t = (g,x,\tau-\Delta t), Y_t = y) \\
&\qquad = \sum_{y' \neq y} \frac{\Phi_g(\tau)}{\Phi_g(\tau-\Delta t)} k_{y' \to y}(x) \Pr(\mathcal{S}^+_t = (g,x,\tau-\Delta t), Y_t = y') \Delta t \\
&\qquad\quad + \frac{\Phi_g(\tau)}{\Phi_g(\tau-\Delta t)} \left(1 - k_{y \to y}(x) \Delta t\right) \Pr(\mathcal{S}^+_t = (g,x,\tau-\Delta t), Y_t = y) \ , \qquad \text{(B2)}
\end{aligned}$$

plus terms of $O(\Delta t^2)$. We Taylor expand $\Phi_g(\tau + \Delta t) = \Phi_g(\tau) - \phi_g(\tau)\Delta t$ to find:

$$\frac{\Phi_g(\tau)}{\Phi_g(\tau - \Delta t)} = 1 - \frac{\phi_g(\tau)}{\Phi_g(\tau)} \Delta t\text{`},$$

plus terms of $O(\Delta t^2)$. And, similarly, assuming differentiability, we write:

$$\Pr(\mathcal{S}^+_t = (g,x,\tau-\Delta t), Y_t = y') = \Pr(\mathcal{S}^+_t = (g,x,\tau), Y_t = y') - \frac{d}{d\tau} \Pr(\mathcal{S}^+_t = (g,x,\tau), Y_t = y') \Delta t\text{`},$$

plus terms of $O(\Delta t^2)$. Substitution into Eq. (B2) then gives:

$$\begin{aligned}
\Pr(\mathcal{S}^+_{t+\Delta t} = (g,x,\tau), Y_{t+\Delta t} = y) = &\left(\sum_{y' \neq y} k_{y' \to y}(x) \Pr(\mathcal{S}^+_t = (g,x,\tau), Y_t = y') \right) \Delta t + \Pr(\mathcal{S}^+_t = (g,x,\tau), Y_t = y) \\
&- \frac{d \Pr(\mathcal{S}^+_t = (g,x,\tau), Y_t = y)}{d\tau} \Delta t - \frac{\phi_g(\tau)}{\Phi_g(\tau)} \Pr(\mathcal{S}^+_t = (g,x,\tau), Y_t = y) \Delta t \\
&- k_{y \to y}(x) \Pr(\mathcal{S}^+_t = (g,x,\tau), Y_t = y) \Delta t \ ,
\end{aligned}$$

plus terms of $O(\Delta t^2)$. For notational ease, we denote:

$$\rho((g,x,\tau), y) := \Pr(\mathcal{S}^+_t = (x,\tau), Y_t = y) \ ,$$

which is equal to $\Pr(\mathcal{S}^+_{t+\Delta t} = (g,x,\tau), Y_{t+\Delta t} = y)$ since we assumed the system is in a NESS. Then we have:

$$\begin{aligned}
\rho((g,x,\tau), y) = &\left(\sum_{y' \neq y} k_{y' \to y}(x) \rho((g,x,\tau), y') \right) \Delta t + \rho((g,x,\tau), y) - \frac{d\rho((g,x,\tau), y)}{d\tau} \Delta t \\
&- \frac{\phi_g(\tau)}{\Phi_g(\tau)} \rho((g,x,\tau), y) \Delta t - k_{y \to y}(x) \rho((g,x,\tau), y) \Delta t
\end{aligned}$$

plus corrections of $O(\Delta t^2)$. We are left equating the coefficient of the $O(\Delta t)$ term to 0:

$$\frac{d\rho((g,x,\tau), y)}{d\tau} = \sum_{y' \neq y} k_{y' \to y}(x) \rho((g,x,\tau), y') - \frac{\phi_g(\tau)}{\Phi_g(\tau)} \rho((g,x,\tau), y) - k_{y \to y}(x) \rho((g,x,\tau), y) \ . \qquad \text{(B3)}$$

Our task is simplified if we separate:

$$\rho((g,x,\tau),y) = p(y|g,x,\tau)\rho(g,x,\tau)$$

and if we recall that:

$$\rho(g,x,\tau) = \mu_g \Phi_g(\tau) p(g) p(x|g) \ .$$

These give:

$$\frac{d\rho(g,x,\tau)}{d\tau} = -\mu_g \phi_g(\tau) p(x) p(x|g). \tag{B4}$$

Plugging Eq. (B4) into Eq. (B3) yields:

$$\begin{aligned}\frac{dp(y|x,\tau)}{d\tau}\rho(g,x,\tau) - \mu_g\phi_g(\tau)p(g)p(x|g)p(y|g,x,\tau) \\ = \sum_{y'\neq y} k_{y'\to y}(x)\rho(g,x,\tau)p(y'|g,x,\tau) - \frac{\phi_g(\tau)}{\Phi_g(\tau)}\rho(g,x,\tau)p(y|g,x,\tau) - k_{y\to y}(x)\rho(g,x,\tau)p(y|g,x,\tau) \ ,\end{aligned}$$

where we note that:

$$\mu_g\phi_g(\tau)p(g)p(x|g)p(y|g,x,\tau) = \frac{\phi_g(\tau)}{\Phi_g(\tau)}\rho(g,x,\tau)p(y|g,x,\tau) \ .$$

Hence, we are left with:

$$\frac{dp(y|g,x,\tau)}{d\tau} = \sum_{y'\neq y} k_{y'\to y}(x)p(y'|g,x,\tau) - k_{y\to y}(x)p(y|g,x,\tau) \ .$$

We can summarize this ordinary differential equation in matrix-vector notation as follows. Let $\vec{v}(g,x,\tau)$ be the vector:

$$\vec{v}(g,x,\tau) := \begin{pmatrix} p(y_1|g,x,\tau) \\ \vdots \\ p(y_{|\mathcal{Y}|}|g,x,\tau)\end{pmatrix} \ .$$

We have:

$$\frac{d\vec{v}}{d\tau} = M(x)\vec{v} \ ,$$

with solution:

$$\vec{v}(g,x,\tau) = e^{M(x)\tau}\vec{v}(g,x,0) \ . \tag{B5}$$

The structure of $M(x)$ guarantees that probability is conserved, as long as $1^\top \vec{v}(g,x,0) = 1$ for all $x \in \mathcal{A}$.

