ESTIMATING INFORMATION FLOW IN SINGLE CELLS

by

Sarah Anhala Cepeda Humerez

May, 2019

A thesis presented to the
Graduate School
of the
Institute of Science and Technology Austria, Klosterneuburg, Austria
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

The thesis of Sarah Anhala Cepeda Humerez, titled *ESTIMATING INFORMATION FLOW IN SINGLE CELLS*, is approved by:

Supervisor: Gašper Tkačik, IST Austria, Klosterneuburg, Austria

Signature: _____

Committee Member: Edouard Hannezo, IST Austria, Klosterneuburg, Austria

Signature: _____

Committee Member: Peter Swain, University of Edinburgh, Edinburgh, United Kingdom

Signature: _____

Defense Chair: Maximilian Jösch, IST Austria, Klosterneuburg, Austria

Signature: _____

signed page is on file

IST Austria Thesis, ISSN: 2663-337X

I hereby declare that this thesis is my own work and that it does not contain other people's work without this being so stated; this thesis does not contain my previous work without this being stated, and the bibliography contains all the literature that I used in writing the dissertation.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee, and that this thesis has not been submitted for a higher degree to any other university or institution.

I certify that any republication of materials presented in this thesis has been approved by the relevant publishers and co-authors.

Signature: _____

Sarah Anhala Cepeda Humerez
May, 2019
signed page is on file

Abstract

Single cells are constantly interacting with their environment and each other, more importantly, the accurate perception of environmental cues is crucial for growth, survival, and reproduction. This communication between cells and their environment can be formalized in mathematical terms and be quantified as the information flow between them, as prescribed by information theory.

The recent availability of real–time dynamical patterns of signaling molecules in single cells has allowed us to identify encoding about the identity of the environment in the time–series. However, efficient estimation of the information transmitted by these signals has been a data–analysis challenge due to the high dimensionality of the trajectories and the limited number of samples. In the first part of this thesis, we develop and evaluate decoding–based estimation methods to lower bound the mutual information and derive model–based precise information estimates for biological reaction networks governed by the chemical master equation. This is followed by applying the decoding-based methods to study the intracellular representation of extracellular changes in budding yeast, by observing the transient dynamics of nuclear translocation of 10 transcription factors in response to 3 stress conditions. Additionally, we apply these estimators to previously published data on ERK and Ca^{2+} signaling and yeast stress response. We argue that this single cell decoding-based measure of information provides an unbiased, quantitative and interpretable measure for the fidelity of biological signaling processes.

Finally, in the last section, we deal with gene regulation which is primarily controlled by transcription factors (TFs) that bind to the DNA to activate gene expression. The possibility that non-cognate TFs activate transcription diminishes the accuracy of regulation with potentially disastrous effects for the cell. This 'crosstalk' acts as a previously unexplored source of noise in biochemical networks and puts a strong constraint on

their performance. To mitigate erroneous initiation we propose an out of equilibrium scheme that implements kinetic proofreading. We show that such architectures are favored over their equilibrium counterparts for complex organisms despite introducing noise in gene expression.

Acknowledgments

To my parents Mabel and Felix: because I owe it all to both of you. Many thanks!
My friend Felix, the idealist, I miss going with you to the mountains, you have definitely taught me more with silence and contemplation than with words, thanks for creating a different world around me so that I could become a different person. My strength Mabel, you have done a great job 24/7 during my entire life, always present with the appropriate advice and permanently connected to my feelings and emotions, thanks for all your daily sacrifices and support.
I am grateful to my siblings and grandmother: Marihe, Abraham and Teresa, for their moral and emotional support during my life. I am also thankful to Jose Hernandez, my life-coach and friend, who has been by my side at every step of my life during the last 15 years, his wisdom and knowledge bring light and clarity to my life.

A very special gratitude to my supervisor Gasper Tkačik: who has been patient, inspiring, encouraging and positive. To my external supervisor and collaborator Peter Swain, to the other members of my Ph.D. committee: Edouard Hannezo and Caroline Uhler.
I am also grateful to my colleagues in the group, especially to Roshan Prizak, Thomas Sokolowsky, Jakob Ruess, Georg Ryeckh and Michal Hledik. To my collaborators: Alejandro Granados, Abraham Martin del Campo and Julian Pietsch.
Finally, last but by no means least, thanks to IST Austria: the group leaders, the grad school and all my colleagues and friends, it has been great sharing the interdisciplinary atmosphere at the institute during the last years.

