Investigating the Fundamental Limits of Biological Sensing

William Gilpin

Advisors: Professor Curtis Callan
Dr. David Schwab

1 Introduction

The physical action of sensing inevitably requires computation. Just as a computer relies upon input data—keystrokes, mouse gestures, or punchcards—to generate meaningful conclusions about external processes, a living being relies upon sensory input—patterns of light, vibrations of molecules, or temperature fluctuations—to provide constant information about the outside world. An organism perceiving this information must at some level perform data processing and statistical analysis in order to convert the scattered, incomplete information it encounters in its environment into a optimal course of action; otherwise it would be impossible for it to reliably adapt to new conditions.

But unlike a computer, in which a specific transistor or capacitor can be held accountable for each decision and each logical step, the mechanics of computation in living organisms remain impenetrable. For simple organisms like bacteria, sensing is a function of a network of coupled chemical reactions within the cell that conspire to produce certain behaviors, like directional swimming.[2] If the environment changes, the rate constants or other parameters of this balance are perturbed, and so the behavior of the cell changes as a result. This metaphor can, in principle, be extended to arbitrarily complex sets of behaviors in higher organisms, but parsing reaction networks that can have millions of interdependent constituents is a prohibitively difficult task.[19]

This paper will outline and develop an information-theoretic formalism for biological sensing, and then explore the implications of the results in the context of a few simple sensing systems. This methodology uses a statistical approach at its core, bypassing analysis of each deterministic component of a network and instead allowing the problem to be recast in terms of the general relationships between inputs and outputs of a network. As a result, issues specific to the biological implementation can be sidestepped, and general relations can be found.

2 Formalism

2.1 Inference and the Shannon Information

Suppose an organism seeks to measure the value of some external variable, x, that characterizes its environment. For a cell undergoing chemotaxis, x could be the concentration of an external nutrient or toxin; for a pigeon or migrating bird, it could be the local magnitude of the earth's magnetic field.[11] The most important feature of x is that *the organism cannot change its value*; it can only measure it repeatedly, and then react appropriately by adjusting its own internal state.

Based on past measurements or known environmental constraints on x, the organism has an initial expectation for what value it will find upon measurement. This expectation may be reflected in the value of some internal state of the organism that the measurement process physically alters, or it may be reflected in the structure of the sensing mechanism itself.[5] If the environment changes erratically in time, however, or if previous measurements of the values of x have been incomplete, then this initial estimate for the measurement outcome is a probability distribution of the possible x values, $p(x)$, which gives the probability (or probability density if x is continuous) that the organism will obtain x when it performs a measurement. If the organism then performs a measurement and receives a value x_0, it can then refine its expectation from its initial estimate (the *prior* distribution) using Bayes' Theorem

$$p(x|x_0) = p(x)\frac{p(x_0|x)}{p(x_0)} \tag{1}$$

The *posterior* probability that the external parameter has various values x given the information that a measurement yielded the value x_0 is $p(x|x_0)$. This is directly proportional to the the initial distribution of x, adjusted by a factor of $p(x_0|x)/p(x_0)$. This term represents the support that observing a value of x_0 lends to the expectation $p(x)$; if observing x_0 is very rare in $p(x)$, then $p(x_0|x)$ is very small and so the posterior distribution $p(x_0|x)$ will broaden. A sequence of measurements would consist of instantiating some prior, retrieving a value by measurement, and then using the resulting posterior as the prior for the next measurement. If x has a constant value c and minimal noise and fluctuations, then $p(x)$ will be refined continuously until it converges to $p(x) = \delta(x - c)$, indicating that the cell is absolutely certain that the external parameter has value c.

The cell thus gains information about the signal x every time it performs a measurement, with which it then refines $p(x)$ using Bayesian inference. A function, $I(x)$, can be defined that quantifies the amount of information gained by the organism in each step of this process. This quantity must satisfy two conditions[7, 20]

1. If there are n possible values for the variable x, and each of these values is equally likely ($p_i = 1/n$), then $I(x)$ should monotonically increase with increasing n. This follows from Bayes' Theorem (1), because if the prior distribution is uniform, $p(x) = 1/n$, then the posterior distribution $p(x|x_0)$ will decrease independently of the value of the Bayesian multiplier, $p(x_0|x)/p(x_0)$, which is independent of n. As a result, the total information that the cell gains by performing a measurement increases as n increases, since the difference between $p(x)$ and $p(x|x_0)$ increases.

2. I must be additive among independent measurements. Suppose the organism has been sensing a stationary signal long enough that its prior distribution p does not greatly change with each new measurement, such that a measurement x_1 is expected with probability p_1 and a different measurement x_2 is expected with probability p_2. The joint probability that the organism will happen to measure both values in succession is $p_1 p_2$ if the two measurements are independent—the measurement of x_1 does not foretell or preclude x_2.[4] Thus the total information gained in taking the two measurements and updating the prior into the posterior distribution should be $I[p_1] + I[p_2]$. No additional information results from the two measurements happening to occur together, and so the same amount of information must be gained when x_2 is measured regardless of the earlier x_1 measurement.[7]

The only function that satisfies both requirements is the logarithmic function, which maps a series of independent measurements $\prod_i p_i$ to a summation of independent packets of learned information $\sum_i I[p_i]$,

$$I(x) \equiv \ln p(x). \tag{2}$$

But because different values of x in the signal domain will impart different amounts of information—a very rare observation is more informative than one that is completely anticipated—it is useful to remove x dependence by averaging over the ensemble $p(x)$, resulting in an average information, the **entropy**, that is a functional of p alone,

$$H[p] \equiv -\langle \ln p(x) \rangle_{p(x)} \tag{3}$$

As its title suggests, the entropy measures the randomness or *uncertainty* within the distribution p. More uncertain distributions will produce more information, on average, for a given measurement. This makes high entropy distributions those that have the least structure; they exhibit minimal preference for certain values of x, and so almost any observation x_0 affects them equally when they are used as priors for Bayesian inference.

The definition of the bracketed average in (3) depends on whether the range of values for x is discrete or continuous,

$$-\sum_{\{x\}} p(x)\ln p(x) \leftrightarrow -\int_{-\infty}^{\infty} p(x)\ln p(x)dx.$$

The latter case implies that $p(x)$ is a probability density, rather than an absolute probability; the total probability of a given value of x instead would be $p(x)dx$. The only problem with the definition of entropy for a continuous variable is that the probability *density* has units of $1/x$. As a result, the integral used to average the information into the entropy in (3) involves taking the logarithm of a dimensional quantity—an unsavory practice that leads to non-conventional units in the final result for entropy.[7] A straightforward solution involves averaging the information over the ensemble while comparing the probability distribution of x at each point to a reference function, $q(x)$,

$$D_{KL}(p||q) = -\int p(x)\ln\left(\frac{p(x)}{q(x)}\right). \tag{4}$$

The resulting relative entropy measure is the **Kullback-Leibler Divergence** of the two distributions.[15] For an organism measuring a signal, the reference distribution q is the prior prediction, and so the KL Divergence of the posterior p gives the total change in uncertainty in the value of x as a result of measurement.

2.2 Thermodynamic Constraints

The entropy of a distribution is a useful concept because, in equilibrium, systems tend to adopt configurations that maximize the entropy of their distribution of states. This constitutes a restatement of the second law of thermodynamics,[12] and it suggests that an equilibrated x has high uncertainty for an organism implementing sensing—each measurement of x provides a large amount of information on average. Truly equilibrated signals are rare in real-world sensing networks, but the equilibrium state of x provides a useful benchmark value from which concepts like relative entropy and free energy can be defined for non-equilibrium systems.

In order to determine the equilibrium distribution, $p(x)$, the entropy functional can be subjected to a series of generalized constraint equations[18, 26]

$$\langle\phi_j(x)\rangle_{p(x)} = a_j, \tag{5}$$

where the set of $\phi_j(x)$ can be any set of functions dependent on x that have definite expected values a_j under the probability distribution. A common set of constraints is $\phi_j(x) = x^j$, which specifies that the raw moments of the distribution (the mean shifted variance, skewness, etc.) take fixed values a_j. Additionally, because p is a probability distribution, the entropy functional is always subject to the constraint $\langle 1\rangle_{p(x)} = 1$, hereafter denoted by $\phi_0 = a_0 = 1$.

The equilibrium distribution can be found by defining a Lagrangian consisting of the entropy and the constraints on the system, and then maximizing the resulting functional over p,

$$\max \mathcal{L} : \mathcal{L} \equiv -\langle \ln p(x)\rangle_{p(x)} + \sum_j \lambda_j \left(\langle\phi_j(x)\rangle_{p(x)} - a_j\right). \tag{6}$$

