Information Bottleneck in Molecular Sensing

Marianne Bauer [1,2] and William Bialek [2,3]
[1]Department of Bionanoscience, Kavli Institute of Nanoscience Delft, Technische Universiteit Delft, Van der Maasweg 9, 2629 HZ Delft, The Netherlands
[2]Joseph Henry Laboratories of Physics and Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey 08544, USA
[3]Center for Studies in Physics and Biology, Rockefeller University, 1230 York Avenue, New York, New York 10065, USA

(Received 13 April 2023; accepted 24 October 2023; published 15 November 2023; corrected 13 December 2023)

Information of relevance to the organism often is represented by the concentrations of particular molecules inside a cell. As outside observers we can now measure these concentrations precisely, but the cell's own mechanisms must be noisier. We show that, in certain limits, there is a universal tradeoff between the information capacity of these concentration measurements and the amount of relevant information that is captured, a version of the information bottleneck problem. This universal tradeoff is confirmed, quantitatively, in an analysis of the positional information encoded by the "gap genes" in the developing fly embryo.

DOI: 10.1103/PRXLife.1.023005

I. INTRODUCTION

In the physics laboratory, and in engineered devices, we are used to information being represented by electrical or optical signals. While the brain also uses electrical signaling, inside living cells information often is represented by the concentrations of particular molecules. The absolute concentrations of these molecules, and even their total number, can be quite small. As a result there has been considerable interest in understanding the physical limits to this molecular signaling [1–9], the strategies that cells can use to maximize information in the face of these limits [8,10–14], and the implications for cellular function.

One approach to understanding signaling via molecular concentrations is to explore increasingly realistic models of the microscopic events [15–21]. Recently, we suggested a different approach, in which we ask abstractly about the implications of noise in the response, or more precisely about the limited information capacity of the cell's "measurements" of concentration [22,23]. To formalize the problem, we imagine that some relevant signal x is represented by the concentrations of K different molecules, which we write as $\mathbf{g} \equiv \{g_1, g_2, \ldots, g_K\}$. The cell does not have access to the exact values of $\mathbf{g}$, but only to some variables that constitute intermediates in the response. As an example, if the molecules act by binding to particular sites along the cell's DNA and thereby regulating the expression of downstream genes, the intermediate variable might be the average occupancy of these binding sites over some relevant time window, or the state of the enhancers built out of groups of these binding sites [16,17,24]. Independent of molecular details, this intermediate variable can carry only a limited amount of information about the real concentrations; in this sense it is a "compressed" representation [25], and we will refer to this representation as C.

Different mechanisms inside the cell will generate different mappings $\mathbf{g} \to C$, and in general we expect this mapping to be noisy, so it will be described by a probability distribution $P(C|\mathbf{g})$. What would be useful for the cell is to capture as much information as possible about the relevant variable x, subject to the constraint that the information about $\mathbf{g}$ is limited. This means that the best mapping $\mathbf{g} \to C$ is one that maximizes

$$\mathcal{U} = I(C;x) - \lambda I(C;\mathbf{g}), \tag{1}$$

where λ is a Lagrange multiplier to implement the constraint on information about $\mathbf{g}$. Here, $I(C;x)$ is the mutual information between C and x,

$$I(C;x) = \int dC \int dx\, P(C|x)P(x) \log\left[\frac{P(C|x)}{P(C)}\right], \tag{2}$$

and $I(C;\mathbf{g})$ is defined analogously. This optimization problem is an example of the information bottleneck problem [26], which arises in contexts ranging from text classification [27] to the analysis of deep networks [28–30] and neural coding [31,32].

Optimizing $\mathcal{U}$ will define a bounding curve, which shows the minimum $I(C;\mathbf{g})$ needed to reach a criterion level of $I(C;x)$. If the cell's measurements of concentration become more precise, then $I(C;\mathbf{g})$ becomes larger, but since C depends on $\mathbf{g}$ and not directly on x we always have $I(C;x) \leqslant I(\mathbf{g};x)$. The question is how close the cell can come to capturing all this available information given limits on the precision of its response.

II. SHARPENING THE QUESTION

The standard approach to solving the information bottleneck problem involves the numerical solution of the self-consistent equation that $P(C|\mathbf{g})$ obeys [26]; to do so, one assumes that C is discrete, so that the mapping $\mathbf{g} \to C$

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

becomes a kind of clustering. There is a separate bounding curve for each choice of the cardinality $||C||$, and the full structure emerges as we let $||C|| \to \infty$. We have applied this approach to the representation of positional information by the gap gene expression levels in the early fly embryo [22]. The results showed, for example, that the optimal mappings for individual genes are close to an intuitive thresholding model, but that optimal compression of multiple genes depends crucially on combinatorial interactions among these molecules, as seen experimentally, for example, in the regulation of the pair rule genes by the gap genes. But this discussion seemed to depend on details of the genetic network in the fly embryo, and missed the possibility that there is something more universal about the tradeoff between $I(C;\mathbf{g})$ and $I(C;x)$.