About the Author

Sarah Cepeda-Humerez completed BSc in Physics at the University UATF in Potosí-Bolivia and MSc in Statistical and Condensed Matter physics at the Abdus Salam international center for theoretical physics (ICTP) in Trieste-Italy, before joining IST Austria in September 2013. Her main research interests include information transmission through dynamical signals in single cells signaling and gene regulation. During her PhD studies she worked on the research project "Distributed and dynamic intracellular organization of extracellular information" with the group of Peter Swain at the University of Edinburgh, and has developed the estimation methodology used in the collaborative project with Swain et al. During her PhD studies, Sarah has also presented her research results in the EMBO conference in Heidelberg in 2017.

List of Publications

1. **Sarah A. Cepeda-Humerez**, Georg Rieckh, and Gašper Tkačik. 2015. Stochastic proofreading mechanism alleviates crosstalk in transcriptional regulation. **Phys. Rev. Lett.**, 115:248101.

2. Abraham Martn del Campo, **Sarah Cepeda**, Caroline Uhler. 2017. Exact goodnessoffit testing for the ising model. **Scandinavian Journal of Statistics**.

3. Alejandro A Granados, Julian M J Pietsch, **Sarah A. Cepeda-Humerez**, Iseabail L Farquhar, Gašper Tkačik, and Peter S Swain. 2018. Distributed and dynamic intracellular organization of extracellular information. **Proceedings of the National Academy of Sciences of the United States of America**, 115(23):6088 6093.

4. **Sarah A. Cepeda-Humerez**, Jakob Ruess and Gašper Tkačik. 2019. Estimating information in time-varying signals (preprint on the arXiv q-bio).

Table of Contents

List of Figures

1 Introduction

1.1 Motivation

The major challenge in the field of biophysics during the last decade has been to find biological principles (similar to the laws of physics), that could provide a unified theory for living systems, in a way addressing the fundamental definition of life itself [Schrödinger and Penrose, 1992]. Along these lines, the recent availability of large and highly reproducible experimental data has enabled some theorists to think about such unifying principles. It is unclear what form such unifying principles could take, but most of the suggested ones take the form of "optimization principles", some of which try to capture the notion of "biological function" mathematically. For example, minimizing the expended energy, maximizing growth, etc. Amongst them, it has been proposed that biological systems are operating near an optimum of information transmission given biophysical constraints [Bialek, 2012; Tkačik and Bialek, 2016], and similar ideas propose that an essential characteristic of life is robustness ensuring that the essential function of the system is carried out precisely despite natural variations [Alon, 2006]. However, the task of describing and understanding the vast variety of life using a handful set of "laws" is just at the beginning [Bialek, 2018].

According to the systemic definition of life: [Maturana R and J. Varela, 1973; Varela *et al.*, 1974] living systems are *autopoietic*, where auto, means 'self' and poiesis, means 'creation or production'. Thus autopoiesis refers to self-generating or self-making systems. To carry out this task, they use, assimilate, and embody the elements available in the environment: nutrients, heat, light, and other forms of energy. Such

permanent exchange with the environment, and the ability to react to environmental changes accordingly, allows organisms to survive, find food and reproduce. Therefore, communication of living systems, which inherently requires mechanisms of information exchange with the environment, is relevant for survival. In some respect, this intuitive idea of information flow being crucial for life –that arises from the systemic definition of life itself– is reflected in the theoretical principle proposed by William Bialek [Bialek, 2018] and is grounded in experimental evidence [Bialek, 2012] [Bialek, 2012]. Our visual system can count single photons [Rieke and Baylor, 1998], bacteria follow the attractive chemical gradients with a reliability so high that they essentially detect every single molecule that arrives to their surface [Berg and Purcell, 1977], while during early development of *Drosophila*, single cells distinguish their position in the embryo with $\sim 1\%$ error (consistent with the statistically optimal use of the available information) [Dubuis *et al.*, 2013b; Dubuis *et al.*, 2013a].
It is important to mention that such claims about optimality can only emerge from appropriate experimental measurements, data analysis, and a theoretical framework based on physical laws.