This quantity is at a maximum when the net entropy (the first term) is at a maximum, as well as when p satisfies each constraint and so the summation vanishes term-for-term. The set λ_j of Lagrange multipliers essentially serves to weight each constraint; if the maximization were conducted by blindly selecting candidate functions p and computing their relative scores $\mathcal{L}$, the constraint equation with the largest value of λ_j would have the largest effect on determining whether a given candidate p has a maximum score.[16] The Lagrangian is maximized by dropping the constant term $\sum_j \lambda_j a_j$ and then finding values for which the integrand of the ensemble is zero

$$\langle \ln p_{eq}(x) + \sum_j \lambda_j\phi_j(x)\rangle_{p_{eq}(x)} = 0$$

$$\implies \quad p_{eq}(x) = \frac{1}{Z(\lambda)}\exp\left(-\sum_{j=1}\lambda_j\phi_j(x)\right)$$

where the constraint $\langle 1\rangle_{p(x)} = 1$ has been pulled out of the exponential and expressed as a normalization constant, $Z \equiv \exp(\lambda_0) = \int \exp\left(-\sum_{j=1}\lambda_j\phi_j(x)\right)dx$. For a biological system, the most general governing constraint on the external signal x is energy conservation,

$$\langle \mu_j x \rangle_p = k_B T,$$

where μ_j represents the energy required to increase x by one unit. If the signal being measured is the concentration of a chemical, then μ represents the *chemical potential*, an intensive measure of free energy content frequently used in chemistry. k_B is Boltzmann's constant and T is the temperature, and so this equation merely constitutes a generalized definition of temperature for an arbitrary probabilistic system. If this equation is re-written in non-dimensional form and the summation is taken over the j configurations that have identical energy, E, the resulting probability distribution is

$$p_{eq}(x) = \frac{1}{Z} \exp\left(-\frac{E(x)}{k_B T}\right) \tag{7}$$

This is the familiar Boltzmann distribution from statistical mechanics. If the environment has energy readily available in the form of a heat reservoir (energy is available to increase the signal x), then the resulting equilibrium distribution will exponentially favor low values of x because they require less heat absorption to access.[12] If other constraints are placed on the signal, however, the functional form of p_{eq} will change accordingly.[18] The original Lagrangian used in the maximization of entropy, when subjected to the energy conservation constraint, can be expressed as a **free energy** equation that applies to any system with a well-defined temperature due to a heat bath,

$$F = E(x) - k_B T\ H = \mu x \tag{8}$$

The derivatives of this function lead to generalized forces on x and other parameters of the system, which drive the system into configurations in which the free energy is minimized. Equivalently, if the value of x is increased or decreased very slowly, so that the system is never far from its equilibrium configuration, then F represents the maximum work that can be extracted from the system.[12] In a real-world sensing network, however, signals that demonstrate this quasi-static behavior would be of little interest to an organism—if the environment is not changing, then there is no need to sense it.

2.3 Time dependent signals

In the preceding discussion, x is a stationary parameter. Because its value is fixed, the posterior distribution will eventually converge to a delta function centered around the true value and the information yielded by additional measurements will approach zero. In a real biological system, however, the parameter x may itself depend on space and time, making the input signal that the cell seeks to process arbitrarily complex. For example, in vision, a basic problem for an organism is to determine the relative brightness of its surroundings, requiring cells to count photon inputs to the retina over time to determine the mean photon flux to high precision.[6] But this information is meaningless to the organism unless it can be compared to previous values, requiring a vision system to successfully store information about the values of the input signal (photons) at earlier times.

In such a case, additional subtlety is introduced to a sensing network because the external environment may correlate with itself at past times—if the organism's surroundings were bright a few seconds earlier, it is likely that they will still be bright a few seconds in the future. Such correlations allow an organism to compress its stored information by identifying structures within the signal that reduce its entropy, as well as to predict future environments by recording long-term trends in the sequence of measurements. But this introduces additional energetic considerations into the sensing network, such as the cost of storing information and taking additional measurements, that can cause the network to adopt non-equilibrium configurations in order to maximize the information gained by its measurements. This section will explore how such tradeoffs affect the measurement of a time-varying signal.

2.4 Information Storage and Mutual Information

Suppose a time-varying signal x takes on a discrete series of values, X, over the time interval Δt. If x is a continuous process like the gradual change in the concentration of a chemical due to the addition of a reagent, X can be taken as a series of sampled measurements of the continuous transition with sample rate determined by the measurement system. For the sensing organism to fully analyze the incident signal, it must record information about X by mapping each value x onto an internal variable, y. Unlike x, the form of y and its dependence on x is controlled by the organism performing the measurement; y is an *internal variable*, whereas its antecedent is an external one. For example, during chemotaxis in a bacterium, the organism's internal concentration of a specific protein conformation changes based on the amount of the chemical of interest in the environment.[17] The protein concentration serves as a stand-in for the actual external chemical concentration, and so it represents the sum of a cell's knowledge about its environment.

The succession of values that comprise a time series X can be mapped onto a series of values, Y, representing the recorded information available to the organism. If X is random, then the distribution of possible sequences $x_1, x_2, ...$ is characterized by the ensemble distribution $P(X)$, which reduces to a product over the set $p(x_i)$ if X has no temporal correlations. In turn, $P(X)$ gives rise to a coupled probability distribution $P(Y)$ due to the dynamics of the network, making the sensing problem for the organism that of maximizing the similarity between the two distributions. But because x may be changing erratically in time, there may exist an upper limit to the amount of information that the state variable y can represent about the signal. This limit is set by the dynamical parameters of the network, which affect how quickly y responds to changes of various magnitudes in x. If the network is slow to respond, but x tends to fluctuate rapidly, then y effectively represents a time average of x over an extended period. Such a network would be useful for calculating large scale trends in x while filtering out short-timescale fluctuations that may not be relevant to the organism's response. For example, if the organism was a tree sensing an approaching rainstorm so that it could upturn its leaves,[10] it would be energetically inefficient for the tree to initiate leaf inversion every time the local humidity increases—otherwise every passing warm front would activate the response. But if an internal system variable, y, that the tree uses to keep track of humidity has a very slow response rate to changes in the humidity, then the network naturally filters the signal, because y will only increase if the humidity increases over a sustained timescale. Equivalent cases, where the response of y to changes in x would need to be very fast, also occur, such as within a motile bacterium that will die if exposed even briefly to sunlight.[25]

As a result, it is useful for the organism to quantify the shared information between the incident time series, X, and a response series, Y. The natural measure of this is the **mutual information**, which is the KL Divergence (4) between the observed joint distribution of the variables and the joint distribution if the two variables were completely independent,

$$I(X;Y) = \left\langle \ln \frac{P(X,Y)}{P(X)P(Y)} \right\rangle_{P(X,Y)} \tag{9}$$

When X and Y are completely uncorrelated, $P(X,Y) = P(X)P(Y)$ and the mutual information goes to zero.

A related quantity for time-series is the **predictive information**,[8] which is simply the mutual information between a past set of measurements and a future set of measurements. The past history can be denoted as X_{past}, a set of observations over the period $(-T, 0)$ (where 0 denotes the present) and the set of measurements taken over the next T can be denoted as X_{fut}. Thus the predictive information is

$$I_{pred} = \left\langle \ln \frac{P(X_{past}, X_{fut})}{P(X_{past})P(X_{fut})} \right\rangle_{P(X_{past}, X_{fut})} \tag{10}$$

An alternative form of this quantity that underscores its deep relationship to Bayes' Theorem can be found. An anticipated set of measurements, conditioned on X_{past}, can be written as $P(X_{fut}|X_{past})$, and its relationship to the prior probability of a set of measurements can be found using (1) to be

$$P(X_{fut}|X_{past}) = P(X_{fut})\frac{P(X_{past}|X_{fut})}{P(X_{past})}.$$

In the discussion of (1), it was noted that the fraction term is known as the *Bayesian multiplier*; it represents the factor by which the data set X_{past} changes an organism's prediction of future measurements. If $P(X_{past}|X_{fut}) = P(X_{past})$ (the organism's past observations are equally consistent with any set of future measurements), then the organism learns nothing from performing a measurement sequence and the prior distribution matches the posterior distribution. The Bayesian multiplier thus contains the information contained in the observation—the more the multiplier differs from 1, the more information is contained in the measurement. If the multiplier is averaged over all possible posterior predictions then this information can be calculated[8]

$$I_{pred} = \left\langle \ln \left(\frac{P(X_{past}|X_{fut})}{P(X_{past})} \right) \right\rangle_{P(X_{past}|X_{fut})}$$

This expression is entirely equivalent to (10), and the earlier form can be regenerated using the relation $p(x,y) = p(x|y)p(y) = p(y|x)p(x)$

With these new definitions in mind, the sensing problem now becomes a matter of maximizing the mutual information $I(X;Y)$ between a succession X of environmental states and a corresponding sequence Y of internal states. But in addition to accuracy, a living organism must also seek efficiency—it should always seek a maximally concise Y. Otherwise it risks wasting resources and energy by encoding useless information about X, such as minute fluctuations due to noise in the system. A recent approach poses this problem as an optimization,[23]

$$\min \mathcal{L} : \mathcal{L} \equiv I(Y_0;Y) - \lambda I(Y;X) \tag{11}$$

where Y_0 represents a highly-accurate representation of the input signal X, and λ is a Lagrange multiplier that reflects the relative cost of storing additional information over reducing the representation quality. The solution to this functional is a transformation can be expressed as $Y = AX + \zeta$ (where ζ represents some form of noise or error in the encoding), and it represents an optimal mapping that simultaneously compresses Y_0 while preserving the essential information it contains about X.[9] Because the high-information representation, Y_0, is squeezed into a lower-dimensional form Y, this method is known as the **information bottleneck**.

But while the sensing problem can thus be addressed purely in terms of the nature of this mapping and the relevant information-theoretic quantities associated with it, for real biological systems it is instructive to analyze how an organism would actually implement the encoding of information about its environment. Real-world systems must employ elaborate, lossy networks of coupled chemical reactions or interwoven neuron bundles in order to transmit and store information, introducing further considerations like the energetic cost of information processing itself.

3 A Simple Sensing Network

With the relevant quantities now defined, a prototype biological sensing network can be investigated. A basic network in which the formalism of information theory can yield unexpected results is a modification of the simple *binary symmetric channel* from classical information theory.[3, 15] The key simplification is that x, the environment variable, and y, the state variable, are both assumed to be binary. x switches back and forth between state 0 and state 1 with respective rates γ_+ and γ_-, but y switches between 0 and 1 at different rates depending on the current state of x. When x is 0, y undergoes transitions $1 \to 0$ at rate k_+ and reverts back at rate k_-; when x is 1, y transitions $0 \to 1$ at rate k_+ and reverts at rate k_-. The network is illustrated in Figure 1.