Our next task is to find expressions for $\vec{v}(g,x,0)$. We do this by considering Eq. (B1) in the limit that $\tau < \Delta t$. More straightforwardly, we consider the equation:

$$\rho((g,x,0),y) = \sum_{g',x'}\int_0^\infty d\tau \ \frac{\phi_{g'}(\tau)}{\Phi_{g'}(\tau)}\delta_{g,\epsilon^+(g',x')}p(x|g)\rho((g',x',\tau),y) \ , \tag{B6}$$

which is based on the following logic. For probability to flow into $\rho((g,x,0),y)$ from $\rho((g',x',\tau),y')$, we need the dwell time for symbol x' to be exactly τ and for $y' = y$. (The latter comes from the unlikelihood of switching both channel

state and input symbol at the same time.) Again decomposing:

$$\begin{aligned}\rho((g',x',\tau),y) &= p(y|g',x',\tau)\rho(g',x',\tau) \\ &= \mu_{g'}\Phi_{g'}(\tau)p(g')p(x'|g')p(y|g',x',\tau)\end{aligned} \tag{B7}$$

and, thus, as a special case:

$$\rho((g,x,0),y) = p(y|g,x,0)p(g)p(x|g)\mu_g \ . \tag{B8}$$

Plugging both Eqs. (B7) and (B8) into Eq. (B6), we find:

$$\begin{aligned}\mu_g p(g)p(x|g)p(y|g,x,0) &= \sum_{g',x'}\int_0^\infty \mu_{g'}p(g')p(x'|g')\phi_{g'}(\tau)\delta_{g,\epsilon^+(g',x')}p(x|g)p(y|g',x',\tau)d\tau \\ \mu_g p(g)p(y|g,x,0) &= \sum_{g',x'}\int_0^\infty \mu_{g'}p(g')p(x'|g')\phi_{g'}(\tau)\delta_{g,\epsilon^+(g',x')}p(y|g',x',\tau)d\tau \ .\end{aligned}$$

Using Eq. (B5), we see that $p(y|g',x',\tau) = \left(e^{M(x')\tau}\vec{v}(g',x',0)\right)_y$ and $p(y|g,x,0) = \left(\vec{v}(g,x,0)\right)_y$. So, we have:

$$\mu_g p(g)\vec{v}(g,x,0) = \sum_{g',x'}\mu_{g'}\delta_{g,\epsilon^+(g',x')}p(g')p(x'|g')\left(\int_0^\infty \phi_{g'}(\tau)e^{M(x')\tau}d\tau\right)\vec{v}(g',x',0) \ .$$

If we form the composite vector:

$$\begin{aligned}\vec{U} &= \begin{pmatrix}\vec{u}(g_1,x_1)\\ \vec{u}(g_1,x_2)\\ \vdots \\ \vec{u}(g_{|\mathcal{G}|},x_{|\mathcal{A}|})\end{pmatrix} \\ &= \begin{pmatrix}\mu_{g_1}p(g_1)\vec{v}(g_1,x_1,0)\\ \vdots \\ \mu_{g_{|\mathcal{G}|}}p(g_{|\mathcal{G}|})\vec{v}(g_{|\mathcal{G}|},x_{|\mathcal{A}|},0)\end{pmatrix}\end{aligned}$$

and the matrix (written in block form) as:

$$\mathbf{C} := \begin{pmatrix}C_{(g_1,x_1)\to(g_1,x_1)} & C_{(g_1,x_2)\to(g_1,x_1)} & \cdots \\ C_{(g_1,x_1)\to(g_1,x_2)} & C_{(g_1,x_2)\to(g_1,x_2)} & \cdots \\ \vdots & \vdots & \ddots\end{pmatrix} \ ,$$

with:

$$C_{(g',x')\to(g,x)} = \delta_{g,\epsilon^+(g',x')}p(x'|g')\int_0^\infty \phi_{g'}(t)e^{M(x')t}dt \ ,$$

we then have:

$$\vec{U} = \mathrm{eig}_1(\mathbf{C}) \ . \tag{B9}$$

Finally, we must normalize $\vec{u}(x)$ appropriately. We do this by recalling that $1^\top\vec{v}(g,x,0) = 1$, since $\vec{v}(g,x,0)$ is a vector of probabilities. Then we have:

$$\vec{u}(g,x) \to \frac{\vec{u}(g,x)}{1^\top\vec{u}(g,x)}\mu_g p(g) \ .$$

for each g, x.