The promising results about biological systems operating close to the optimum information transmission performance and the perspectives towards a unified theory of biological systems motivates this thesis to i) provide tools for data analysis of recently available single cell dynamic signals and ii) contribute to the understanding of mechanisms that biological systems may use for increasing their response accuracy to external signals.

1.2 A framework to study communication in living systems

The large variety of organisms in nature have evolved an enormous number of communication systems. We are primarily familiar with: language, sounds, images, smells and taste, and tactile perception. It is remarkable that all those signals are simultaneously integrated by our nervous system. Understanding how these signals are transmitted,

combined and how they eventually influence our behaviour, is one of the most challenging questions of our time. It would be out of the scope of a single work to study of a system of such complexity. Therefore, here we focus on single-cell organisms which can be studied experimentally and are also capable of: sensing their environment, performing computations, sending out signals and making decisions.
Many intriguing examples for such mechanisms can be found in different bacteria species: although their size is limited to a few microns, magnetotactic bacteria can use the earth's magnetic field to find the optimum concentration of oxygen in their habitat [S., 1963; Blakemore, 1975], *E. coli* can anticipate changes in the carbon sources of the environment, by observing subsequent changes of temperature and oxygen in its surroundings [Tagkopoulos *et al.*, 2008; Mitchell *et al.*, 2009]. In addition, apart from detection and prediction they perform several other functions that involve complex computations.

While single-cell organisms can respond to a large variety of signal types, as portrayed above, their primary way of environment perception is through biochemical signals. Namely, the receptors on the surface of a cell interact with extracellular signals and propagate the information towards the inside of the cell, e.g., to the nucleus. Such sequences of biochemical interactions are called signaling cascades; their main task consists of triggering gene regulatory networks further downstream which activate the production of biomolecules, called proteins, required by the cell to adequately respond to input signals. Although life scientists have classified gene regulatory networks as separate from signaling cascades according to our understanding of their functions, in nature they are not essentially different and share fundamental characteristics. For instance, the stochastic nature of the interactions is reflected by the fact that signaling molecules are randomly diffusing in the cell in order to find their cognate receptor to propagate the signals. This search process could be reasonably fast and reliable, if the copy number of each molecule would be large. However, this is often not the case since many molecules are present at very low copy numbers (i.e. as low as 10 molecules of some chemical species per bacterial cell), representing a major obstacle for efficient information transmission in cells.

The fluctuations in genetic and signaling networks is subject to the central limit theorem, with a consequence that the relative size of fluctuations decreases as the mean response grows large. In turn this means that for low copy numbers the fluctuations around the mean are relevant. By now, the noise in gene expression has been studied extensively both in theory and in the experiment. For the first time in the early 2000's direct measurements of noise in gene expression [Elowitz *et al.*, 2002; Blake *et al.*, 2003; Ozbudak *et al.*, 2002] and cell-to-cell variability [Feinerman *et al.*, 2008; Spencer *et al.*, 2009] became possible. The notion of efficient and reliable information transmission, discussed in the motivation, is obviously in stark contrast to the fact that the information transduction signals inside cells are highly stochastic. How it is possible for optimal performance and high precision to arises from individual stochastic components? To answer this question, we first need to know how noise limits the transmission of information through the system. In 1948, Claude Shannon introduced the appropriate mathematical formalism to address this question in his seminal paper: "The mathematical theory of communication" [Shannon and Weaver, 1949]. In this work, which lay the foundation of information theory, he presented a measure of information transmission through a noisy channel, called mutual information[1].

Shortly after the introduction of information theory, the ideas of efficient information transmission where adopted in biology, explicitly in the field of neuroscience [MacKay and McCulloch, 1952; Barlow, 1959]. Theoretical questions about efficient coding and signal's redundancy reduction were addressed already in these original studies. During the early 1990's, neural activity measurements of entropies and mutual information in response to complex, dynamic stimuli were carried out [Bialek *et al.*, 1991] followed by initial analysis tools for direct and unbiased estimates of information transmission from limited experimental data [Strong *et al.*, 1998; Borst and Theunissen, 1999].

Thanks to the availability of highly accurate data [Gregor *et al.*, 2007a; Gregor *et al.*, 2007b], in the late 2000's, similarly motivated studies explored genetic networks by focusing on the Bicoid/Hunchback system in the early *Drosophila* embryo [Tkačik *et al.*, 2008].