As a result of these dynamics, the value of the state variable reflects the value of the environment variable, making the network a primitive model of a basic sensing process in which a threshold value of the external parameter is needed to discern meaningful changes in the environment from small perturbations due to fluctuations. For example, for a bacterium that seeks to sense the concentration of an external chemical, the network in Figure 1 is a primitive model of a cellular receptor for chemotaxis. As the bacterium navigates its environment, it encounters molecules of the chemical of interest with a frequency proportional to the concentration. These molecules bind and occupy a receptor for time intervals related to the inverse of the "on-rate," γ_+, and then then fall off with a constant rate γ_-, reopening the site for other molecules and allowing the sensor to take additional measurements. When the receptor is "on," or occupied, the internal chemical network of the cell has a constant probability per unit time k_+ of registering the change by activating some internal protein,

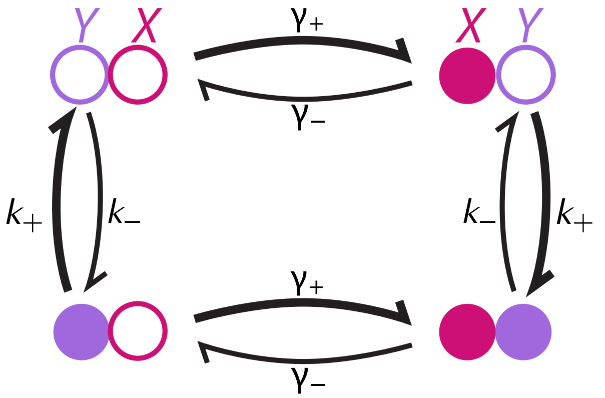

Figure 1: A simple Markovian network with a binary control signal, x, and a binary state variable, y. The rate constants for y are dependent on the value of x, but x is independent of the value of y

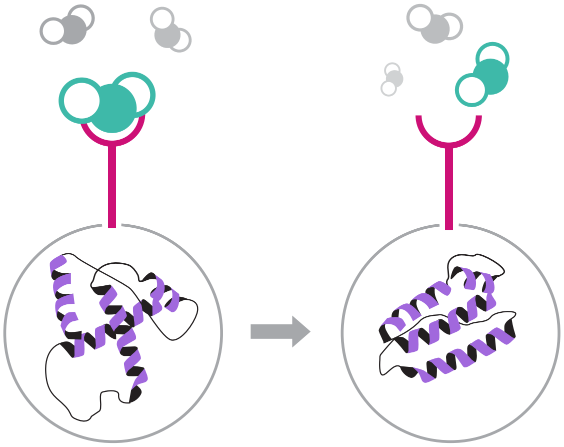

Figure 2: The four-state conditional Markovian network can be implemented in a biological context as a receptor with two states, "bound" and "unbound," and a protein that is activated or deactivated at rates that change based on the state of the receptor.

y, and this acknowledgement has a decay timescale set by $1/k_-$.

For a basic phosphorylation network, k_+ would represent the rate at which phosphate groups are added to an internal protein, activating it by inducing an allosteric rearrangement of the molecule.[1] The activated protein can then be used to transmit the information that the receptor is occupied to other parts of the cell, allowing the bacterium to adjust its behavior in response to the stimulus. In this respect, the value of the off-rate for the activated protein, k_-, sets the timescale for which information is retained by the cell; if the cell only occasionally encounters molecules of the chemical of interest, then even if x and y consecutively turn "on," y will tend to relax back to its "off" state unless the receptor releases its current molecule and and then captures another. A schematic of a simple biological implementation of the network in Figure 1 is given in Figure 2

Discrete network models for chemotaxis and other signaling pathways illustrate how information-theoretic limits can bound biological processes.[17] They also provide concrete systems in which thermodynamic quantities can be explicitly calculated. If the two x values are indexed by α and the two y values are indexed by i, then each node of the network in Figure 1 can be represented by a probability distribution, $p_{\alpha i} \in \{p_{00}, p_{01}, p_{10}, p_{11}\}$, with p_{00} being taken to represent the probability that the system is in the upper left hand corner in the figure. A master equation for probability conservation can thus be written for each of the four state probabilities; for the first state it reads

$$\dot{p}_{00} = k_+ p_{01} + \gamma_- p_{10} - (k_- + \gamma_+) p_{00}.$$

The first two terms represent the increase in probability that the system occupies state 00 during the interval dt due to the chance that it was an adjacent node in the network and transitioned into 00. The last term represents the decrease in probability that the system is in 00 due to the chance that it started at 00 and then transitioned out to an adjacent state during dt. Three similar equations can be written for the other nodes, and so the dynamics can be summarized with a transition matrix

$$\mathbb{W} = \begin{bmatrix} -(\gamma_+ + k_-) & k_+ & \gamma_- & 0 \\ k_- & -(\gamma_+ + k_+) & 0 & \gamma_- \\ \gamma_+ & 0 & -(\gamma_- + k_+) & k_- \\ 0 & \gamma_+ & k_+ & -(\gamma_- + k_-) \end{bmatrix}$$

The system evolves according to $\dot{\mathbf{p}} = \mathbb{W}\mathbf{p}$, where $\mathbf{p}$ is the probability vector $[p_{00}\ p_{01}\ p_{10}\ p_{11}]^T$. In the steady state, the time derivative vanishes and so the solution becomes the nullspace, $\mathbf{p}^{ss}$, of the matrix $\mathbb{W}$,

$$\mathbf{p}^{ss} = \begin{bmatrix} \frac{\gamma_- \left(\gamma_+ \ k_- + k_+ \left(\gamma_- + k_- + k_+\right)\right)}{(\gamma_- + \gamma_+)(k_- + k_+)(\gamma_- + \gamma_+ + k_- + k_+)} \\ \frac{\gamma_- \ k_- \left(\gamma_- + k_- + k_+\right) + \gamma_- \ \gamma_+ \ k_+}{(\gamma_- + \gamma_+)(k_- + k_+)(\gamma_- + \gamma_+ + k_- + k_+)} \\ \frac{\gamma_+ \left(k_-^2 + k_- \left(\gamma_+ + k_+\right) + \gamma_- \ k_+ \right)}{(\gamma_- + \gamma_+)(k_- + k_+)(\gamma_- + \gamma_+ + k_- + k_+)} \\ \frac{\gamma_+ \left(k_- \left(\gamma_- + k_+\right) + k_+ \left(\gamma_+ + k_+\right)\right)}{(\gamma_- + \gamma_+)(k_- + k_+)(\gamma_- + \gamma_+ + k_- + k_+)} \end{bmatrix}$$

In the case in which the internal transition rates are not affected by the external variable, $k_+ = k_-$, this expression simplifies to a distribution that depends only on the x transition rates,

$$\mathbf{p}^{ss} = \frac{1}{2(\gamma_+ + \gamma_-)} [\gamma_- \ \ \gamma_- \ \ \gamma_+ \ \ \gamma_+]^T$$

In this case, $p(y|x)$ is the same for both values of y, and so the internal variable carries no information about the value of x. Thus the network can only generate mutual information between the two variables if y out of equilibrium and detailed balance.

In any case, the relationship between the two variables can be found by first defining the entropy of each variable,

$$H_X = \sum_{\alpha} p_{\alpha y} \ln(p_{\alpha y}),$$

where the "*ss*" label has been suppressed to emphasize that this relation would apply even if the system was not in its steady state. The steady state entropy of the state variable y can be defined analogously, $H_Y = \sum_i p_{xi} \ln(p_{xi})$. This results in the pair of expressions

$$H_X = -\frac{\gamma_- \ln\left(\frac{\gamma_-}{\gamma_- + \gamma_+}\right) + \gamma_+ \ln\left(\frac{\gamma p}{\gamma_- + \gamma_+}\right)}{\gamma_- + \gamma_+}$$

$$\begin{aligned} H_Y = &- \frac{(\gamma_+ k_- + \gamma_- k_+) \ln\left(\frac{\gamma_+ k_- + \gamma_- k_+}{(\gamma_- + \gamma_+)(k_- + k_+)}\right)}{(\gamma_- + \gamma_+)(k_- + k_+)} \\ &- \frac{(\gamma_- k_- + \gamma_+ k_+) \ln\left(\frac{\gamma_- k_- + \gamma_+ k_+}{(\gamma_- + \gamma_+)(k_- + k_+)}\right)}{(\gamma_- + \gamma_+)(k_- + k_+)} \end{aligned}$$

As expected, the entropy of y depends on the rates for both x and y transitions, but the entropy of x is independent of the transition rates of y; the structure of the signal is unaffected by the presence of internal states. In addition to these separate expressions, the entropy can be defined for *all pairs* of values X, Y

$$H_{X,Y} = \sum_{\alpha,i} p_{\alpha y} \ln(p_{\alpha y})$$

The resulting quantity, $\mathbf{p}^{ss} \cdot \ln \mathbf{p}^{ss}$, goes to H_X when $k_+ = k_-$ because y adds no entropy to the system when it is in detailed balance. From these three expressions for entropy, the instantaneous information available about x given a value of the state y can be calculated as the mutual information between the two variables x and y at time t. The logarithmic product in the mutual information formula (9) can be expanded in order to yield the convenient form

$$I(X;Y) = H_X + H_Y - H_{X,Y} \equiv I_{mem}.$$

For standard values of the kinetic parameters, this is a value greater than zero. The quantity I_{mem} is used to denote the mutual information because, in this network, it represents the stored information about the signal X that can be garnered from the steady state distributions determined by x, y. Knowledge of the system's state y can yield at most I_{mem} information about the current value of the signal, x.