Our goal is to uncover this (asymptotically) universal tradeoff, which emerges in an intuitive limiting regime. This has a direct physical meaning, discussed below, but it also provides a solution to the information bottleneck problem without requiring numerics, which can become challenging with multiple, complex inputs. Perhaps the most similar analytic approach has been to assume that $P(x;\mathbf{g})$ is Gaussian [33], but we find we can make progress more generally by assuming that the solution to the problem $P(C|\mathbf{g})$ is a narrow Gaussian. This simplification is self-consistent if the signals $\mathbf{g}$ are sufficiently informative about x, and if the relevant mappings are sufficiently smooth, but we do not need to make assumptions about the form of $P(x;\mathbf{g})$.

To be concrete, we write the mean concentrations at each x as $\langle g_\mu\rangle_x$, which are smooth functions of x, and fluctuations around these means are described by a covariance matrix $\langle \delta g_\mu \delta g_\nu\rangle_x$. To be clear, we denote by $\langle\cdots\rangle_x$ an average over fluctuations at fixed x, while $\langle\cdots\rangle^{(x)}$ denotes an average over x. If $\langle g_\mu\rangle_x$ is smooth, it is reasonable to think that the compressed variable C also will have a smooth relation to other variables; this means that the moments of the distributions $P(C|\mathbf{g})$ and $P(C|x)$ will be smooth. Further, although a single measurement of concentration might be noisy, one can imagine that the noise level is smaller if we think about the encoding of the relevant variable x. These observations suggest that we search for optima in which C is a continuous variable, and that at appropriate points we can take a small noise limit. Concretely, let us assume that

$$P(C|\mathbf{g}) = \frac{1}{\sqrt{2\pi\sigma_C^2(\mathbf{g})}} \exp\left(-\frac{[C-\bar{C}(\mathbf{g})]^2}{2\sigma_C^2(\mathbf{g})}\right), \tag{3}$$

so that optimization of $\mathcal{U}$ now corresponds to finding the optimal functions $\bar{C}(\mathbf{g})$ and $\sigma_C(\mathbf{g})$.

III. UNIVERSALITY, ASYMPTOTICALLY

With the Gaussian approximation of Eq. (3) we can immediately write [25,34]

$$I(C;\mathbf{g}) = S(C) - \tfrac{1}{2}\left\langle \log_2\left[2\pi e\sigma_C^2(\mathbf{g})\right]\right\rangle^{(\mathbf{g})}, \tag{4}$$

where $S(C)$ is the entropy of the variable C, the second term the conditional entropy $S(C|\mathbf{g})$, and $\langle\cdots\rangle^{(\mathbf{g})}$ denotes an average over the distribution of $\mathbf{g}$. Importantly, $S(C)$ is finite in the limit of small noise, since C must be tied, even if implicitly, to the relevant variable x. In particular, if the effective noise in estimating x from $\mathbf{g}$ is small compared with the scale on which the distribution $P(x)$ varies, then this relationship becomes nearly deterministic [10], and we can write

$$P(C) = P(x)\left|\frac{d\bar{C}}{dx}\right|^{-1}, \tag{5}$$

$$S(C) = S(x) + \left\langle \log_2\left|\frac{d\bar{C}}{dx}\right|\right\rangle^{(x)}, \tag{6}$$

where the dependence of $\bar{C}$ on x is through $\mathbf{g}$,

$$\frac{d\bar{C}}{dx} = \sum_\mu \frac{d\langle g_\mu\rangle_x}{dx}\frac{\partial\bar{C}}{\partial g_\mu}\bigg|_{\mathbf{g}=\langle\mathbf{g}\rangle_x}, \tag{7}$$

again working in a small noise limit, and $\langle\cdots\rangle^{(x)}$ again denotes an average over the distribution of x.

To compute the information which C conveys about x, $I(C;x)$, we need the distribution

$$P(C|x) = \int d\mathbf{g} P(C|\mathbf{g})P(\mathbf{g}|x), \tag{8}$$

which in general is complicated, but if noise is small we can again make a Gaussian approximation. Then, we can expand $\bar{C}$, perform the integral over $\mathbf{g}$, and obtain the variance of C at fixed x; this variance has two components, one from the variance at fixed $\mathbf{g}$, and one from the (co)variance of $\mathbf{g}$ at fixed x, $\langle\delta g_\mu \delta g_\nu\rangle_x$:

$$\langle(\delta C)^2\rangle_x = \sigma_C^2(\mathbf{g})\big|_{\mathbf{g}=\langle\mathbf{g}\rangle_x} + A(x), \tag{9}$$

$$A(x) = \sum_{\mu\nu}\frac{\partial\bar{C}}{\partial g_\mu}\langle\delta g_\mu\delta g_\nu\rangle_x\frac{\partial\bar{C}}{\partial g_\nu}\bigg|_{\mathbf{g}=\langle\mathbf{g}\rangle_x}. \tag{10}$$

With this we have

$$I(C;x) = S(C) - \tfrac{1}{2}\langle\log_2[2\pi e\langle(\delta C)^2\rangle_x]\rangle^{(x)}, \tag{11}$$

and all the ingredients needed to express the objective function $\mathcal{U}$.