[1] The concept and formal definition of Mutual information will be further extended in section 2.1.3

Gene regulation – the ability of cells to modulate the expression levels of genes to match their current needs – is fundamental to normal growth, development and survival of an organism. Regulation of gene expression is primarily achieved through transcriptional regulation, where special regulatory proteins known as transcription factors (TFs) bind to specific sites on the DNA to either enhance or inhibit transcription of nearby genes. Because TF molecules are often present at low concentrations [Milo and Philips, 2016; Li *et al.*, 2014], transcriptional regulation is highly stochastic; this manifests itself as temporally variable gene expression even when the environmental conditions are held fixed. The stochasticity in gene expression has been extensively studied experimentally [Elowitz *et al.*, 2002; Blake *et al.*, 2003; Ozbudak *et al.*, 2002], with various theoretical and data-analysis frameworks suggesting how the total observed variability could be apportioned to different mechanistic noise sources [Swain *et al.*, 2002; Paulsson, 2004].

High-precision measurements of noise in gene expression coupled with the application of information-theoretic analyses have quantified the role of noise in gene regulation. In the simplest setup, we consider gene regulation as a control process that maps certain input transcription factor concentrations into the expression level of the regulated gene. Such input/output relations have a limited dynamic range due to biophysical constraints on gene expression (e.g, the maximal rate of transcriptional initiation), and given that they are noisy as experimentally demonstrated, their "power" to transmit information must be limited [Tkačik *et al.*, 2008]. Starting with precise measurements of noisy input/output relationships, it is possible to use Shannon's information theory to quantify the regulatory power by computing the information flow through transcriptional regulatory elements, or even ask about the capacity, i.e., the maximal achievable flow given the measured noise [Shannon and Weaver, 1949]. Information-theoretic measures of regulatory power are not arbitrary: they are well-founded, theoretically unique measures satisfying certain basic expectations, for example, that information from independent sources is additive. Furthermore, there have been promising indications that it is possible to use information theory not only to analyze data, but also to generate de-novo predictions about how biological networks should be wired together, by postulating that the networks have evolved to maximize information transmission given irreducible sources of noise in biological signaling [Tkačik *et al.*, 2008b;

Tkačik *et al.*, 2008]. Recent evidence for such signatures of optimality has been provided in the system of gap genes in fruit fly (*Drosophila melanogaster*) development [Tkačik *et al.*, 2015].

Two essential limitations to the application of information theory to biochemical signaling are addressed in this thesis:

- **Restriction to treatment of either the steady state of nonlinear networks, or dynamic linearized networks with Gaussian signals.** In terms of theory, biological signaling has been analyzed in the limit of temporal signals with jointly Gaussian statistics [Tostevin and ten Wolde, 2009]; or in the limit of steady state, where the input/output relations can be nonlinear and the distributions non-gaussian [Tkačik and Walczak, 2011]. For toy models of discrete systems, these restrictions have recently been partly lifted, allowing the calculation of information rate in a nonlinear noisy system [Barato *et al.*, 2013]. In terms of applications to data, however, the methods proposed to estimate the information between environmental signals and the resulting full gene activity trajectories have been limited to large number of samples [Selimkhanov *et al.*, 2014]. Here, we first develop and evaluate estimation methods for the mutual information: i) for biological reaction networks governed by the chemical master equation, we derive model based information estimates and ii) decoding-based estimators that lower bound the mutual information between a finite set of inputs and the single cell time series data. Then we apply the methods to single cell data: i) on previously published data of Erk and Ca^{2+} and ii) on nuclear translocation of transcription factors in yeast.

- **Restriction to theoretical models of biochemical processes in equilibrium.** Most models of cellular signaling to which information theory has been applied are considered to be in thermodynamic equilibrium. For example, gene regulation through binding of TFs to their regulatory sites is almost exclusively considered as an equilibrium process. In the field of biochemical signaling (in particular, detection of ligand concentrations), the equilibrium results of Berg and Purcell [Berg and Purcell, 1977; Bialek and Setayeshgar, 2005; Kaizu *et al.*, 2014] have recently been generalized to out-of-equilibrium situations, with interesting and

universal results about how the sensing precision of biochemical sensors can be enhanced by energy expenditure [Lang *et al.*, 2014]. For gene regulation, no such systematic generalization exits, although the first steps have been taken [Rieckh and Tkačik, 2014]. Here we focus on transcriptional regulation, where crosstalk –the possibility that non-cognate TFs could initiate transcription – has been neglected in previous models. We study how an out of equilibrium process could help reduce erroneous initiation, due to crosstalk, and the conditions when it is advantageous.