But just as the current state of y reveals information about the current state of x, the current state of y can also be used to predict future values of x if the signal has significant correlations in time. Suppose that, for some small time interval Δt, the total transition of the system $x(t), y(t) \rightarrow x(t + \Delta t), y(t + \Delta t)$ can be divided into two half-reactions, one representing the change in x and the other representing the change in y during Δt. The signal change probability can then be expressed as a discrete master equation in x only,

$$p_{00}^{x(t+\Delta t)} = p_{00}(1 - \gamma_+ \Delta t) + p_{10}\gamma_- \Delta t. \qquad (12)$$

The total probability that the system is still in state 00 after Δt is the total probability $1 - \gamma_-$ that no transition occurred during the time interval, added to which is the probability that the system evolved to 00 from 10 along the upper horizontal segment in Figure 1. Identical half-reaction equations can be found for the other three combinations of signal and state value, and so the dynamics of x can be summarized as $\mathbf{p}^{x(t+\Delta t)} = \mathbb{W}^{x(t+\Delta t)}\mathbf{p}$ in a similar manner to the original master equation for the system. Because the evolved distribution $\mathbf{p}^{x(t+\Delta t)}$ represents the probability distribution of x after an infinitesimal time step, it can be used to quantify the predictive capacity of the network: First, the entropy of the evolved distribution can be found in a similar manner to the entropy of the steady state,

$$H'_X = \sum_{\alpha} p_{\alpha y}^{x(t+\Delta t)} \ln\left(p_{\alpha y}^{x(t+\Delta t)}\right).$$

Since the system never drifts far from its steady state probability distributions, each $p_{\alpha y}^{x(t+\Delta t)}$ can be calculated using the steady state values. If similar quantities H'_Y and $H'_{X,Y}$ are defined using the formulae used to calculate H_Y and $H_{X,Y}$ for the steady state, then an instantaneous predictive information (10) can be calculated in a manner directly analogous to the mutual information,

$$I_{pred} = H'_X + H'_Y - H'_{X,Y}.$$

This represents the maximum amount of information that an organism, given the current value of the state y, could deduce about what value the external signal x might take in the next time step. It differs from the general predictive information given in (10) in that it gives the mutual information between the current value of the state and the next value of the signal, whereas the traditional definition of predictive information gives the mutual information between the past and present values of a single variable. This form of the predictive information is of particular interest because it, presumably, gives an estimate of the total predictive capacity of the system; the organism likely can better guess what value x will take in the immediate future than in the far future, and so this form of I_{pred} estimates the overall capability of the network. This is underscored by the specific structure of the network in Figure 1, which is Markovian in x and conditionally Markovian in y, implying that the next step is determined by the current state.

The predictive information has an explicit dependence on the timestep, Δt, by which the transitions of the network are discretized. In the general definition of predictive information (10), the timestep gives the spacing between successive values of the signal X_{t+1}, X_t. For small timesteps (approaching the continuous time limit), I_{pred} can be expanded in the order of Δt, and the linear term can be isolated to yield a predictive information *rate*, denoted by $\hat{I}_{pred}$. It is convenient to relate this quantity to the mutual information by defining a "nostalgia rate" for the network,[22]

$$I_{mem}(t) - \hat{I}_{pred}(t) \tag{13}$$

This represents information about the signal that is contained in the instantaneous state of y but that does not aid in predicting the next value that x will take. If x has the form $x(t) = f(t) + \zeta(t)$, where $\zeta(t)$ is a random force (such as white noise) and $f(t)$ is systematic, meaningful drift in the value of x, then nostalgia quantifies the change in the internal state at time t due to just the non-correlated fluctuations from ζ, which do not predict future values of x. It can be inferred that a real biological system would seek to minimize the value of the nostalgia, since this information is of little use to the organism. A plot of the nostalgia rate for the binary network is shown in Figure 3. It is apparent that the non-predictive information increases when y varies much faster than x; this makes physical sense, because if y is updating its value more often than x, then it must be encoding noise rather than meaningful transitions because x is not changing. Likewise, the undersampling regime where $\gamma_+ \gg k_+$ minimizes the nostalgia because y does not have time to update its value for every x transition, and so it can only encode the slowly-varying components of the signal x—eliminating high-frequency random noise.

3.1 Entropy Production in a Steady State

The organism implementing the sensing network in Figure 1 cannot freely minimize nostalgia to arbitrarily small values however; optimizing the response of the network by changing the kinetic parameters inevitably imposes an energetic cost. This manifests as heat that is produced by the network and then dissipated, introducing a tradeoff between energy efficiency and information quality for the organism implementing the network. Entropy production

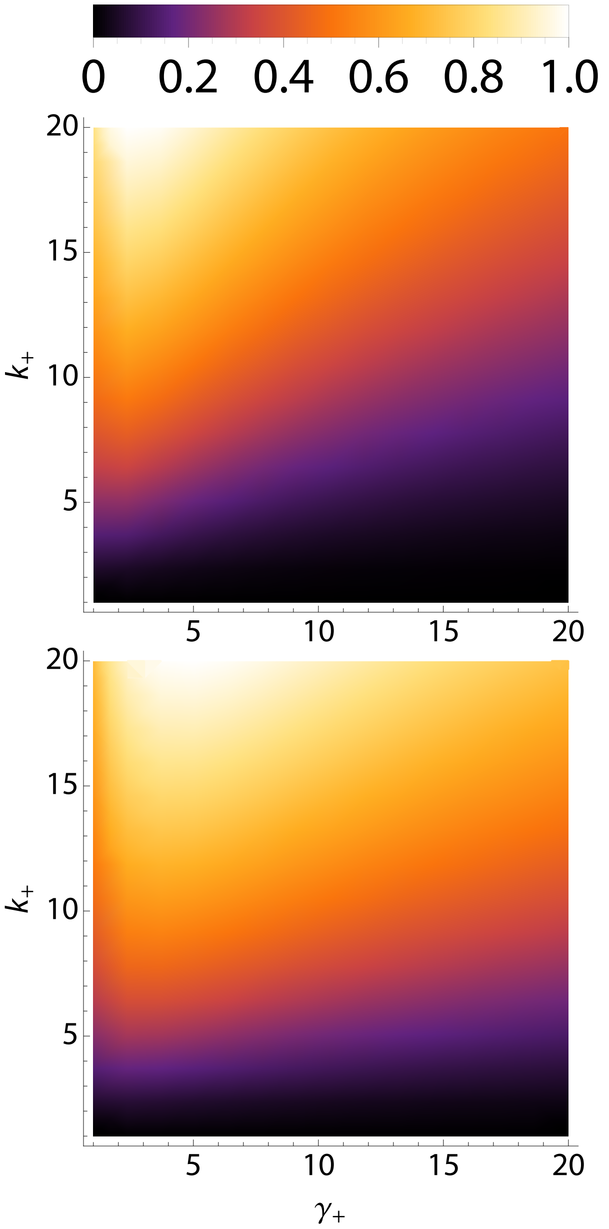

Figure 3: A plot of the normalized nostalgia (top) and entropy (bottom) generated per unit time as a function of the kinetic parameters, where rates are expressed in units of γ_-. As expected, regions of high nostalgia tend to have high energy dissipation.

is a general property of any reaction network containing closed loops,[4] and so it can be readily calculated for a general network and then applied to the specific case of a binary network:

For a general network, the overall entropy can be defined as

$$H[p] = \sum_{\sigma} p(\sigma) \ln p(\sigma),$$

where each σ represents a unique node of the network, such as each of the four pairs x, y in Figure 1. Generally, σ represents one set of unique coordinates for a state in a network of arbitrary dimension. The time derivative of this total entropy must go to zero in the steady state,

$$0 = \sum_{\sigma} \dot{p}(\sigma)(1 + \ln p(\sigma)).$$

The left hand side of this expression can be decomposed into the difference between two separate terms that have physical meaning.[14] This first requires the identification of the general master equation for this network,

$$\dot{p}(\sigma) = \sum_{\sigma'} k_{\sigma' \to \sigma} p(\sigma') - k_{\sigma \to \sigma'} p(\sigma).$$

Inserting the master equation into the time derivative of entropy yields the expression

$$0 = \sum_{\sigma} \sum_{\sigma'} (k_{\sigma' \to \sigma} p(\sigma') - k_{\sigma \to \sigma'} p(\sigma))(1 + \ln p(\sigma)).$$

The master equation vanishes when $k_{\sigma' \to \sigma} p(\sigma') = k_{\sigma \to \sigma'} p(\sigma)$. Inserting this into the equation and grouping terms, the total entropy production in the steady state can be expressed as

$$0 = \sum_{\sigma} \sum_{\sigma'} (k_{\sigma' \to \sigma} p(\sigma') - k_{\sigma \to \sigma'} p(\sigma))(1 + \ln p(\sigma)).$$

$$\begin{aligned} 0 = \sum_{\sigma,\sigma'} k_{\sigma' \to \sigma} \left(p(\sigma') \ln p(\sigma) + p(\sigma) \ln p(\sigma') + p(\sigma') \right) \\ - \sum_{\sigma,\sigma'} k_{\sigma \to \sigma'} \left(p(\sigma) \ln \left(\frac{k_{\sigma \to \sigma'}}{k_{\sigma' \to \sigma}} \right) + p(\sigma) \right) \end{aligned} \tag{14}$$

From the expansion, a term that depends linearly on p can be isolated, allowing it to be expressed as a property of the ensemble,

$$\dot{H}_{diss} = \left\langle \sum_{\sigma'} k_{\sigma\to\sigma'} \ln\left(\frac{k_{\sigma\to\sigma'}}{k_{\sigma'\to\sigma}}\right) \right\rangle_{p(\sigma)} . \tag{15}$$

This quantity corresponds to the energy that the system dissipates when it is not in detailed balance, $k_{\sigma'\to\sigma} = k_{\sigma\to\sigma'}$ for all σ'. In such a case, there is a net flux of material across closed loops in the network (such as the single loop in Figure 1), which results in energy being dissipated in order to switch between steady states.[4] The remaining terms in the expanded entropy equation (14) that were not included in the definition (15) represent the entropy increase within the network due to the transitions among states.