To calculate prediction metrics—i.e., I_{mem} and I_{fut}—we need $p(x, y)$ and $p(y, \sigma^-)$. The former is a marginalization of $p(\sigma^+, y)$ that we just calculated. The second can be calculated via:

$$p(\sigma^-, y) = \sum_{\sigma^+} p(\sigma^-|\sigma^+) p(y, \sigma^+) \ ,$$

where:

$$\begin{aligned} p(\sigma^-|\sigma^+) &= p((g_-, x_-, \tau_-)|(g_+, x_+, \tau_+)) \\ &= \delta_{x_+, x_-} p(g_-|g_+, x_+) \mu_{g_+} \phi_{g_+}(\tau_+ + \tau_-) \ . \end{aligned}$$

Hence, we turn our attention to calculating dissipation metrics, for which we only need:

$$\frac{\delta p}{\delta t} = \lim_{\Delta t \to 0} \frac{\Pr(X_{t+\Delta t} = x, Y_t = y) - \Pr(X_t = x, Y_t = y)}{\Delta t} \ .$$

Moreover, we can use the Markov chain $Y_t \to \mathcal{S}_t^+ \to X_{t+\Delta t}$ to compute it:

$$\Pr(X_{t+\Delta t} = x, Y_t = y) = \sum_{\sigma^+} \Pr(X_{t+\Delta t} = x | \mathcal{S}_t^+ = \sigma^+) \Pr(Y_t = y, \mathcal{S}_t^+ = \sigma^+) \ .$$

We have:

$$\begin{aligned} \Pr(X_{t+\Delta t} = x | \mathcal{S}_t^+ = \sigma^+) &= \Pr(X_{t+\Delta t} = x | \mathcal{S}_t^+ = (g', x', \tau')) \\ &= \begin{cases} \frac{\Phi_{g'}(\tau' + \Delta t)}{\Phi_{g'}(\tau')} & x = x' \\ \frac{\phi_{g'}(\tau')}{\Phi_{g'}(\tau')} p(x|\epsilon^+(g', x')) \Delta t & x \neq x' \end{cases} \ . \end{aligned}$$

This, combined with $p(\sigma^+, y)$, gives:

$$\begin{aligned} \Pr(X_{t+\Delta t} = x, Y_t = y) = & \sum_{g', x' \neq x} \int d\tau' \ \rho((g', x', \tau'), y) \frac{\phi_{g'}(\tau')}{\Phi_{g'}(\tau')} \Delta t \ p(x|\epsilon^+(g', x')) \\ & + \sum_{g'} \int d\tau' \frac{\Phi_{g'}(\tau' + \Delta t)}{\Phi_{g'}(\tau')} \rho((g', x', \tau'), y) \\ = & \Pr(X_t = x, Y_t = y) + \Delta t \Big(\sum_{g', x' \neq x} \int d\tau' p(x|\epsilon^+(g', x')) \frac{\phi_{g'}(\tau')}{\Phi_{g'}(\tau')} \rho((g', x', \tau'), y) \\ & - \sum_{g'} \int d\tau' \frac{\phi_{g'}(\tau')}{\Phi_{g'}(\tau')} \rho((g', x, \tau'), y) \Big) \ , \end{aligned}$$

correct to $O(\Delta t)$. Recalling that:

$$\begin{aligned} \rho((g', x', \tau'), y) &= \rho(g', x', \tau') p(y|g', x', \tau') \\ &= p(x'|g') \Phi_{g'}(\tau') \left(e^{M(x')\tau'} \vec{u}(g', x') \right)_y \ , \end{aligned}$$

gives:

$$\frac{\delta p}{\delta t} = \lim_{\Delta t \to 0} \frac{\Pr(X_{t+\Delta t} = x, Y_t = y) - \Pr(X_t = x, Y_t = y)}{\Delta t}$$

$$= \sum_{g', x' \neq x} \int d\tau' p(x|\epsilon^+(g', x')) \frac{\phi_{g'}(\tau')}{\Phi_{g'}(\tau')} \rho((g', x', \tau'), y) - \sum_{g'} \int d\tau' \frac{\phi_{g'}(\tau')}{\Phi_{g'}(\tau')} \rho((g', x, \tau'), y) \tag{B10}$$

$$= \sum_{g', x' \neq x} \int d\tau' \; p(x|\epsilon^+(g', x')) p(x'|g') \phi_{g'}(\tau') \left(e^{M(x')\tau'} \vec{u}(g', x') \right)_y$$

$$- \sum_{g'} \int d\tau' \; p(x|g') \phi_{g'}(\tau') \left(e^{M(x)\tau'} \vec{u}(g', x) \right)_y \; . \tag{B11}$$

From this, Eqs. (6) and (9) can be used to calculate $\dot{\mathrm{I}}_{\mathrm{np}}$ and βP.

Appendix C: Specialization to Semi-Markov Input

Up to this point, we wrote expressions for the general case of unifilar hidden semi-Markov environment inputs to the sensor. We now specialize to the semi-Markov input case: the environment's states are directly observed, not hidden. Not surprisingly, a great simplification ensues: hidden states g are the current emitted symbols x. Recall that, in an abuse of notation, $q(x|x')$ is now the probability of observing symbol x after seeing symbol x'.

Hence, forward-time causal states are given by the pair (x, τ). The analog of Eq. (B5) is:

$$\vec{p}(y|x, \tau) = e^{M(x)\tau} \vec{p}(y|x, 0) \; ,$$

and we define vectors:

$$\vec{u}(x) := \mu_x p(x) \vec{p}(y|x, 0) \; .$$

The large vector:

$$\vec{U} := \begin{pmatrix} \vec{u}(x_1) \\ \vdots \\ \vec{u}(x_{|\mathcal{A}|}) \end{pmatrix}$$

is the eigenvector $\mathrm{eig}_1(\mathbf{C})$ of eigenvalue 1 of the matrix:

$$\mathbf{C} = \begin{pmatrix} 0 & q(x_1|x_2) \int_0^\infty \phi_{x_2}(\tau) e^{M(x_2)\tau} d\tau & \dots \\ q(x_2|x_1) \int_0^\infty \phi_{x_1}(\tau) e^{M(x_1)\tau} d\tau & 0 & \dots \\ \vdots & \vdots & \ddots \end{pmatrix},$$

where normalization requires $1^\top \vec{u}(x) = \mu_x p(x)$.