We notice that $\mathcal{U}$ is a local functional of $\bar{C}(\mathbf{g})$ and $\sigma_C^2(\mathbf{g})$, so we can use the calculus of variations in a familiar way. Optimizing with respect to $\sigma_C^2(\mathbf{g})$ is especially straightforward, and we find, using Eqs. (4), (6), and (11),

$$\frac{\partial\mathcal{U}}{\partial\sigma_C^2(\mathbf{g})} = 0 \tag{12}$$

$$\Rightarrow \sigma_C^2(\mathbf{g})\big|_{\mathbf{g}=\langle\mathbf{g}\rangle_x} = \frac{\lambda}{1-\lambda}A(x), \tag{13}$$

with A from Eq. (10). This makes sense, since it tells us that the precision of encoding $\mathbf{g}$ in C should be related to the scale of the fluctuations in $\mathbf{g}$ when the relevant variable x is fixed. At this optimum we have

$$I(C;\mathbf{g}) = S(x) - \frac{1}{2}\left\langle\log_2\left[\frac{2\pi e}{B(x)}\frac{\lambda}{1-\lambda}\right]\right\rangle^{(x)} \tag{14}$$

$$I(C;x) = S(x) - \frac{1}{2}\left\langle\log_2\left[\frac{2\pi e}{B(x)}\frac{1}{1-\lambda}\right]\right\rangle^{(x)}, \tag{15}$$

with $B = (1/A)[d\bar{C}/dx]^2$. If we can maximize $B(x)$ locally at each x, then we will have optimized $\mathcal{U}$. In what follows we simplify notation by leaving out the explicit x dependences, and understand that functions of $\mathbf{g}$ are all evaluated at $\langle\mathbf{g}\rangle_x$.

Before continuing with the general case, we can anticipate the universal tradeoff by considering the simpler case where there is only one variable g. Then,

$$A = A_1 = \left|\frac{d\bar{C}}{dg}\right|^2 \langle(\delta g)^2\rangle, \tag{16}$$

$$B = B_1 = \frac{1}{A_1}\left|\frac{d\bar{C}}{dx}\right|^2. \tag{17}$$

Since we are working in the limit where noise is small,

$$\frac{d\bar{C}}{dx} = \frac{d\bar{C}}{dg}\cdot\frac{d\langle g\rangle_x}{dx} \tag{18}$$

$$\Rightarrow B = \frac{1}{\langle(\delta g)^2\rangle}\left|\frac{d\langle g\rangle_x}{dx}\right|^2 = \frac{1}{\sigma_x^2(g)}, \tag{19}$$

where in the last step we recognize σ_x as the error bar in estimating x from the vector of concentration g [35,36]; this identification itself is correct only in the low-noise limit. Substituting into Eq. (14), we have

$$I(C;g) = S(x) - \frac{1}{2}\left\langle \log_2\left[2\pi e\sigma_x^2(g)\frac{\lambda}{1-\lambda}\right]\right\rangle^{(x)}, \tag{20}$$

and we identify the mutual information between g and x,

$$I(g;x) = S(x) - \tfrac{1}{2}\left\langle \log_2\left[2\pi e\sigma_x^2(g)\right]\right\rangle^{(x)}. \tag{21}$$

This allows us to write

$$\Rightarrow I(C;g) = I(g;x) - \frac{1}{2}\log_2\left(\frac{\lambda}{1-\lambda}\right). \tag{22}$$

Similarly, we find

$$I(C;x) = I(g;x) - \frac{1}{2}\log_2\left(\frac{1}{1-\lambda}\right). \tag{23}$$

Thus, the optimized $I(\mathbf{g};x)$ and $I(C;x)$ that form the bounding curve of the information bottleneck problem are related to one another in a way that is independent of the detailed structure of the problem, as encoded in the distribution $P(\mathbf{g};x)$. Perhaps surprisingly, we will see that with multiple variables we arrive at the same answer by optimizing the function $\bar{C}(\mathbf{g})$.