For a system in which the external signal is Markovian, but the internal state is conditionally Markovian (transition rates are governed by the values of x), the entropy production equation (15) must be altered to take into account the conditional internal states

$$\dot{H}_{diss} \left\langle \sum_{\sigma'} k_{\sigma\to\sigma'} \ln\left(\frac{k_{\sigma\to\sigma'}}{k_{\sigma'\to\sigma}}\right) + \sum_{i} k^{\sigma}_{i\to j} \ln\left(\frac{k^{\sigma}_{i\to j}}{k^{\sigma}_{j\to i}}\right) \right\rangle_{p(\sigma)}$$

The internal state transition rates are superscripted by the external variable state, σ, because the set of transition rates between a given pair of allowed internal states depends on the value of the external signal, which modulates and facilitates reactions.

For the conditionally Markovian network in Figure 1, this yields the steady state entropy dissipation rate

$$\dot{H} = p^{ss}_{00}\left(k_- \ln\tfrac{k_-}{k_+} + \gamma_+ \ln\tfrac{\gamma_+}{\gamma_-}\right) + p^{ss}_{01}\left(k_+ \ln\frac{k_+}{k_-} + \gamma_+ \ln\frac{\gamma_+}{\gamma_-}\right)$$
$$+ p^{ss}_{10}\left(k_+ \ln\frac{k_+}{k_-} + \gamma_- \ln\tfrac{\gamma_-}{\gamma_+}\right) + p^{ss}_{00}\left(k_- \ln\tfrac{k_-}{k_+} + \gamma_- \ln\tfrac{\gamma_-}{\gamma_+}\right) \tag{16}$$

At a fixed temperature, this entropy dissipation rate corresponds to heat that permanently leaves the system, $\dot{Q} = k_B T \dot{H}$. This quantity is plotted in the second part of Figure 3; it behaves in a similar manner to the nostalgia when the kinetic parameters are varied, suggesting that additional nostalgia in the system results in additional energy being wasted as heat generation. However, because the two graphs are not identical, there exists a unique location in parameter space, $(k_+, k_-, \gamma_-, \gamma_+)$, such that the nostalgia is minimal for a given heat production. A natural sensing network likely would evolve towards this minimum, since it represents the set of parameters for which the organism can gain the most predictive information about X for each unit of energy it expends.

3.2 Optimizing the Network

The network in Figure 1 can be optimized over its parameter space, $\mathbf{a} \equiv (k_+, k_-, \gamma_-, \gamma_+)$, by fixing the entropy production (16) at a series of values, $\dot{H}_i$, and introducing a set of Lagrangians for each fixed value of the entropy rate. This results in a set of optimization problems,

$$\max \mathcal{L}_i : \mathcal{L}_i \equiv I_{mem}(\mathbf{a}) - \lambda(\dot{H}(\mathbf{a}) - \dot{H}_i). \tag{17}$$

The solutions correspond to sets of parameters that generate the largest mutual information between X and Y for fixed entropy production. This form of the problem is well-suited for numerical analysis, and so the results of a pointwise numerical optimization are shown in in Figure 4, a plot of mutual information versus entropy production per unit timestep. Initially, as the kinetic parameters are increased, the system rapidly gains additional mutual information for each additional unit of heat expenditure. But eventually the network plateaus at a maximal mutual information (corresponding to ln 2, the most information possible about the binary variable x), and so investing additional energy into the network yields diminishing returns. For other, non-optimal sets of kinetic parameters, the location of the curve in the I_{mem} versus $\dot{H}$ plot is bounded below the curve and above the horizontal axis, and its location is insensitive to small changes in parameters.

Unsurprisingly, the highest-information solution for every value of entropy production occurs when $\gamma_+ = \gamma_-$, physically corresponding to the case in which the receptor is occupied half of the time. This corresponds to the maximum-entropy probability distribution $p(x)$, and so a given measurement yields maximal information because the prior prefers no value of x. If the values of γ_+ and γ_- are fixed to not be equal, however, and the optimization is performed over the y rates only, then the maximal information approaches a lower asymptote determined by the logarithm of γ_+/γ_-. This is because, when $\gamma_+/\gamma_- \neq 1$, there is less information in the signal itself. An organism implementing a sensing network could tune the structure and binding potentials of its receptors to ensure that the point $\gamma_+ = \gamma_-$ occurs at the most common values of x, thus allowing it to extract maximal information most of the time.

Because the parameter space is essentially two-dimensional (k_+ and k_-), alternative two-parameter parameterizations exist. An identical optimal curve is generated if the network is instead expressed in terms of the parameters, p and q, representing, respectively, the fraction of the time that y is on when x is on, and the fraction of the time that both are off. The corresponding probabilities $1-p$ and $1-q$ thus represent the error rates of the internal state for each state of the receptor. The equivalence of the parameterization suggests that the details of the measurement and recording process do not affect the maximum information rate of the network—the only figures of merit are the conditional accuracies, $p(y=0|x=0)$ and $p(y=1|x=1)$. As a result, the same relation between I_{mem} and $\dot{H}$ can be constructed without reference to the y transition rates, $k_\pm$, and the identical result is overlaid in Figure 4.

This result suggests that the optimization can be posed alternatively as an information bottleneck. Because $I_{mem} \equiv I(X_t;Y_t)$ and $I_{pred} \equiv I(X_{t+\Delta t};Y_t)$, the optimization (11) can be expressed as

$$\min \mathcal{L}_i : \mathcal{L}_i \equiv I_{mem} - \lambda I_{pred}. \tag{18}$$

The mutual information represents the high-accuracy representation of the input signal, and I_{pred} represents a compression of I_{mem} that stores only features that will correlate with values at later timesteps. This formulation illustrates the relevance of the earlier definition of nostalgia, (3), which represents this Lagrangian in the case $\lambda = 1$, or when decreasing I_{mem} is equally as important to the optimization as increasing I_{pred}. An optimization of this Lagrangian yields an identical plot to that shown in Figure 4. The critical parameter values, p_c and q_c, for which (17) is maximized (or (18) is minimized), are shown in Figure 5 as a function of the nostalgia rate $\dot{N}$ in the network. Because of symmetry, at each $\dot{N}$ there are two possible locations at which the maximum occurs, (p_c, q_c) and $(1-p_c, 1-q_c)$. For low energy budgets both p_c and q_c stay near zero, but p_c approaches 1 as the nostalgia rate (and thus energy used) increases. q_c, however, increases until it reaches a maximum when $p_c = .5$, corresponding to the case where y is wrong just as often as it is right. The overall behavior of the graph suggests that the system prioritizes high certainty about one state of x at the cost of certainty about the other state.

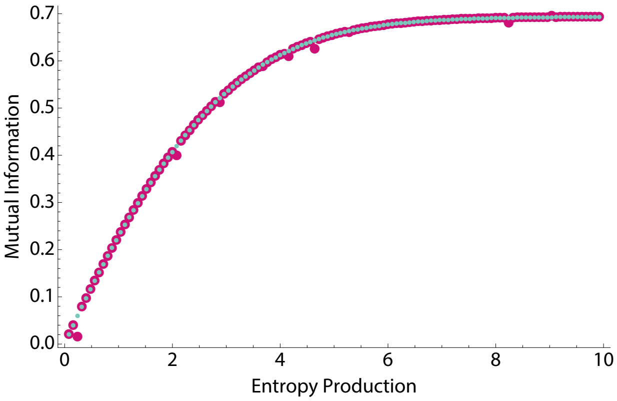

Figure 4: The maximal mutual information rate for various values of the entropy production rate in the network. The optimization was performed numerically using the Nelder-Mead method, and scattered points are due to numerical error. The magenta points represent the optimization over the set of kinetic rates for the network, and the turquoise points represent optimization over two generic error rates.

Together, these results suggest that the entropy production and information gained by the network obey a simple functional relation that accounts for the monotonic, asymptotic relation between the two quantities in Figure 4. It remains unclear, however, what "enforces" the asymptote—as the entropy production of the network increases, so must some other quantity that siphons energy as I_{mem} approaches $\ln 2$. This behavior can be explained by exploring a further subtlety of the network, the free energy change induced by the transitions in x.

3.3 Free Energy Change

Whenever the external parameter x changes value, it provokes a change in y that causes the system to use energy to adapt to the new environment. This energy can be quantified by considering the free energy change provoked by the half-reaction in which y adapts to a change in x.

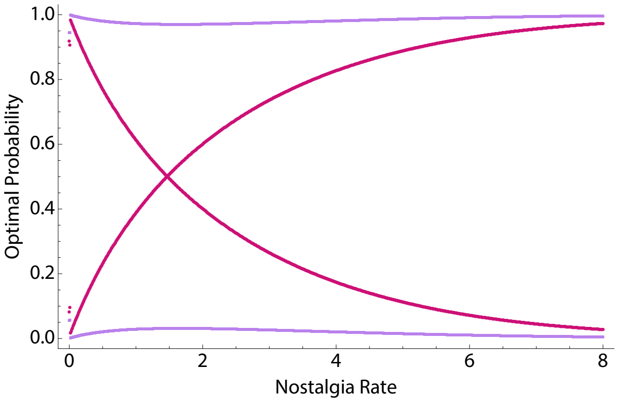

Figure 5: The values of the optimal conditional probabilities, $p_c = p_{11}$ (magenta) and $q_c = p_{00}$ (purple), parametrized by the fixed nostalgia rate in the network. Optimization was performed using the Nelder-Mead method, and the discretization is finer than the resolution of the image.

If the value of the variable x is held fixed at 1 or 0, the variable y will eventually enter an equilibrium steady state that does not require a continued expenditure of energy to maintain. This result can be seen by noting that the network of Figure 1 contains no continuous loops when the value of x is held fixed; y can switch back-and-forth between its two states, but it only has one route to each point on the network. This means that the network is in detailed balance, a strong thermodynamic condition[12] that implies that the system will not produce entropy. The entropy production rate (15) vanishes because $k_{\sigma' \to \sigma} = k_{\sigma \to \sigma'}$ for all σ.