2 Background

2.1 Mathematical theory of communication

2.1.1 Introduction

When Shannon and Weaver first introduced their theory [Shannon and Weaver, 1949], they were referring specifically to human and electronic communication, and had to deal with three levels of communications problems:

- The technical problem: How accurately can the communication symbols be transmitted.
- The semantic problem: How precisely they convey the desired meaning.
- The effectiveness problem: How effectively meaning influence desirable functions.

They limited the theory to the first question and shortly discussed possible relations between the three problems.
Throughout its 80 years of existence, the theory was shown to be generally applicable to several areas of science. In particular, in this study, when we talk about single cells, we might not need to contemplate the three levels of communication. In fact, we can consider the semantic and effectiveness problems to be embedded in the internal circuitry that leads to behaviour. Thus, the technical approach becomes sufficient for the analysis of cellular communication in a broad sense.

In this study, communication has a particular sense, different from ordinary usage: it refers to the information flow between the source of information and the destination.

For instance, if the correlation between some extracellular signal (the source) and the behaviour of cells (the destination) is greater than zero, that implies the presence of information flow and hence communication between the environment and the cell. On the other hand, it should be pointed out that in this case communication does not refer to a single event, but rather it considers all possible input messages from the source, all possible outcomes and their frequency of occurrence. Overall, it's providing a macroscopic[1], general and statistical measure for communication systems, that is valid for the complete set of input messages and outcomes.

To clarify these ideas and frame the definitions in mathematical terms we first introduce the concept of information, then information transmission and lastly channel capacity.

2.1.2 Information and entropy (H)

The information provided by a source is intuitively equivalent to the amount of choice that it offers. If the source emits a discrete set of possible values, with probabilities: $p(u_1), p(u_2), ..., p(u_z)$, then the information of the source u is defined as:

$$H(u) = -k \sum_{i=1}^{z} p(u_i) \log_2 \Big(p(u_i) \Big),$$

where k is a positive number. Notice that the equation has the same form as the Boltzmann entropy for $k = k_B T$.

To avoid ambiguity, we will refer to this measure as entropy, but one should keep in mind that it alludes to a measure of information and choice offered by the source, as Shannon originally established it.

Before we continue, we will define the notation and parameters: let's set the scaling parameter $k = 1$ for simplicity, use the short notation $p(u) = p(u_i)$ inside the summation, which always runs over all possible values of u (unless specified), and use the logarithm base 2 ($\log_2$), so that the entropy is measured in bits and has an intuitive interpretation[2].

Two intuitive ideas that illustrate why the entropy is a sensible measure of information are:

[1] In the statistical physics sense.

[2] Such interpretation is further explained in section 2.1.3

- When all possible values of u have the same probability $\frac{1}{z}$, we expect that the entropy grows with the number of possible choices, therefore, any monotonic increasing function of z does the job. Our measure, becomes a logarithmic function of z.

$$H(u) = \log_2(z).$$

- Regarding the occurrence frequency of u: if only one message u^* shows up, we expect the information of the source to be zero (there is no choice offered by the source). In that case $p(u^*) = 1$ and then $H(u) = 0$. In fact, $H(u)$ is maximal when $p(u)$ is uniform.

As an example, the entropy of a binary source, where the probabilities of the two possible values of u are $p(u_1)$ and $p(u_2) = 1 - p(u_1)$, is plotted in Fig 2.1 as a function of $p(u_1)$.

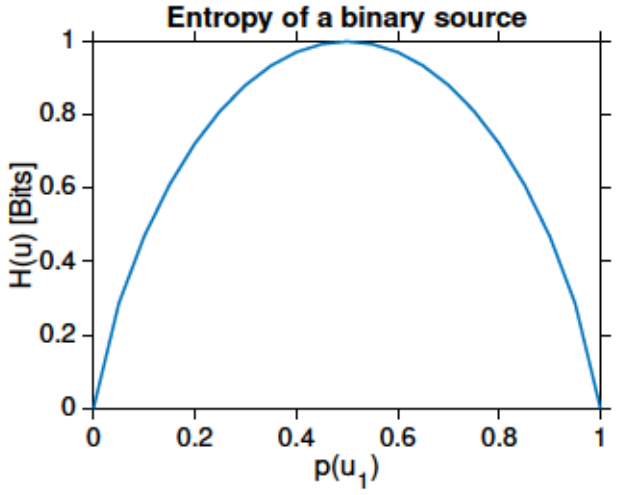

Figure 2.1: **Entropy of a binary source**, where the probabilities are $p(u_1)$ and $p(u_2) = 1 - p(u_1)$.