Returning to the general case, we have seen that optimizing $\mathcal{U}$ is equivalent to optimizing B, where

$$B = \frac{1}{A}\left[\sum_\mu \frac{\partial\bar{C}}{\partial g_\mu}\frac{d\langle g_\mu\rangle_x}{dx}\right]^2, \tag{24}$$

with A from Eq. (10). If we shift $\bar{C}(\mathbf{g}) \to \bar{C}(\mathbf{g}) + \delta\bar{C}(\mathbf{g})$, then $B \to B + \delta B$, where

$$\delta B = \frac{2}{A}\sum_\mu V_\mu\frac{\partial\bar{C}}{\partial g_\mu},$$

$$V_\mu = -B\sum_\nu \frac{\partial\bar{C}}{\partial g_\nu}\langle\delta g_\nu\delta g_\mu\rangle_x + \left[\sum_\nu \frac{\partial\bar{C}}{\partial g_\nu}\frac{d\langle g_\nu\rangle_x}{dx}\right]\frac{d\langle g_\mu\rangle_x}{dx}. \tag{25}$$

Because we need to know functions of $\mathbf{g}$ only along the one-dimensional trajectory $\mathbf{g} = \langle\mathbf{g}\rangle_x$, we have enough freedom for $\partial\bar{C}/\partial g_\mu$ to be an arbitrary vector. Thus to find an extremum $\delta B = 0$ we need $V_\mu = 0$, which we can rewrite as

$$\sqrt{AB}\frac{d\langle g_\mu\rangle_x}{dx} = B\sum_\nu\langle\delta g_\mu\delta g_\nu\rangle_x\frac{\partial\bar{C}}{\partial g_\nu}$$

$$\Rightarrow \frac{\partial\bar{C}}{\partial g_\nu} = \sqrt{\frac{A}{B}}\sum_\mu[(\langle\delta g\delta g\rangle_x)^{-1}]_{\nu\mu}\frac{d\langle g_\mu\rangle_x}{dx}. \tag{26}$$

This yields a simple expression for B,

$$B = \sum_{\mu\nu}\frac{d\langle g_\mu\rangle_x}{dx}[(\langle\delta g\delta g\rangle_x)^{-1}]_{\mu\nu}\frac{d\langle g_\nu\rangle_x}{dx} \tag{27}$$

$$= \frac{1}{\sigma_x^2(\mathbf{g})}, \tag{28}$$

where $\sigma_x(\mathbf{g})$ now is the error bar in estimating x from the entire vector of concentrations $\mathbf{g}$ [35,36].

As in the case of one variable, we identify the information that the concentrations $\mathbf{g}$ provide about x,

$$I(\mathbf{g};x) = S(x) - \tfrac{1}{2}\left\langle \log_2\left[2\pi e\sigma_x^2(\mathbf{g})\right]\right\rangle^{(\mathbf{g})}. \tag{29}$$

Finally, putting the different terms together we recover, analogously to Eqs. (22) and (23),

$$I(C;x) = I(\mathbf{g};x) - \frac{1}{2}\log_2\left(\frac{1}{1-\lambda}\right), \tag{30}$$

$$I(C;\mathbf{g}) = I(\mathbf{g};x) - \frac{1}{2}\log_2\left(\frac{\lambda}{1-\lambda}\right). \tag{31}$$

These equations imply that with $0 < \lambda < 1$, the information $I(C;x)$ that is captured about the relevant variable is less than the available information $I(\mathbf{g};x)$, as it must be. As $\lambda \to 0$ this gap closes, but at the expense of requiring an increasing information capacity in the mapping $\mathbf{g} \to C$. Taken together, Eqs. (30) and (31) define the bounding curve in the information plane, as shown in Fig 1.

The result in Eqs. (30) and (31) is surprising, because all details of the underlying system have disappeared. This asymptotically universal bounding curve suggests there is a tradeoff between the capacity of the cell to measure the concentration of signaling molecules and the resulting ability to capture relevant information, independent of molecular mechanisms, at least in some regime. Are real cells in this regime?

IV. THE EARLY FLY EMBRYO

The gap genes form a network that is crucial to the early events of fly development [37,38]. These four genes take inputs from primary maternal morphogens and in turn drive the striped patterns of pair-rule gene expression. The local concentrations of gap gene proteins provide enough information to specify position to ∼1% accuracy along the anterior-posterior axis, and this is the precision with which the stripes are positioned [35,39,40]. The algorithm that achieves optimal readout of this positional information predicts, quantitatively, the distortions of the pair-rule stripes in mutant flies where individual maternal inputs are deleted [41].