In such an arrangement, which might occur if the concentration of a chemical reaches a threshold high or low enough that the receptor x will always be on or off, the resulting equilibrium distribution of y can be calculated. For the initial state $x = 0, y = 0$, the waiting time (the average time that y waits before transitioning to 1) is proportional to $1/k_-$. Thus after very many time steps the probability that y is in state 0 will be proportional to $1/k_-$. This quantity can then be normalized by dividing by the total time that it takes y to return to 0. Applying this method to each node of the binary network yields the equilibrium set

$$\mathbf{p}^{eq} = \begin{bmatrix} \frac{k_+}{k_- + k_+} & \frac{k_-}{k_- + k_+} & \frac{k_-}{k_- + k_+} & \frac{k_+}{k_- + k_+} \end{bmatrix}^T$$

In addition to these equilibrium distributions, the conditional probability, $p(y|x)$, of each value can be calculated using the relation $p(y,x) = p(y|x)p(x) = p(x|y)p(y)$ and the nonequilibrium steady state joint probability distributions. This yields a set of probability distributions for y conditioned on a known value of x,

$$\mathbf{p}^c = \begin{bmatrix} \frac{p_{00}}{p_{00} + p_{01}} & \frac{p_{01}}{p_{00} + p_{01}} & \frac{p_{10}}{p_{10} + p_{11}} & \frac{p_{11}}{p_{10} + p_{11}} \end{bmatrix}^T$$

Unlike the members of the joint distribution, the elements of the vector $\mathbf{p}^c$ are arranged with the first index representing y, such that the first element represents $p_{0|0}$, the second $p_{1|0}$ and so forth.

Finally, for the half-reaction in which the value of x changes but the value of y does not evolve, an analogous set of conditional distributions can be defined based on the half-reaction steady states (12), $\mathbf{p}^{c,x(t+\Delta t)}$. These have the same form as $\mathbf{p}^c$, but use the half-reaction steady states instead of the true steady states.

For reasons that will be proven generally in the next section, a free energy function can be defined in terms of $\mathbf{p}_{eq}$, $\mathbf{p}^c$, and $p^{c,x(t+\Delta t)}$ for the network that takes the form

$$F = k_B T \sum_{x,y} p_x p^c_{y|x} \ln \left(\frac{p^c_{y|x}}{p^{eq}_{x,y}} \right),$$

where Boltzmann's constant and the temperature convert the quantity into a true energy, as occurs with entropy/heat production. Qualitatively, this function behaves as free energy for the system should behave. In a network with no signal discrimination, $k_+ = k_-$, this quantity goes to zero. This corresponds to the origin in Figure 4, at which no information is stored about the signal and no entropy is generated. Energy conservation thus dictates that the free energy function must also vanish. In the other limit $k_+ \gg k_-$, x and y will always have the same value, and so the sequence Y will perfectly record the signal X. This corresponds to a case very far out along the horizontal axis in Figure 4, where the mutual information has attained the maximum ln 2, at the cost of high heat production. In this regime, the conditions $p_{00}, p_{11} \rightarrow, p_{01}, p_{10} \rightarrow 0$ cause F to diverge—a sensible result, since F serves the function (described in the previous section) of compensating for the difference between the stored information and the entropy generated, which diverges when the entropy goes to infinity and I_{mem} approaches a finite constant. F thus represents how far *out of equilibrium* the system lies when it is in its steady state; the form of F is similar to the KL Divergence (4) between the conditional distribution and the equilibrium, averaged over x.

4 Generalizations

As the information bottleneck assumed and the binary network illustrated, any organism must at some point make a tradeoff between accuracy and efficiency when measuring a time-varying signal. Thus there must be a general relationship between information, heat production, and free energy, that applies to more complicated sensing systems. This section will present a simplified version of a recent derivation by Still et al.[22] that provides a relationship among I_{mem}, H, and F for a general class of conditionally Markovian networks. While this principle may not scale to networks with more complicated temporal correlations between the two sequences X and Y, it nonetheless illustrates the importance of these quantities in characterizing a biological sensing network.

The approach taken by Still et al. simplifies the complex, interrelated kinetics of two interrelated variables by first discretizing time and then dividing the response of the system into a series of work phases and relaxation phases, as illustrated in Figure 6. If a system starts out in equilibrium and with a well-defined Hamiltonian, work must be applied in order to drive the change $x_0 \rightarrow x_1$. For the binary network in Figure 1, this represents the energy difference between the two states of the receptor; more generally, if x has a chemical potential μ, this work represents the extensive free energy change, $W = \mu\Delta x$. For a general Hamiltonian, x could be another parameter of interest, such as the stiffness of a spring constant or the pressure of a gas. Once x has completed its transformation, the system then has some time to adjust its coupled state variable, y, during the relaxation step. Before y has time to completely equilibrate, however, the next change in x occurs, once again resetting y to a position far from equilibrium. In this manner, the state variable can be driven to different values depending on the size and frequency of jumps in x.

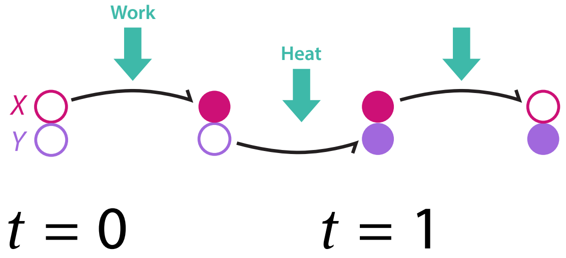

Figure 6: The first two half-steps of the general network mechanism proposed by Still et al.[22] The dynamics of the system are separated into two phases; the first represents work done to change the external variable, and the second represents heat absorbed when the internal workings of the system respond. The timestep size is arbitrary, and the system can be driven arbitrarily far from equilibrium using repeated iterations of this sequence.

The jumps in x inevitably cause changes in the free energy of the system, which result in energy being dissipated into the environment. The relaxation half-step of each section of the work protocol is characterized by the distribution $p(y|x)$, the probability that the system enters configuration y given that the environment is now in x. The free energy change invoked when the system adopts a given probability distribution $p(y|x)$ can be found using the definition of free energy (8) and the first law of thermodynamics to be

$$F_t^{neq} = \langle E(y_t, x_t)\rangle_{p(y_t|x_t)} + k_B T\langle \ln(p(y_t|x_t))\rangle_{p(y_t|x_t)}$$

This expression is identical to the standard relationship between free energy, entropy, and internal energy, but it has been explicitly denoted F_t^{neq} to clarify that it applies to the system with distribution $p(y|x)$at time t, which is generally not in equilibrium after a step in the work protocol takes place. If this quantity is averaged over all possible values that x could have taken, $p(x)$, then

$$\langle F_t^{neq}\rangle_{p(x)} = \langle E(y_t, x_t)\rangle_{p(x_t,y_t)} - k_B T\ H[p(y_t|x_t)] \tag{19}$$

This represents the expected free energy when a given probability distribution $p(x)$ governs the dynamics of the work protocol. The average energy in the joint configuration x, y, less the entropy, represents the non-equilibrium free energy of the arrangement. In the first half-reaction in Figure 6, work is done on the system which increases its internal energy

$$W^{y}_{t\to t+1} = E(y_{t+1}, x_t) - E(y_t, x_t). \tag{20}$$

The upper index indicates that y does not change during the half-step. Of this added work on the system, a portion is dissipated into the heat bath, and the remainder is invested in changing the non-equilibrium free energy during the relaxation half-step during which y updates. The average dissipated work in a single step (both half-steps) can be found by averaging the difference over all possible sequences of states,

$$\langle W^{diss}_{t\to t+1}\rangle_{P_{X,Y}} = \langle W^{y}_{t\to t+1}\rangle - \langle F^{neq}_{t+1} - F^{neq}_{t}\rangle$$

This equation can now be combined with (19) and (20) to yield the relation

$$\langle W^{diss}_{t\to t+1}\rangle_{P_{X,Y}} = k_B T\left(H[p(y_{t+1}|x_{t+1})] - H[p(y_t|x_t)]\right)$$

Using the previous definitions of mutual information, (9), and predictive information (10), this simplifies to the meaningful form

$$\langle W^{diss}_{x_\tau\to x_{\tau+1}}\rangle = k_B T(I_{mem} - I_{pred}), \tag{21}$$

where $W^{diss}_{x_\tau\to x_{\tau+1}}$ represents the total energy lost as heat when x completes one step of its work protocol. This result establishes in general the earlier claim that nostalgia in a sensing network is equivalent to inefficiency. In the binary network, the amount of work done on the system that was wasted as heat scales with the amount of nostalgia in the system, suggesting that a organism employing a sensing network might always seek to avoid nostalgia in order to minimize its energy expenditure.