We continue by finding $p(y)$, since from this we obtain $\mathrm{H}[Y]$. We do this via straightforward marginalization:

$$p(y) = \sum_{\sigma^+} \rho(\sigma^+, y) = \sum_{\sigma^+} p(y|\sigma^+) \rho(\sigma^+)$$

$$= \sum_x \int_0^\infty \; p(y|x, \tau) \rho(x, \tau) \; d\tau$$

$$= \sum_x \int_0^\infty \left(e^{M(x)\tau} \vec{v}(x, 0) \right)_y \mu_x p(x) \Phi_x(\tau) d\tau$$

$$= \sum_x \left(\left(\int_0^\infty e^{M(x)\tau} \Phi_x(\tau) d\tau \right) \vec{u}(x) \right)_y \; .$$

This implies that:

$$\vec{p}(y) = \sum_x \left(\int_0^\infty e^{M(x)\tau} \Phi_x(\tau) d\tau \right) \vec{u}(x) \ .$$

From earlier, recall that $\vec{u}(x) := \mu_x p(x) \vec{p}(y|x,0)$.

Next, we aim to find $p(x,y)$, again via marginalization:

$$\begin{aligned} p(x,y) &= \int_0^\infty \rho((x,\tau),y) d\tau \\ &= \int_0^\infty \mu_x p(x) \Phi_x(\tau) p(y|x,\tau) d\tau \\ &= \int_0^\infty \mu_x p(x) \Phi_x(\tau) \left(e^{M(x)\tau} \vec{v}(x,0) \right)_y d\tau \\ &= \left(\left(\int_0^\infty e^{M(x)\tau} \Phi_x(\tau) d\tau \right) \vec{u}(x) \right)_y \ . \end{aligned} \tag{C1}$$

From the joint distribution $p(x,y)$, we easily numerically obtain $\mathrm{I}[X;Y]$, since $|\mathcal{A}| < \infty$ and $|\mathcal{Y}| < \infty$.

For notational ease, we introduced $\mathcal{T}_t$ in this section as the random variable for the time since last symbol, whose realization is τ. Finally, we require $p(y|\sigma^-)$ to calculate $\mathrm{H}[Y|\mathcal{S}^-]$, which we can then combine with $\mathrm{H}[Y]$ to get an estimate for $\mathrm{I}_{\mathrm{fut}}$. We utilize the Markov chain $Y \to \mathcal{S}^+ \to \mathcal{S}^-$, as stated earlier, and so have:

$$\begin{aligned} p(y|\sigma^-) &= \sum_{\sigma^+} \rho(y, \sigma^+ | \sigma^-) \\ &= \sum_{\sigma^+} p(y|\sigma^+, \sigma^-) \rho(\sigma^+|\sigma^-) \\ &= \sum_{\sigma^+} p(y|\sigma^+) \rho(\sigma^+|\sigma^-) \ . \end{aligned}$$

Eq. (B5) gives us $p(y|\sigma^+)$ as:

$$\begin{aligned} p(y|\sigma^+) &= p(y|x_+, \tau_+) \\ &= \left(e^{M(x_+)\tau_+} \vec{v}(x_+, 0) \right)_y \end{aligned}$$

and Eq. (2) gives us $\rho(\sigma^+|\sigma^-)$ after some manipulation:

$$\begin{aligned} \rho(\sigma^+|\sigma^-) &= \rho((x_+,\tau_+)|(x_-,\tau_-)) \\ &= \delta_{x_+,x_-} \frac{\phi_{x_-}(\tau_+ + \tau_-)}{\Phi_{x_-}(\tau_-)} \ . \end{aligned}$$

Combining the two equations gives:

$$\begin{aligned} p(y|x_-,\tau_-) &= \sum_{x_+} \int_0^\infty \delta_{x_+,x_-} \frac{\phi_{x_-}(\tau_+ + \tau_-)}{\Phi_{x_-}(\tau_-)} \left(e^{M(x_+)\tau_+} \vec{v}(x_+,0) \right)_y d\tau_+ \\ &= \frac{1}{\Phi_{x_-}(\tau_-)} \left(\left(\int_0^\infty \phi_{x_-}(\tau_+ + \tau_-) e^{M(x_-)\tau_+} d\tau_+ \right) \vec{v}(x_-,0) \right)_y \ . \end{aligned}$$

From this conditional distribution, we compute $\mathrm{H}[Y|\mathcal{S}^- = \sigma^-]$, and so $\mathrm{H}[Y|\mathcal{S}^-] = \langle \mathrm{H}[Y|\mathcal{S}^- = \sigma^-] \rangle_{\rho(\sigma^-)}$. In more detail, define:

$$D_x(\tau) := \int_0^\infty \phi_x(\tau + s) e^{M(x)s} ds \ ,$$

and we have:

$$\vec{p}(y|x_-,\tau_-) = D_{x_-}(\tau_-)\vec{u}(x_-)/\mu_{x_-}p(x_-)\Phi_{x_-}(\tau_-)\ .$$

This conditional distribution gives:

$$\begin{aligned}
\mathrm{H}[Y|X_- = x_-, \mathcal{T}_- = \tau_-] &= -\sum_y p(y|x_-,\tau_-)\log p(y|x_-,\tau_-) \\
&= -1^\top\left(\frac{D_{x_-}(\tau_-)\vec{u}(x_-)}{\mu_{x_-}p(x_-)\Phi_{x_-}(\tau_-)}\log\left(\frac{D_{x_-}(\tau_-)\vec{u}(x_-)}{\mu_{x_-}p(x_-)\Phi_{x_-}(\tau_-)}\right)\right) \\
&= -\frac{1}{\mu_{x_-}p(x_-)\Phi_{x_-}(\tau_-)}\Big(1^\top\left((D_{x_-}(\tau_-)\vec{u}(x_-))\log(D_{x_-}(\tau_-)\vec{u}(x_-))\right) \\
&\quad - 1^\top(D_{x_-}(\tau_-)\vec{u}(x_-))\log(\mu_{x_-}p(x_-)\Phi_{x_-}(\tau_-))\Big)\ .
\end{aligned}$$