The concentrations of the gap gene proteins ($\mathbf{g}$) encode information about position (x), providing an example of the problem we have been discussing. This positional information

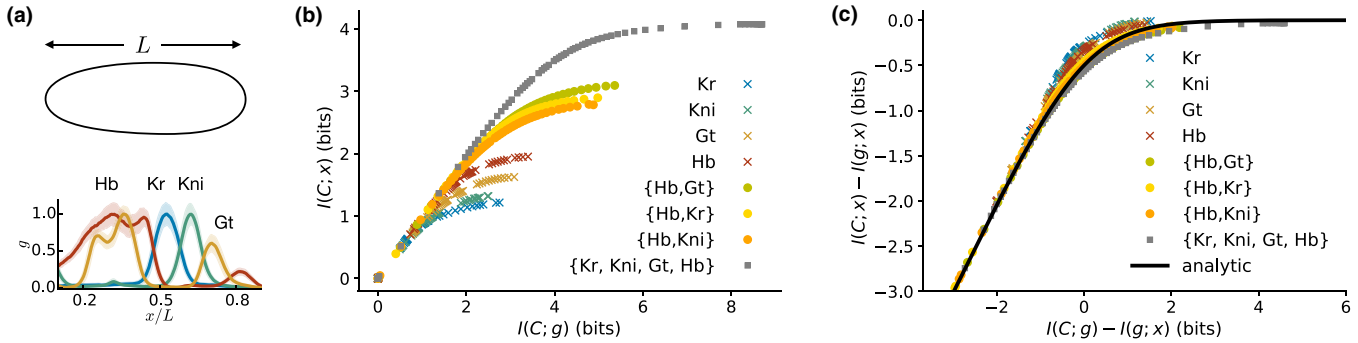

FIG. 1. (a) The four gap genes Hb, Kr, Kni, and Gt are expressed in varying concentrations g along the anterior-posterior axis of the fly embryo x [41]; $L \sim 0.5$ mm. Mean profiles $\langle g_i \rangle_x$ in solid lines, shading shows standard deviations $\sqrt{\langle (\delta g_i)^2 \rangle_x}$. (b) The information plane $I(C;x)$ vs $I(C;\mathbf{g})$ for optimized C when $\mathbf{g}$ is each of the four gap gene proteins separately, some exemplary pairs, or all gap gene proteins together; computations follow the methods of Ref. [22]. (c) Collapse of the data onto the universal information tradeoff, as predicted from Eqs. (30) and (31), in black. In these coordinates, both $I(C;x)$ and $I(C;\mathbf{g})$ are measured relative to the maximum available positional information, $I(\mathbf{g};x)$.

is important to the organism, as nuclei must make distinct cell-fate decisions in their further development; $I(\mathbf{g};x)$ bounds the (log) number of distinct cell fate decisions can be placed in a reliable relation to position. Figure 1(a) shows the means and standard deviations of $\mathbf{g}$ as a function of x. With four gap genes, we can analyze the information that they carry individually, or in groups. To test the predictions of our theory we need to compare with the full numerical solution of the bottleneck problem in each of these cases.

We emphasize that the solution of the full information bottleneck problem is determined by the joint distribution $P(\mathbf{g}, x)$, which must be estimated from experiment. Thus, while it emerges as the solution of an optimization problem, the true bounding curve in the plane $I(C;x)$ vs $I(C;\mathbf{g})$ is a property of the data, and should be seen as a experimental result that we can compare to the theory developed here. Following previous analyses [35,41], we make use of experiments that measure the concentrations $\mathbf{g}$ at each point x in a large number of embryos. We approximate $P(\mathbf{g}|x)$ as Gaussian, and estimate $\langle g_\mu \rangle_x$ and $\langle \delta g_\mu \delta g_\nu \rangle_x$ in a small window of time ∼42 min into nuclear cycle 14. The mapping $P(C|\mathbf{g})$ that optimizes $\mathcal{U}$ is the solution of the self–consistent equation [26]

$$P(C|\mathbf{g}) = \frac{P(C)}{Z(\mathbf{g},\lambda)} \exp\left[-\frac{1}{\lambda}\int dx P(x|\mathbf{g}) \ln\left(\frac{P(x|\mathbf{g})}{P(x|C)}\right)\right],$$

where Z is a normalization constant. To solve this numerically we assume that C is a discrete variable. For each cardinality $||C||$ we find a curve of $I(C;x)$ vs $I(C;\mathbf{g})$, and the true bounding curve is obtained at large $||C||$ [42]; details are in the supplement to Ref. [22]. Figure 1(b) shows the optimal $I(C;x)$ vs $I(C;\mathbf{g})$ for the numerically optimized $P(C|\mathbf{g})$ with λ varying along the curves. We emphasize that *any* molecular mechanism by which the embryo responds to the concentrations $\mathbf{g}$ must fall on a point below the bounding curve.

Figure 1(c) shows that these curves collapse when shifted by the mutual information $I(\mathbf{g};x)$, as predicted in Eqs. (22) and (23). Notice that these shifts vary by up to ∼2.5 bits across the different groups. We see that these real examples follow the predictions of the universal tradeoff (in black) quite accurately.

Our theory predicts that the universal tradeoff is true asymptotically, which means that it does not provide an upper (or lower) bound for the data. If we look at the gap genes individually, there are regions along the x axis where the small noise approximation fails, and we expect deviations from the predicted behavior; it is perhaps surprising that these deviations [points versus line in Fig. 1(c)] are so small. One of the important features of these data is that as we include more of the gap genes in our analysis the effective positional noise becomes uniformly small along the full length of the embryo [35,41]. Thus the predictions of a universal tradeoff should be most accurate in the case of all four genes, and this is what we see.