But while this derivation gives the relationship between dissipation and information for a single complete step $\tau \to \tau + 1$ in the network, establishing the relationship for the entire sequence of trajectories that leads to the final configuration $p(y_\tau|x_\tau)$ requires additional finesse. The entire sequence of driving forces and incomplete relaxations affects the final state, y_τ. In general, this final state will be out of equilibrium; the system will have a distribution of possible states $p(y_\tau|x_\tau)$ that fails to maximize the entropy and minimize the energy of the Lagrangian of the form in (6). As a result, the free energy is not at the minimum value that it can possibly attain; it instead has an excess term corresponding to the energy it would dissipate into the environment if it was given a chance to do so. This term is proportional to the entropy difference between the final distribution $p(y_\tau|x_\tau)$ and the true minimum $p_eq(y_\tau|x_\tau)$, making the magnitude of the excess free energy simply the KL Divergence, (4), between the distributions

$$F^{add}_{\tau}[p(y_\tau|x_\tau)] = k_B T\left\langle \ln\left(\frac{p(y_\tau|x_\tau)}{p_{eq}(y_\tau|x_\tau)}\right)\right\rangle_{p(y_\tau|x_\tau)} \tag{22}$$

This quantity is a functional of the conditional distribution of the internal state variable, and the equilibrium distribution p_{eq} can be any function that minimizes (6) subject to an arbitrary set of constraint equations. But because an arbitrary point in the trajectory in Figure 6 could have been the endpoint of the sequence, each y_τ has an associated F^{add}_{τ}. This makes the quantity of interest in a sensing network the *relative amount* by which two subsequent states differ from equilibrium,

$$\Delta F^{add}_{y_{\tau-1}\to y_\tau} = F^{add}_{\tau} - F^{add}_{\tau-1}$$

This quantity aggregates as each step in the work protocol is performed, and so if it is totaled for the entire sequence and then averaged over all possible trajectories that the system could have taken,

$$\langle \Delta F^{add}\rangle = \left\langle \sum_{t=0}^{\tau-1} \Delta F^{add}_{y_{\tau-1}\to y_\tau}\right\rangle_{P(Y|X)}$$

This represents the average (taken over all possible sequences that the system variable could have taken) amount by which the free energy will differ from equilibrium at the completion of the work protocol $x_1, x_2, ..., x_\tau$, provided that the system was in equilibrium before the sequence began. This term can now be attached to the previous relation for a single timestep to yield a governing equation for the entire process,

$$\langle W^{diss}\rangle = \tau(I_{mem} - I_{pred}) - \langle \Delta F^{add}\rangle. \tag{23}$$

All averages are taken over the joint probability density, $P(X,Y)$. This expression accounts for the loose proportionality between mutual information and entropy production seen in the network of Figure 1; the nostalgia, or the part of the mutual information that captures long-term behavior in X, is proportional to be the relaxation free energy of the system and the heat dissipated. Thus optimization of the sensing process constitutes maximizing the predictive information while minimizing the overall mutual information, since the latter introduces wasted energy in the form of heat production, leading to limitating behavior such as that seen in Figure 3.

4.1 Landauer's principle

The relation (23) found by Still et. al. is important because it acts as a refinement of Landauer's principle, an early result from information theory that links the computational capacity of a system to the energy available to it.[13]

In the network described in Figure 1, the environmental signal has Markovian dynamics—the value of x at one time point is strongly correlated only with its value at the previous timepoint. If the sequence of values X did not even have this structure, and instead each value x was completely independent from its predecessors, then the signal x would essentially be a random number generator with replacement, and predictive information would be impossible to generate. In this scenario, the dynamics of the x transitions cannot be conveniently summarized with a master equation. However, the distribution of the equilibrated internal state can be determined, since this depends only on the rate constants for y transitions and general manner in which the state of x affects their values. As before, the complete equilibrium probability density is

$$\mathbf{p}_{eq} = \left[\frac{k_+}{k_- + k_+} \quad \frac{k_-}{k_- + k_+} \quad \frac{k_-}{k_- + k_+} \quad \frac{k_+}{k_- + k_+} \right]^T,$$

where the same indexing convention has been used as in previous sections. Now, if the system initially begins in its equilibrium configuration, the entropy of the internal state can be found to be

$$H[Y] = -\frac{k_- \ln(k_+) + k_+ \ln(k_-)}{k_- + k_+}$$

The maximum of this function occurs when $k_+ = k_- = .5$; in this limit, observing the state of y would provide no information about the state of x because y has a 50% chance of being on or off independently of the value of x. The maximal entropy is thus $\ln 2$. The opposite limit is the case when $k_+ = 1, k_- = 0$, and so the state variable always copies the value of x. In this case, the entropy goes to zero, since there is no uncertainty in the value of y.

These limits together set the necessary entropy change for a perfect sensing network that alternates between two configurations: an uncertain state in which the internal variable is evenly dividing among all of its possible values, and a state in which y has a single value determined by x. The entropy generated in switching y from a pure state to a maximal entropy state is $\ln 2$, which can be expressed as the heat that the network generates in changing its configuration, $k_B T \ln 2$

This result is a specific example of **Landauer's principle**,[13] which states that, whenever a single bit of information is erased from a system, the total heat dissipated into the environment must be at least $k_B T \ln 2$. In the ideal non-Markovian binary network described above, the transition from the recorded value $y = 1$ to the indecisive configuration $k_+ = k_-$ constitutes the erasure of a single bit of information; the network evolves from y containing little information about x to y containing full information about the environment.

Landauer's principle thus dictates that the energetic costs of a sensing network originate in its mechanism for erasing the state of its internal variable, y. A perfectly adiabatic sensing network can only be realized if the network has impossibly infinite memory, making the principle a necessary relation between the energy burned by a sensing network and its information transmission capacity. In a network less ideal than the one described above, in which the transitions in rate constants provoked by a change in x are far more subtle, the system will burn less energy per sensing event, but it will also carry less information about x.

The relationship between Landauer's principle and the relation of Still et al. arises from relaxing the assumption above that each x be entirely independent of all previous values. If, instead, x is once again allowed to be Marko-

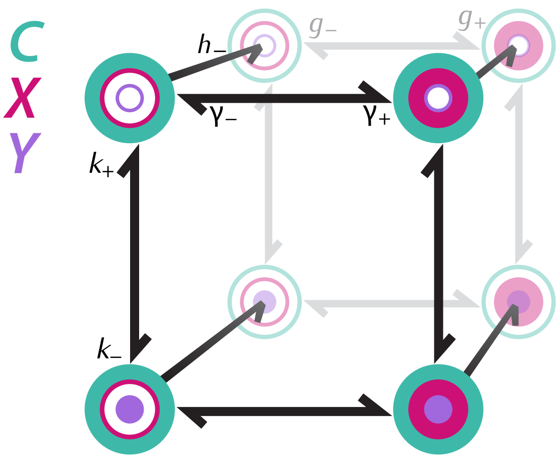

Figure 7: An extension of the binary network where the transition rates for x are themselves controlled by a binary variable, c. The front plane of the network is identical to the network in Figure 1, but the back face has modified x transition rates, $\gamma_\pm \to g_\pm$

vian, then Landauer's principle will no longer apply to the network because the sequence Y can portend future values of X. In this case, the non-equilibirum free-energy emerges as an additionally change in the overall energy of the state, resulting in an extra term in the relationship between the energetic cost and information gain of the network.

5 Extensions

Landauer's principle and the relation of Still et al. suggest that there is an upper limit to the information that a sensing network can gain for each unit of energy it expends, but it is unclear whether the previously-introduced binary network attains this limit. A useful test is whether a more sophisticated network can achieve a steeper curve than that shown in Figure 4:

It is apparent that a more accurate rendering of the biological sensor shown in Figure 2 must introduce a third variable, c, representing a time-varying concentration of the external species that intermittently occupies the receptor x. For simplicity, c may itself be treated as a binary variable. Such a scenario might occur when an organism is interested only in whether the environment variable is above or below a certain critical threshold, such as an *E. coli* cell that is trying to avoid a toxic repellent.[24] Alternatively, such thresholding occurs in vision networks, in which a certain, critical number of photons simultaneously incident on the retina are needed to trigger registration of light.[6]

A network implementing an additional binary concentration variable is shown in Figure 7. The additional variable increases the dimension of the network's state space, resulting in a cubical layout with eight allowed joint configurations (c, x, y). Both the front and back faces of the cube represent implementations of the original binary network, but now at each of the four points there is an additional option of moving to the opposite face at a rate h_+ or h_-. The only difference between the front and back faces is the rate of x transitions: when $c = 0$, x has rate constants γ_-, γ_+, and when $c = 1$, x has rate constants g_- and g_+. Physically, this models a network in which the presence of a high concentration of attractant ($c = 1$) catalyzes the occupation process of the receptor, resulting in faster transitions, $g_+ > \gamma_+$. As a result, the relevant information theoretic-quantity becomes $I(C;Y)$, and x is reduced to the role of an intermediate between the external drive and internal memory.

As before, the dynamics of the network can be summarized by an 8×8 transition matrix, $\mathbb{W}$. The associated steady state and entropies of the system have prohibitive analytic solutions, and so instead the problem can be addressed numerically by choosing random values for each of the eight rate constants and then calculating the value of I_{mem} and $\dot{H}$ for the resulting, specific transition matrix. If this procedure is applied for sufficiently many iterations, the resulting data can be reduced by taking the maximum value of I_{mem} found for a given entropy production rate. Figure 8 shows the result of performing this process on a simulation of 100 million points, where during each iteration each rate constant was randomly selected from the range $[1 \times 10^{-8}, 1 \times 10^{8}]$.

The figure shows the same asymptotic, concave behavior as the four-state network, but it also clearly exhibits a much lower maximal information gain than the original, four-state network. This suggests that the additional heat invested in running the network (which increases due to the higher number of closed loops of states through which the system can pass) does not allow the network to generate additional mutual information. Instead, the additional processing step introduced by the intermediate x actually costs the network information, leading to a lower information yield at high energy expenditure.

This result may seem to follow from a basic theorem in information theory known as the **data processing inequality**.[15] This principle states the intuitive notion that the mutual information between two variables cannot be increased through local processing operations alone—the maximum information that can be extracted about a signal is just some representation of the signal itself, and information can only be lost in any intermediate steps between the observation and encoding of data. If $I(X;Y)$ is the mutual information for the four-state network, and $I(C;Y)$ is the mutual information for the eight-state network, then the data processing inequality suggests that $I(X;Y) \geq I(C;Y)$. However, the assumptions behind this theorem subtly break down for the cube network because the state Y_{t+1} is conditioned on not only the value of X_t, but also on Y_t, and so the network contains more predictive information than a standard Markov chain $C \to X \to Y$. This suggests that there may exist regimes where the mutual information $I(C;Y)$ can be made to transcend the limiting curve in Figure 4, although this behavior is not observed within the 16 orders of magnitude that were searched to generate Figure 8.