We recognize the factor $\mu_{x_-}p(x_-)\Phi_{x_-}(\tau_-)$ as $\rho(x_-,\tau_-)$ and so we find that:

$$\begin{aligned}
\mathrm{H}[Y|X_-,\mathcal{T}_-] &= \sum_{x_-}\int_0^\infty \rho(x_-,\tau_-)\,\mathrm{H}[Y|X_- = x_-,\mathcal{T}_- = \tau_-]d\tau_- \\
&= -\int_0^\infty\left(\sum_{x_-}1^\top\left((D_{x_-}(\tau_-)\vec{u}(x_-))\log(D_{x_-}(\tau_-)\vec{u}(x_-))\right)\right)d\tau_- \\
&\quad + \int_0^\infty\left(\sum_{x_-}1^\top D_{x_-}(\tau_-)\vec{u}(x_-)\log(\mu_{x_-}p(x_-)\Phi_{x_-}(\tau_-))\right)d\tau_-\ .
\end{aligned}$$

This, combined with earlier formula for $\mathrm{H}[Y]$, gives $\mathrm{I}_{\mathrm{fut}}$.

Finally, we wish to find an expression for the nonpredictive information rate $\dot{\mathrm{I}}_{\mathrm{np}}$. We review the somewhat compact derivation of $\delta p/\delta t$ in the more general case, specialized for semi-Markov input. This requires finding an expression for $\Pr(Y_t = y, X_{t+\Delta t} = x)$ as an expansion in Δt. We start as usual:

$$\Pr(Y_t = y, X_{t+\Delta t} = x) = \sum_{x'}\int_0^\infty \Pr(Y_t = y, X_{t+\Delta t} = x, X_t = x', \mathcal{T}_t = \tau)d\tau$$

and utilize the Markov chain $Y_t \to \mathcal{S}_t^+ \to X_{t+\Delta t}$, giving:

$$\Pr(Y_t = y, X_{t+\Delta t} = x) = \sum_{x'}\int_0^\infty \Pr(Y_t = y|X_t = x', \mathcal{T}_t = \tau)\Pr(X_{t+\Delta t} = x|X_t = x', \mathcal{T}_t = \tau)\rho(x',\tau)d\tau\ . \tag{C2}$$

We have $\Pr(Y_t = y|X_t = x, \mathcal{T}_t = \tau)$ from Eq. (B5). So, we turn our attention to finding $\Pr(X_{t+\Delta t} = x|X_t = x', \mathcal{T}_t = \tau)$. Some thought reveals that:

$$\Pr(X_{t+\Delta t} = x|X_t = x', \mathcal{T}_t = \tau) = \begin{cases} \Delta t q(x|x')\phi_{x'}(\tau)/\Phi_{x'}(\tau) & x \neq x' \\ \Phi_{x'}(\tau+\Delta t)/\Phi_{x'}(\tau) & x = x' \end{cases}\ , \tag{C3}$$

plus corrections of $O(\Delta t^2)$. We substitute Eq. (C3) into Eq. (C2) to get:

$$\begin{aligned}
\Pr(Y_t = y, X_{t+\Delta t} = x) &= \left(\sum_{x'\neq x}\int_0^\infty \Pr(Y_t = y|X_t = x', \mathcal{T}_t = \tau)q(x|x')\frac{\phi_{x'}(\tau)}{\Phi_{x'}(\tau)}\rho(x',\tau)d\tau\right)\Delta t \\
&\quad + \int_0^\infty \Pr(Y_t = y|X_t = x, \mathcal{T}_t = \tau)\frac{\Phi_x(\tau+\Delta t)}{\Phi_x(\tau)}\rho(x,\tau)d\tau\ ,
\end{aligned}$$

plus corrections of $O(\Delta t^2)$. Recalling:

$$\frac{\Phi_x(\tau+\Delta t)}{\Phi_x(\tau)} = 1 - \frac{\phi_x(\tau)}{\Phi_x(\tau)}\Delta t \ ,$$

plus corrections of $O(\Delta t^2)$, we simplify further:

$$\begin{aligned}\Pr(Y_t = y, X_{t+\Delta t} = x) = &\left(\sum_{x'\neq x}\int_0^\infty \Pr(Y_t = y|X_t = x', \mathcal{T}_t = \tau)q(x|x')\frac{\phi_{x'}(\tau)}{\Phi_{x'}(\tau)}\rho(x',\tau)d\tau\right)\Delta t \\ &+ \int_0^\infty \Pr(Y_t = y|X_t = x, \mathcal{T}_t = \tau)\rho(x,\tau)d\tau \\ &- \left(\int_0^\infty \Pr(Y_t = y|X_t = x, \mathcal{T}_t = \tau)\frac{\phi_x(\tau)}{\Phi_x(\tau)}\rho(x,\tau)d\tau\right)\Delta t \ ,\end{aligned}$$

plus $O(\Delta t^2)$ corrections. We notice that:

$$\int_0^\infty \Pr(Y_t = y|X_t = x, \mathcal{T}_t = \tau)\rho(x,\tau)d\tau = \Pr(Y_t = y, X_t = x) \ ,$$

so that:

$$\begin{aligned}\lim_{\Delta t\to 0}\frac{\Pr(Y_t = y, X_{t+\Delta t} = x) - \Pr(Y_t = y, X_t = x)}{\Delta t} = &\sum_{x'\neq x}\int_0^\infty \Pr(Y_t = y|X_t = x', \mathcal{T}_t = \tau)q(x|x')\frac{\phi_{x'}(\tau)}{\Phi_{x'}(\tau)}\rho(x',\tau)d\tau \\ &- \int_0^\infty \Pr(Y_t = y|X_t = x, \mathcal{T}_t = \tau)\frac{\phi_x(\tau)}{\Phi_x(\tau)}\rho(x,\tau)d\tau \ .\end{aligned}$$