To explore the applicability of our universal tradeoff, we generate artificial data that is a perturbation of the real data from the fly embryo. Specifically we hold the means $\langle g_i \rangle_x$ fixed and vary the structure of the noise, then analyze the tradeoff between $I(C;x)$ and $I(C;\mathbf{g})$ in Fig. 2. We perform this analysis both for all four genes together [Fig. 2(a)] and for Hb alone [Fig. 2(b)].

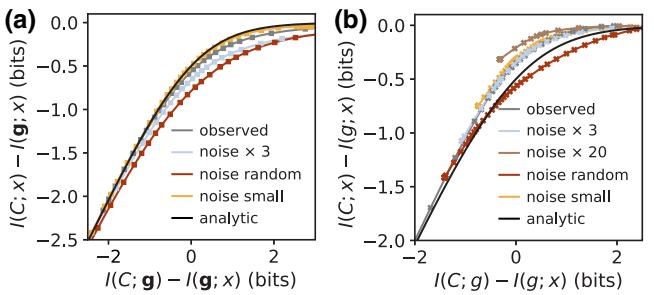

FIG. 2. Applicability of the universal tradeoff. We generated modified gene expression profiles, inspired by all genes (a) and Hb alone (b), but changed the noise profiles. The IB compression from the observed expression data in the fly embryo are shown in gray in both panels. Light blue and light brown [only in (b)] corresponds to increasing the naturally observed noise three and 20 times, respectively; for the latter the IB curve in rescaled units is no longer captured by the universal tradeoff. A randomized noise profile (multiplied at each x by a random number up to 9) lowers the IB curve below the tradeoff. Small, uniform noise $[\langle (\delta g_i)^2 \rangle = (0.03)^2]$ shows perfect alignment with the optimal compression for Hb and perfect alignment with our universal tradeoff for all genes.

We expect our tradeoff to be valid when data means are monotonic and smooth, and when the noise is smooth and small compared to changes in the mean. For the data inspired by all four genes, a smooth and small noise profile matches the tradeoff exactly [Fig. 2(a)]. On the other hand, threefold increases in natural noise as well as a randomly chosen, heterogeneous noise profile push the IB curves below the real data and our universal tradeoff.

For the case of a single gene (Hb), deviations from the universal tradeoff are larger in the real data and persist even when we lower the noise level [Fig. 2(b)]. As expected, deviations are even larger if we increase the noise level by a large factor, and have the opposite sign if we replace the observed noise levels with a random function of x. We conclude that our tradeoff is valid asymptotically and relies on the smooth, small-noise regime.

V. IMPLICATIONS

The fact that a real system follows the universal tradeoff in Eqs. (22) and (23) invites us to consider the implications. We notice that the theoretical prediction is close to being a line of unit slope, $I(C;x) \sim I(C;\mathbf{g})$, ending in saturation at $I(C;x) = I(\mathbf{g};x)$. We can never have $I(C;x) > I(C;\mathbf{g})$, and it is interesting that, under fairly general conditions, it is possible to approach this maximal efficiency of $I(C;x) \sim I(C;\mathbf{g})$: If noisy mechanisms can keep only I bits of information about the signaling molecule concentrations, then it is possible for all of these I bits to be relevant for the organism, even when the mappings among signals are complicated, as in the pattern of gap gene expression versus position. This proximity of the information bottleneck solution to the bound $I(C;x) \leqslant I(C;\mathbf{g})$ is not true in general [43], and in many problems $I(C;x)$ diverges from $I(C;\mathbf{g})$ already close to the origin. We can see how close this tradeoff is to the diagonal $I(C;x) = I(C;\mathbf{g})$ by calculating $I(C;x)$ at the point $I(C;\mathbf{g}) = I(\mathbf{g},x)$; from Eq. (23) this happens at $\lambda = 0.5$, where Eq. (22) predicts that $I(C;x) = I(\mathbf{g},x) - 0.5$ bits. Thus, in the regime we are considering, sensors with "just enough" capacity to transmit all the information provided by the signaling molecules can come close to deploying all this capacity for the relevant information.

On the other hand, it is worth emphasizing that no real mechanism can extract *all* of the information that is available about x from a perfect measurement of $\mathbf{g}$. If the available information is I, then to get within ϵ bits Eq. (23) tells us that we need a mechanism with capacity $I(C;\mathbf{g}) \sim I - (1/2)\log_2(2\epsilon \ln 2)$, where we expand the logarithm in Eq. (31) in $\lambda \approx 2\epsilon \ln 2$ at small ϵ. If the cell needs to make a binary decision, then making errors with probability q causes an information loss ϵ which is just the entropy of these errors. At small q this entropy is $\epsilon \sim q\log_2(e/q)$ bits. This means, for example, that if the initial signals $\mathbf{g}$ are just sufficient to provide one bit of information, the cell would need to read these signals with $\sim$2.5 bits of accuracy in order to keep errors below $q \sim 1\%$. The requirements are even more stringent if we imagine that the initial signal $\mathbf{g}$ is processed through several layers. The perhaps surprising conclusion is that mechanisms with one bit of information capacity are not sufficient for cells to make reliable binary decisions. More generally, cells must sense concentrations with mechanisms that have nearly 2 bits more capacity than the relevant information that they need to extract.