In order to examine whether this network can be further adapted to generate better results, set of unique on and off rates can be defined for each edge in Figure 7. The c transition rates are kept as $h_{\pm}$, but for the front face separate transition rates $\gamma_{\pm} \to \gamma_{\pm 1}, \gamma_{\pm 2}$ are defined for the top and bottom edges of the graph. A similar set of rates is used for the g rates on the back face, and four sets of rates $k_{\pm i}$ are defined for the four locations in which the y transitions appear. This results in an 18-dimensional parameter space (eight from y, eight from x, two from c), over which the mutual information between the concentration and y can be optimized.

The results of a random simulation for 16 parameters (scattered over 14 orders of magnitude) are shown in Figure 9. As with $\gamma_{\pm}$ from the four state-network, the rates $h_{\pm}$ are beyond the organism's control, and so changing their ratio only affects the location of the upper asymptote, with $h_{+}/h_{-} = 1$ corresponding to a signal bearing maximal information. Thus $h_{-} = h_{+} = 1$ for the simulation. The remaining parameters were allowed to vary over many orders of magnitude, but the optimal curve is nearly identical to that for the four-state network in Figure 4.

The results indicated in the Figure suggest that the relationship (23) found for conditionally Markovian networks by Still et al. is a specific case of an even more general governing relationship for sensing networks. The network in Figure 7 does not obey the relation of Still et al. due to the presence of closed loops of sequential states when c is fixed, which use energy even if the environment is not changing—this additional energy loss is the **housekeeping heat**, in contrast to the **excess heat** generated in switching between two different steady states, which was found in (15).[21] The fact that the cube network is substantially different from a pure Markovian network due to the presence of this extra quantity, yet exhibits the same maximal mutual information rate per unit of heat production, suggests that the free energy term used by Still et al. has an analogue in more complex networks that causes them to behave the same way as the simple binary network. Additional investigation must be done in order to determine whether the parameter search used to generate Figure 8 was suitably extensive, as well as whether natural sensing systems even implement networks that obey dynamics as complicated as the 18-rate system. But these early results suggest that the tradeoff between heat production and information capacity found in basic networks may be a meaningful constraint in natural sensing systems.

6 Next Steps

If the internal variable was allowed more possible states, or if the topology of the network was extended in such a way as to allow the binary y to store its values from previous timesteps, then a conditional Markovian network could be generated that captures predictive information about x with even greater accuracy. Such developments would be well-suited to scenarios in which the input sequence X itself has more complicated structure, such as long-range temporal correlations or additional allowed values. De-

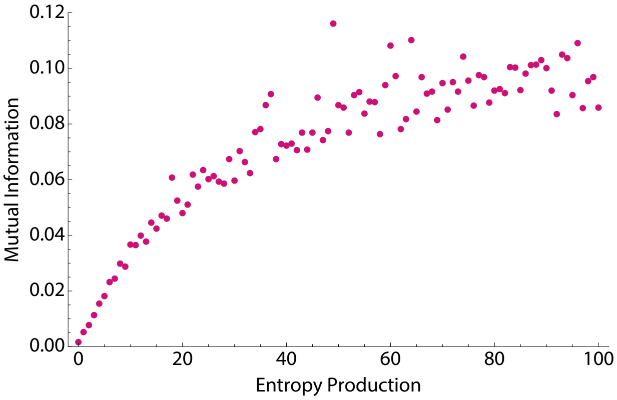

Figure 8: The optimal mutual information rate versus entropy production rate for the eight-state network. The parameters of the network were each randomly scattered over ten orders of magnitude, and for each set of random parameters the entropy production and mutual information were calculated. At the conclusion of the simulation, the maximum mutual information for each entropy production was recorded. Vertical scatter is likely due to insufficient sampling.

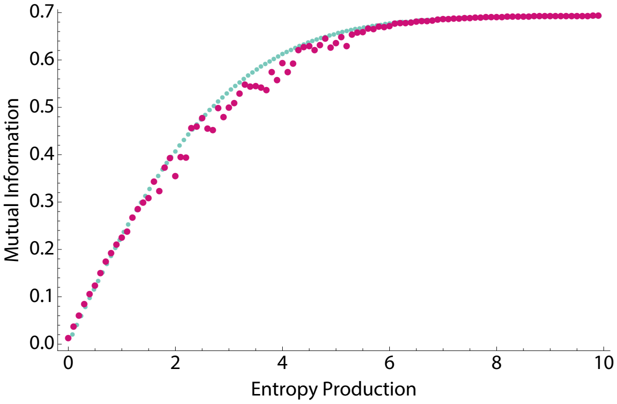

Figure 9: The optimal mutual information rate versus entropy production rate for the for the cube network in which all parameters are free (magenta). The optimal curve for the four state network is underlaid in turquoise. Despite its additional degrees of freedom, the cube network does not offer greater information gain than the simpler network.

pending on the arrangement of the network, in most cases some form of Landauer's principle or the relation of Still et al., would be expected to apply, as additional measurements inevitably would impose additional energetic costs.

The recurrence of such basic thermodynamic constraints in different network structures is unsurprising, but the fact that general Markov networks can be constructed and analyzed without explicit reference to the type of sensing process suggests that these fundamental constraints guide the formation of sensing networks in the first place. With genetic code to define the problem, and natural selection to optimize parameters, organisms have developed natural networks capable of efficiently and accurately measuring environmental changes ranging from single molecules of toxins to single photons of light.[6] While the mechanisms (coupled chemical reactions, electrical impulses) behind such networks differ widely, all must fundamentally answer to the same basic conservation laws—making the information-theoretic approach an important means of consolidating natural diversity.

The author thanks Professor Callan for formulating the concept behind the paper, as well as for providing extensive advice on the structure and approach. He also thanks Dr. Schwab for originating many of the approaches taken, as well as for elucidating many of the mathematical concepts presented.

References

[1] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell.* Garland Science, 1997.

[2] Uri Alon, Michael G Surette, Naama Barkai, and Stanislas Leibler. Robustness in bacterial chemotaxis. *Nature*, 397(6715):168–171, 1999.

[3] AC Barato and U Seifert. Information-theoretic vs. thermodynamic entropy production in autonomous sensory networks. *arXiv preprint arXiv:1212.3186*, 2012.

[4] Daniel A Beard and Hong Qian. *Chemical biophysics: quantitative analysis of cellular systems.* Cambridge University Press, 2008.

[5] Howard C Berg and Edward M Purcell. Physics of chemoreception. *Biophysical journal*, 20(2):193–219, 1977.

[6] William Bialek. Physical limits to sensation and perception. *Annual review of biophysics and biophysical chemistry*, 16(1):455–478, 1987.

[7] William Bialek. *Biophysics: searching for principles.* Princeton University, 2010.

[8] William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, complexity, and learning. *Neural Computation*, 13(11):2409–2463, 2001.

[9] Felix Creutzig and Henning Sprekeler. Predictive coding and the slowness principle: An information-theoretic approach. *Neural computation*, 20(4):1026–1041, 2008.

[10] DA Grantz. Plant response to atmospheric humidity. *Plant, Cell & Environment*, 13(7):667–679, 2006.

[11] Sönke Johnsen and Kenneth J Lohmann. The physics and neurobiology of magnetoreception. *Nature Reviews Neuroscience*, 6(9):703–712, 2005.

[12] Charles Kittel and Herbert Kroemer. *Thermal physics.* WH Freeman, 1980.

[13] Rolf Landauer. Irreversibility and heat generation in the computing process. *IBM journal of research and development*, 5(3):183–191, 1961.

[14] Joel L Lebowitz and Herbert Spohn. A Gallavotti–Cohen-type symmetry in the large deviation functional for stochastic dynamics. *Journal of Statistical Physics*, 95(1):333–365, 1999.

[15] David JC MacKay. *Information theory, inference and learning algorithms.* Cambridge University Press, 2003.

[16] Jerry B Marion and Stephen T Thornton. *Classical dynamics of particles and systems.* Saunders College Pub., 1995.

[17] Pankaj Mehta and David J Schwab. The Energetic Costs of Cellular Computation. *arXiv preprint arXiv:1203.5426*, 2012.

[18] Sung Y Park and Anil K Bera. Maximum entropy autoregressive conditional heteroskedasticity model. *Journal of Econometrics*, 150(2):219–230, 2009.

[19] Grzegorz Rozenberg, Thomas Bck, and Joost N Kok. *Handbook of natural computing.* Springer Publishing Company, Incorporated, 2011.

[20] Claude Elwood Shannon, Warren Weaver, Richard E Blahut, and Bruce Hajek. *The mathematical theory of communication*, volume 117. University of Illinois press Urbana, 1949.

[21] T Speck and U Seifert. Integral fluctuation theorem for the housekeeping heat. *Journal of Physics A: Mathematical and General*, 38(34):L581, 2005.

[22] Susanne Still, David A Sivak, Anthony J Bell, and Gavin E Crooks. Thermodynamics of Prediction. *Physical Review Letters*, 109(12):120604, 2012.

[23] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[24] Wung-Wai Tso and Julius Adler. Negative chemotaxis in escherichia coli. *Journal of bacteriology*, 118(2):560–576, 1974.

[25] Hanjing Yang, Hachiro Inokuchi, and Julius Adler. Phototaxis away from blue light by an Escherichia coli mutant accumulating protoporphyrin IX. *Proceedings of the National Academy of Sciences*, 92(16):7332–7336, 1995.

[26] Arnold Zellner and Richard A Highfield. Calculation of maximum entropy distributions and approximation of marginal posterior distributions. *Journal of Econometrics*, 37(2):195–209, 1988.