Substituting Eqs. (B5) and (1) into the above expressions yields:

$$\sum_{x'\neq x}\int_0^\infty \Pr(Y_t = y|X_t = x', \mathcal{T}_t = \tau)q(x|x')\frac{\phi_{x'}(\tau)}{\Phi_{x'}(\tau)}\rho(x',\tau)d\tau = \sum_{x'} q(x|x')\left(\left(\int_0^\infty \phi_{x'}(\tau)e^{M(x')\tau}d\tau\right)\vec{u}(x')\right)_y$$

and:

$$\int_0^\infty \Pr(Y_t = y|X_t = x, \mathcal{T}_t = \tau)\frac{\phi_x(\tau)}{\Phi_x(\tau)}\rho(x,\tau)d\tau = \left(\left(\int_0^\infty \phi_x(\tau)e^{M(x)\tau}d\tau\right)\vec{u}(x)\right)_y \ ,$$

so that we have:

$$\begin{aligned}&\lim_{\Delta t\to 0}\frac{\Pr(Y_t = y, X_{t+\Delta t} = x) - \Pr(Y_t = y, X_t = x)}{\Delta t} \\ &\qquad = \left(\sum_{x'} q(x|x')\left(\int_0^\infty \phi_{x'}(\tau)e^{M(x')\tau}d\tau\right)\vec{u}(x') - \left(\int_0^\infty \phi_x(\tau)e^{M(x)\tau}d\tau\right)\vec{u}(x)\right)_y \ .\end{aligned}$$

For notational ease, denote the lefthand side as $\delta p(x,y)/\delta t$. The nonpredictive information rate is given by:

$$\begin{aligned}\dot{\mathrm{I}}_{\mathrm{np}} &= \lim_{\Delta t\to 0}\frac{\mathrm{I}[X_t;Y_t] - \mathrm{I}[X_{t+\Delta t};Y_t]}{\Delta t} \\ &= \lim_{\Delta t\to 0}\frac{(\mathrm{H}[X_t] + \mathrm{H}[Y_t] - \mathrm{H}[X_t,Y_t]) - (\mathrm{H}[X_{t+\Delta t}] + \mathrm{H}[Y_t] - \mathrm{H}[X_{t+\Delta t},Y_t])}{\Delta t} \\ &= \lim_{\Delta t\to 0}\frac{\mathrm{H}[X_{t+\Delta t},Y_t] - \mathrm{H}[X_t,Y_t]}{\Delta t} \ ,\end{aligned}$$

where we utilize stationarity to assert $\mathrm{H}[X_t] = \mathrm{H}[X_{t+\Delta t}]$. Then, correct to $O(\Delta t)$, we have:

$$\begin{aligned}
\mathrm{H}[X_{t+\Delta t}, Y_t] &= -\sum_{x,y} \left(p(x,y) + \frac{\delta p(x,y)}{\delta t}\Delta t \right) \log \left(p(x,y) + \frac{\delta p(x,y)}{\delta t}\Delta t \right) \\
&= -\sum_{x,y} p(x,y) \log p(x,y) - \sum_{x,y} p(x,y) \frac{\delta p(x,y)/\delta t}{p(x,y)} \ \Delta t - \sum_{x,y} \frac{\delta p(x,y)}{\delta t} \log p(x,y) \Delta t \\
&= \mathrm{H}[X_t; Y_t] - \sum_{x,y} \frac{\delta p(x,y)}{\delta t} \log p(x,y) \Delta t \ ,
\end{aligned}$$

which implies:

$$\dot{\mathrm{I}}_{\mathrm{np}} = \sum_{x,y} \frac{\delta p(x,y)}{\delta t} \log p(x,y) \ ,$$

with:

$$\frac{\delta p(x,y)}{\delta t} = \left(\sum_{x'} q(x|x') \left(\int_0^\infty \Phi_{x'}(\tau) e^{M(x')\tau} d\tau \right) \vec{u}(x') - \left(\int_0^\infty \phi_x(\tau) e^{M(x)\tau} d\tau \right) \vec{u}(x) \right)_y \tag{C4}$$

and $p(x,y)$ given in Eq. (C1).

[1] C. H. Bennett. The thermodynamics of computation: a review. *Intl. J. Theo. Physics*, 21(12):905–940, 1982.

[2] J. C. Maxwell. *Theory of Heat*. Longmans, Green and Co., London, United Kingdom, ninth edition, 1888.

[3] L. Szilard. On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings. *Z. Phys.*, 53:840–856, 1929.

[4] D. Mandal and C. Jarzynski. Work and information processing in a solvable model of Maxwell's demon. *Proc. Natl. Acad. Sci. USA*, 109(29):11641–11645, 2012.

[5] S. Deffner and C. Jarzynski. Information processing and the second law of thermodynamics: An inclusive, Hamiltonian approach. *Phys. Rev. X*, 3(4):041003, 2013.

[6] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Correlation-powered information engines and the thermodynamics of self-correction. *Phys. Rev. E*, 95(1):012152, 2017.

[7] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Leveraging environmental correlations: The thermodynamics of requisite variety. *J. Stat. Phys.*, 167(6):1555–1585, 2016.

[8] A. Chapman and A. Miyake. How an autonomous quantum Maxwell demon can harness correlated information. *Phys. Rev. E*, 92(6):062125, 2015.

[9] T. McGrath, N. S. Jones, P. R. ten Wolde, and T. E. Ouldridge. Biochemical machines for the interconversion of mutual information and work. *Phys. Rev. Let.*, 118(2):028101, 2017.