The entropy is maximal at the uniform distribution ($p(u_1) = p(u_2) = 0.5$), corresponding to the highest uncertainty.

The notion of entropy generalizes to continuous distributions and to functions of several variables, such as a set of source signals $\vec{u}$:

$$H(\vec{u}) = -\int_{\vec{u}} p(\vec{u}) \log_2(p(\vec{u}))d\vec{u}.$$

As appreciated for a long time [Cover and Thomas, 2005], there are technical challenges in defining entropy for continuous state spaces, much like issues with the ab-

solute offset for entropy in statistical physics. For example, as illustrated in [Tkačik and Walczak, 2011] with a Gaussian distribution, the entropy of a continuous variable depends on the choice of units: if the units change, the value of the entropy changes as well. Therefore, the (discrete) number of possibilities must depend on how finely u is measured; nominally, if u were known with arbitrary precision, the number of states would be infinite. However, we will find that the relevant quantity that measures information transmission is defined as a difference of entropies, solving the unit/offset problem. On the other hand if we specify the measurement precision and discretize u by binning, no practical problems arise [Cover and Thomas, 2005].

We have defined here the information of a source as the entropy or uncertainty of the source's distribution, but we are interested in how the information is transmitted from the source to the receiver; as Shannon initially formulated:

"The fundamental problem of communication is that of reproducing at one point either exactly or aproximately a message selected at an other point."

For that purpose, he proposed the mutual information or information transmission, which measures the information flow between two "points".

2.1.3 Mutual information (MI)

Let's consider a communication system that has several components: first the information source (u), then the transmitter (responsible for encoding the message), the channel, the receiver (responsible for decoding the signal) and at the end, the information recipient that receives the final signals (x) (see Fig 2.2).

To measure the information flow through the communication system, we consider the discrete random variables u (input signal) and x (output signal), with probability distributions $p(u)$ and $p(x)$ correspondingly. At the source, we know that the information is measured with the entropy $H(u)$. However, if the channel is noisy, in general, the information transmitted to the recipient will be lower. In fact, we can measure the uncertainty introduced by the noise, by measuring the conditional entropy of u knowing

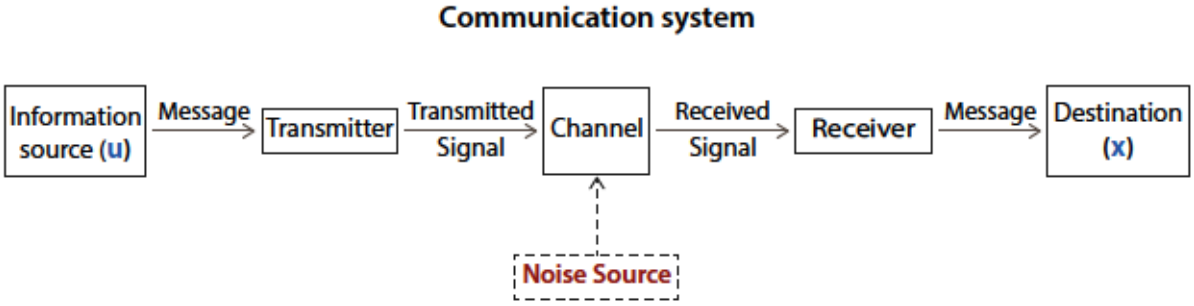

Figure 2.2: **Communication system**, based on the scheme proposed by Shannon in [Shannon and Weaver, 1949]

the output signal x, that is $H(u|x)$. Following this idea, the information transmitted is obtained by subtracting from the source entropy the average conditional entropy:

$$I(u;x) = H(u) - \langle H(u|x)\rangle_{p_x} \tag{2.1}$$

where $\langle . \rangle_{p_x}$ represent the average over the distribution $p(x)$, for a short notation we will ignore the brackets and refer to the average conditional entropy as $\langle H(u|x)\rangle_{p_x} = H(u|x)$.