VI. CONCLUSION

In conclusion, we have derived a universal tradeoff between $I(C;x)$ and $I(C;\mathbf{g})$ in the limit where mappings are smooth and the effective noise level is small. This provides analytic control over the information bottleneck problem in a regime that is different from the case where $P(\mathbf{g};x)$ is Gaussian [33]. Importantly, we find that for the encoding of positional information by transcription factors in the fly embryo, our universal tradeoff captures the results from full numerical optimization very well. We can think of the compressed variable C as representing the state of the enhancers that respond directly to the transcription factors and determine the expression of "downstream" genes. The universal tradeoff gives us a path to analytic understanding of the information capacity that these enhancers must achieve in order to extract the limited information available about the ultimate body plan of the organism [22,23].

ACKNOWLEDGMENTS

We thank our experimental colleagues T. Gregor, M. D. Petkova, and E. F. Wieschaus, whose results inspired these ideas. This work was supported in part by the National Science Foundation through the Center for the Physics of Biological Function (PHY–1734030), by fellowships from the Simons Foundation and the John Simon Guggenheim Memorial Foundation (W.B.), and by a start-up grant from the Bionanoscience Department at TU Delft (M.B.).

[1] H. C. Berg and E. M. Purcell, Physics of chemoreception, Biophys. J. **20**, 193 (1977).
[2] W. Bialek and S. Setayeshgar, Physical limits to biochemical signaling, Proc. Natl. Acad. Sci. USA **102**, 10040 (2005).
[3] W. Bialek and S. Setayeshgar, Cooperativity, sensitivity, and noise in biochemical signaling, Phys. Rev. Lett. **100**, 258101 (2008).
[4] R. G. Endres and N. S. Wingreen, Maximum likelihood and the single receptor, Phys. Rev. Lett. **103**, 158101 (2009).
[5] T. Mora and N. S. Wingreen, Limits of sensing temporal concentration changes by single cells, Phys. Rev. Lett. **104**, 248101 (2010).
[6] K. Kaizu, W. H. de Ronde, J. Paijmans, K. Takahashi, F. Tostevin, and P. R. ten Wolde, The Berg–Purcell limit revisited, Biophys. J. **106**, 976 (2014).
[7] T. Mora, Physical limit to concentration sensing amid spurious ligands, Phys. Rev. Lett. **115**, 038102 (2015).
[8] P. R. ten Wolde, N. B. Becker, T. E. Ouldridge, and A. Mugler, Fundamental limits to cellular sensing, J. Stat. Mech. **162**, 1395 (2016).
[9] T. Mora and I. Nemenman, Physical limit to concentration sensing in a changing environment, Phys. Rev. Lett. **123**, 198101 (2019).