[10] J. M. Horowitz, T. Sagawa, and J.M.R. Parrondo. Imitating chemical motors with optimal information motors. *Phys. Rev. Let.*, 111(1):010602, 2013.

[11] S. Toyabe, T. Sagawa, M. Ueda, E. Muneyuki, and M. Sano. Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality. *Nat. Physics*, 6(12):988, 2010.

[12] Here, when analyzing sensory information processing in biological systems, we take care to distinguish intrinsic, functional, and useful computation [74–76]. *Intrinsic computation* refers to how a physical system stores and transforms its historical information. We take *functional computation* as information processing in a physical device that promotes the performance of a larger, encompassing system. Whereas, we take *useful computation* as information processing in a physical device used to achieve an external user's goal. The first is well-suited to analyzing structure in physical processes and determining if they are candidate substrates for any kind of information processing. The second is well-suited for discussing biological sensors, while the third is well-suited for discussing the benefits of contemporary digital computers.

[13] J. M. R. Parrondo, J. M. Horowitz, and T. Sagawa. Thermodynamics of information. *Nat. Physics*, 11(2):131–139, 2015.

[14] C. Aghamohammdi and J. P. Crutchfield. Thermodynamics of random number generation. *Phys. Rev. E*, 95(6):062139, 2017.

[15] M. Hinczewski and D. Thirumalai. Cellular signaling networks function as generalized Wiener-Kolmogorov filters to suppress noise. *Phys. Rev. X*, 4(4):041017, 2014.

[16] S. Still, D. A. Sivak, A. J. Bell, and G. E. Crooks. Thermodynamics of prediction. *Phys. Rev. Lett.*, 109:120604, Sep 2012.

[17] A. C. Barato, D. Hartich, and U. Seifert. Efficiency of cellular information processing. *New J. Physics*, 16(10):103024, 2014.

[18] J. M. Horowitz and M. Esposito. Thermodynamics with continuous information flow. *Phys. Rev. X*, 4:031015, Jul 2014.

[19] S. Goldt and U. Seifert. Stochastic thermodynamics of learning. *Phys. Rev. Lett.*, 118(1):010601, 2017.

[20] A. B. Boyd, D. Mandal, P. M. Riechers, and J. P. Crutchfield. Transient dissipation and structural costs of physical information transduction. *Phys. Rev. Lett.*, 118:220602, 2017.

[21] A. B. Boyd and J. P. Crutchfield. Maxwell demon dynamics: Deterministic chaos, the Szilard map, and the intelligence of thermodynamic systems. *Phys. Rev. Lett.*, 116:190601, 2016.

[22] R. P. N. Rao and D. H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.*, 2(1):79–87, 1999.

[23] S. E. Palmer, O. Marre, M. J. Berry, and W. Bialek. Predictive information in a sensory population. *Proc. Natl. Acad. Sci. USA*, 112(22):6908–6913, 2015.

[24] D. B. Chklovskii and A. A. Koulakov. Maps in the brain: what can we learn from them? *Annu. Rev. Neurosci.*, 27:369–392, 2004.

[25] A. Hasenstaub, S. Otte, E. Callaway, and T. J. Sejnowski. Metabolic cost as a unifying principle governing neuronal biophysics. *Proc. Natl. Acad. Sci. USA*, 107(27):12329–12334, 2010.

[26] We only consider online or real-time computations, so that the oft-considered energy-speed-accuracy tradeoff [77, 78] reduces to an energy-accuracy tradeoff.

[27] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT Press, Cambridge, Massachusetts, 1998.

[28] M. L. Littman, R. S. Sutton, and S. P. Singh. Predictive representations of state. In *NIPS*, volume 14, pages 1555–1561, 2001.

[29] N. Brodu. Reconstruction of ϵ-machines in predictive frameworks and decisional states. *Adv. Complex Sys.*, 14(05):761–794, 2011.

[30] D. Y. Little and F. T. Sommer. Learning and exploration in action-perception loops. *Closing the Loop Around Neural Systems*, page 295, 2014.

[31] N. B. Becker, A. Mugler, and P. R. ten Wolde. Optimal prediction by cellular signaling networks. *Phys. Rev. lett.*, 115(25):258103, 2015.

[32] F. Creutzig and H. Sprekeler. Predictive coding and the slowness principle: An information-theoretic approach. *Neural Comp.*, 20(4):1026–1041, 2008.

[33] F. Creutzig, A. Globerson, and N. Tishby. Past-future information bottleneck in dynamical systems. *Phys. Rev. E*, 79(4):041925, 2009.

[34] W. Bialek, I. Nemenman, and N. Tishby. Predictability, complexity, and learning. *Neural Computation*, 13:2409–2463, 2001.

[35] I. Nemenman, F. Shafee, and W. Bialek. Entropy and inference, revisited. In *Advances in neural information processing systems*, pages 471–478, 2002.

[36] S. E. Marzen and J. P. Crutchfield. Predictive rate-distortion for infinite-order Markov processes. *J. Stat. Phys.*, 163(6):1312–1338, 2016.

[37] S. Marzen and J. P. Crutchfield. Structure and randomness of continuous-time discrete-event processes. *J. Stat. Phys.*, 169(2):303–315, 2017.

[38] P. Mehta and D. J. Schwab. Energetic costs of cellular computation. *Proc. Natl. Acad. Sci. USA*, 109(44):17978–17982, 2012.

[39] C. C. Govern and P. R. ten Wolde. Energy dissipation and noise correlations in biochemical sensing. *Phys. Rev. Let.*, 113(25):258102, 2014.

[40] A. H. Lang, C. K. Fisher, T. Mora, and P. Mehta. Thermodynamics of statistical inference by cells. *Phys. Rev. Let.*, 113(14):148103, 2014.

[41] S. Bo, M. Del Giudice, and A. Celani. Thermodynamic limits to information harvesting by sensory systems. *J. Stat. Mech.: Th. Expt.*, 2015(1):P01014, 2015.

[42] F. Mancini, M. Marsili, and A. M. Walczak. Trade-offs in delayed information transmission in biochemical networks. *J. Stat Phys.*, 162(5):1088–1129, 2016.

[43] S. Ito and T. Sagawa. Maxwell's demon in biochemical signal transduction with feedback loop. *Nature Comm.*, 6, 2015.

[44] P. Sartori, L. Granger, C. F. Lee Fan, and J. M. Horowitz. Thermodynamic costs of information processing in sensory adaptation. *PLoS Comp. Bio.*, 10(12):e1003974, 2014.

[45] R. A. Brittain, N. S. Jones, and T. E. Ouldridge. What we learn from the learning rate. *J. Stat. Mech*, 2017:063502, 2017.

[46] E. M. Izhikevich. *Dynamical systems in neuroscience*. MIT press, Cambridge, Massachusetts, 2007.

[47] C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *J. Stat. Phys.*, 104:817–879, 2001.

[48] R. G. James, C. J. Ellison, and J. P. Crutchfield. Anatomy of a bit: Information in a time series observation. *CHAOS*, 21(3):037109, 2011.

[49] S. Still, J. P. Crutchfield, and C. J. Ellison. Optimal causal inference: Estimating stored information and approximating causal architecture. *CHAOS*, 20(3):037111, 2010.

[50] R. W. Yeung. A new outlook on Shannon's information measures. *IEEE Trans. Info. Th.*, 37(3):466–474, 1991.

[51] H. Jaeger. *Short term memory in echo state networks*, volume 5. GMD-Forschungszentrum Informationstechnik, 2001.

[52] S. E. Marzen. Difference between memory and prediction in linear recurrent networks. *Phys. Rev. E*, 96(3):032308, 2017.

[53] S. Marzen and J. P. Crutchfield. Optimized bacteria are environmental prediction engines. *Phys. Rev. E*, in press, 2018. arXiv:1802.03105.

[54] S. Still. Information-theoretic approach to interactive learning. *EuroPhys. Lett.*, 85:28005, 2009.

[55] N. Tishby and D. Polani. Information theory of decisions and actions. In *Perception-action cycle*, pages 601–636. Springer, 2011.

[56] J. P. Crutchfield, C. J. Ellison, and J. R. Mahoney. Time's barbed arrow: Irreversibility, crypticity, and stored information. *Phys. Rev. Lett.*, 103(9):094101, 2009.

[57] S. G Das, M. Rao, and G. Iyengar. Universal lower bound on the free-energy cost of molecular measurements. *Phys. Rev. E*, 95(6):062410, 2017.

[58] C. L. Vestergaard and M. Génois. Temporal Gillespie algorithm: Fast simulation of contagion processes on time-varying networks. *PLoS Comp. Bio.*, 11(10):e1004579, 2015.

[59] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comp. Phys.*, 22(4):403–434, 1976.

[60] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, 1991.

[61] S. E. Marzen and S. DeDeo. Weak universality in sensory tradeoffs. *Phys. Rev. E*, 94(6):060101, 2016.

[62] S. Marzen, H. G. Garcia, and R. Phillips. Statistical mechanics of Monod-Wyman-Changeux (MWC) models. *J. Mole. Bio.*, 425(9):1433–1460, 2013.

[63] G. Tkačik, A. M. Walczak, and W. Bialek. Optimizing information flow in small genetic networks. *Phys. Rev. E*, 80(3):031920, 2009.

[64] A. M. Walczak, G. Tkačik, and W. Bialek. Optimizing information flow in small genetic networks. II. Feed-forward interactions. *Phys. Rev. E*, 81(4):041905, 2010.

[65] B. M. C. Martins and P. S. Swain. Trade-offs and constraints in allosteric sensing. *PLoS Comput. Bio.*, 7(11):e1002261, 2011.

[66] S. H. Strogatz. *Nonlinear Dynamics and Chaos: with applications to physics, biology, chemistry, and engineering*. Addison-Wesley, Reading, Massachusetts, 1994.

[67] L. R. Rabiner. A tutorial on hidden Markov models and selected applications. *IEEE Proc.*, 77:257, 1989.

[68] D. Pfau, N. Bartlett, and F. Wood. Probabilistic deterministic infinite automata. In *Adv. Neural Info. Proc. Sys.*, pages 1930–1938. MIT Press, 2011.

[69] C. C. Strelioff and J. P. Crutchfield. Bayesian structural inference for hidden processes. *Phys. Rev. E*, 89:042119, 2014.

[70] L. Arnold. *Random dynamical systems*. Springer Science & Business Media, New York, New York, 2013.

[71] J Casas-Vázquez and D Jou. Temperature in non-equilibrium states: a review of open problems and current proposals. *Rep. Prog. Physics*, 66(11):1937, 2003.

[72] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, second edition, 2006.

[73] R. Landauer. Irreversibility and heat generation in the computing process. *IBM J. Res. Develop.*, 5(3):183–191, 1961.

[74] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105–108, 1989.

[75] J. P. Crutchfield. The calculi of emergence: Computation, dynamics, and induction. *Physica D*, 75:11–54, 1994.

[76] J. P. Crutchfield and M. Mitchell. The evolution of emergent computation. *Proc. Natl. Acad. Sci.*, 92:10742–10746, 1995.

[77] G. Lan, P. Sartori, S. Neumann, V. Sourjik, and Y. Tu. The energy-speed-accuracy trade-off in sensory adaptation. *Nat. Physics*, 8(5):422–428, 2012.

[78] S. Lahiri, J. Sohl-Dickstein, and S. Ganguli. A universal tradeoff between power, precision and speed in physical communication. *arXiv:1603.07758*, 2016.