[10] G. Tkačik, Jr., C. G. Callan, Jr., and W. Bialek, Information flow and optimization in transcriptional regulation, Proc. Natl. Acad. Sci. USA **105**, 12265 (2008).
[11] F. Tostevin and P. R. ten Wolde, Mutual information between input and output trajectories of biochemical networks, Phys. Rev. Lett. **102**, 218101 (2009).
[12] G. Tkačik and A. M. Walczak, Information transmission in genetic regulatory networks: A review, J. Phys.: Condens. Matter **23**, 153102 (2011).
[13] E. D. Siggia and M. Vergassola, Decisions on the fly in cellular sensory systems, Proc. Natl. Acad. Sci. USA **110**, E3704 (2013).
[14] N. B. Becker, A. Mugler, and P. R. ten Wolde, Optimal prediction by cellular signaling networks, Phys. Rev. Lett. **115**, 258103 (2015).
[15] G. K. Ackers, A. D. Johnson, and M. A. Shea, Quantitative model for gene regulation by λ phage repressor, Proc. Natl. Acad. Sci. USA **79**, 1129 (1982).
[16] M. Ptashne and A. Gann, *Genes and Signals* (Cold Spring Harbor Press, New York, 2002).
[17] L. Bintu, N. E. Buchler, H. G. Garcia, U. Geraland, T. Hwa, J. Kondev, and R. Phillips, Transcriptional regulation by the numbers: Models, Curr. Opin. Genet. Dev. **15**, 116 (2005).
[18] D. Hnisz, K. Shrinivas, R. A. Young, A. K. Chakraborty, and P. A. Sharp, A phase separation model for transcriptional control, Cell **169**, 13 (2017).
[19] M. Morrison, M. Razo-Mejia, and R. Phillips, Reconciling kinetic and thermodynamic models of bacterial transcription, PLoS Comput. Biol. **17**, e1008572 (2021).
[20] B. Zoller, T. Gregor, and G. Tkačik, Eukaryotic gene regulation at equilibrium, or non? Curr. Opin. Sys. Biol. **31**, 100435 (2022).
[21] R. Martinez-Corral, M. Park, K. Biette, D. Friedrich, C. Scholes, A. S. Khalil, J. Gunawardena, and A. H. DePace, Transcriptional kinetic synergy: A complex landscape revealed by integrating modelling and synthetic biology, Cell Syst. **14**, 324 (2023).
[22] M. Bauer, M. D. Petkova, T. Gregor, E. F. Wieschaus, and W. Bialek, Trading bits in the readout from a genetic network, Proc. Natl. Acad. Sci. USA **118**, e2109011118 (2021).
[23] M. Bauer, How does an organism extract relevant information from transcription factor concentrations? Biochem Soc. Trans. **50**, 1365 (2022).
[24] E. E. M. Furlong and M. Levine, Developmental enhancers and chromosome topology, Science **361**, 1341 (2018).
[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
[26] N. Tishby, F. C. Pereira, and W. Bialek, The information bottleneck method, in *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, edited by B. Hajek and R. S. Sreenivas (University of Illinois, Urbana, 1999), pp. 368–377; arXiv:physics/0004057.
[27] N. Slonim and N. Tishby, Document clustering using word clusters via the information bottleneck method, in *Proceedings of the 23rd Annual International ACM SIGIR Conference* (ACM, New York, 2000), pp. 208–215.
[28] N. Tishby and N. Zaslavsky, Deep learning and the information bottleneck principle, in *2015 IEEE Information Theory Workshop (ITW)* (IEEE, New York, 2015), pp. 1–5.
[29] A. Saxe, Y. Bansai, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox, On the information bottleneck theory of deep learning, J. Stat. Mech. (2019) 124020.
[30] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, Deep variational information bottleneck, ICLR (2017), arXiv:1612.00410.
[31] E. Schneidman, W. Bialek, and M. J. Berry II, An information theoretic approach to the functional classification of neurons, in *Advances in Neural Information Processing 15*, edited by S. Becker, S. Thrun, and K. Obermayer (MIT Press, Cambridge, MA, 2003), pp. 197–204.
[32] L. Buesing and W. Maass, A spiking neuron as information bottleneck, Neural Comput. **22**, 1961 (2010).
[33] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, Information bottleneck for Gaussian variables, in *Advances in Neural Information Processing Systems 16*, edited by S. Thrun, L. K. Saul, and B. Schölkopf (MIT Press, Cambridge, MA, 2004), pp. 1213–1220.
[34] W. Bialek, *Biophysics: Searching for Principles* (Princeton University Press, Princeton, NJ, 2012).
[35] J. O. Dubuis, G. Tkačik, E. F. Wieschaus, T. Gregor, and W. Bialek, Positional information, in bits, Proc. Natl. Acad. Sci. USA **110**, 16301 (2013).
[36] G. Tkačik, J. O. Dubuis, M. D. Petkova, and T. Gregor, Positional information, positional error, and readout precision in morphogenesis: A mathematical framework, Genetics **199**, 39 (2015).
[37] C. Nüsslein-Volhard and E. F. Wieschaus, Mutations affecting segment number and polarity in Drosophila, Nature (London) **287**, 795 (1980).
[38] J. Jaeger, The gap gene network, Cell. Mol. Life Sci. **68**, 243 (2011).
[39] F. Liu, A. H. Morrison, and T. Gregor, Dynamic interpretation of maternal inputs by the Drosophila segmentation gene network, Proc. Natl. Acad. Sci. USA **110**, 6724 (2013).
[40] V. Antonetti, W. Bialek, T. Gregor, G. Muhaxheri, M. Petkova, and M. Scheeler, Precise spatial scaling in the early fly embryo, arXiv:1812.11384.
[41] M. D. Petkova, G. Tkačik, W. Bialek, E. F. Wieschaus, and T. Gregor, Optimal decoding of cellular identities in a genetic network, Cell **176**, 844 (2019).
[42] We use $||C|| = 70$ for the individual genes and $||C|| = 800$ for the combination of all four genes. We have checked that all information theoretic quantities have saturated once $||C||$ is this large.
[43] V. Ngampruetikorn and D. J. Schwab, Perturbation theory for the information bottleneck, in *Advances in Neural Information Processing Systems 34*, edited by M. Ranzato *et al.* (NeurIPS, San Diego, 2021), pp. 21008–21018.

Correction: Panels (a) and (b) in the previously published Figure 2 were interchanged and have been set right.