If we write the conditional distributions in terms of the joint distribution $p(u,x)$ using $p(u,x) = p(u)p(x|u) = p(x)p(x|u)$, it can be shown that the information transmission is symmetric in u and x:

$$I(u;x) = \sum_{i,j} p(u_i, x_j) \log_2 \left(\frac{p(u_i, x_j)}{p(u_i)p(x_j)} \right), \tag{2.2}$$

hence, it is called "mutual information". Consequently, Eq 2.1 can likewise be written as:

$$I(u;x) = H(x) - \langle H(x|u)\rangle_{p_u}. \tag{2.3}$$

If the two variables are independent $P(u,x) = P(u)P(x)$, then the conditional entropy $H(u|x) = H(u)$, and no information is transmitted $I(u;x) = 0$. From this, we clearly see that when there is no statistical dependency there is no mutual information. Therefore, I is an appropriate measure of statistical dependency between two random variables. As such, I is free of assumptions about the nature of dependency, and encompasses both linear and nonlinear correlations. This quantity has convenient properties to serve as a proper measure of information transmission as enlisted in [Tkačik and Walczak, 2011].

- It is always a positive quantity, and zero when no information is transmitted.
- It is well defined for continuous or discrete quantities. The variables u and x can be either continuous, discrete or a mixture thereof.
- It is reparametrization invariant. Mutual information between u and x is the same as mutual information between any invertible transformation of the variables, that is $I(u;x) = I(f(u);h(x))$. This is very convenient in the context of biological data analysis, where where experiments often report, e.g. raw intensities or log-transformed intensities on the microarray chips or in fluorescence activated cell sorting.
- It obeys data processing inequality. Suppose that x depends on u and y depends on x (but not directly on u). In other words, one can imagine a Markov process, $u \to x \to y$, where arrows denote a noisy mapping from one value to the next. Then $I(u;x) \geq I(u;y)$, that is, information necessarily either gets lost or stays the same in the transmission process, but it is never spontaneously created.
- It has a clear interpretation and units. For the logarithm base 2 used in Eq 2.2 the mutual information is measured in bits (binary units). If there are I bits of mutual information between input c and output g, we can interpret that on average $2^{I(u,x)}$ distinguishable levels of x can be reached by modulating the values of u.

Continuous representation

The corresponding continuous version for mutual information is:

$$\hat{I}(u;x) = \int_x \int_u p(x,u) \log_2 \left(\frac{p(x,u)}{p(x)p(u)} \right) du dx, \tag{2.4}$$

which can also be expressed in terms of the entropy as in equations 2.1 and 2.3, where the corresponding continuous version of the average conditional entropy $H(x|u)$ is

$$\begin{aligned} H(x|u) &= \int_u H(x|u)p(u)du \\ &= -\int_u \int_x p(x,u) \log_2(p(x|u))dxdu. \end{aligned} \tag{2.5}$$

The $H(x)$ (similarly to $H(u)$), the differential entropy of x is defined as

$$H(x) = -\int_x p(x) \log_2(p(x))dx. \tag{2.6}$$

Channel capacity

The mutual information defined over the joint probability distribution $p(u, x)$ is not an intrinsic property of the channel. If we examine Eq 2.1, the first term ($H(u)$) is totally independent of it, while the second term ($H(x|u)$) measures the signal distortion due to the noise through the channel. To maximize the information transmission through a specific channel, we could think that a uniform probability distribution of the source would maximize $H(u)$ and therefore the mutual information. However, the uncertainty of x (introduced by the channel) for each value of u is not necessarily the same. Therefore, to find the maximum information transmission we need to find the probability distribution of inputs $p(u)$ that maximizes $I(u; x)$ through a specific channel.
In information theory this quantity is known as channel capacity $C(u; x)$ and is mathematically defined as

$$C = \max_{p(u)}(H(x) - H(x|u)). \tag{2.7}$$

In other words, the channel capacity is the mutual information maximized over all possible distributions of the signal which is an intrinsic property of the channel itself.
Finally, if the channel is noiseless then $H(x|u) = 0$. However the majority of real-world communication channels are noisy, thus the noisy channel coding theorem becomes relevant. This theorem states that if the entropy of the source is lower than the channel capacity, despite some degree of noise, with an adequate coding, the information can be transmitted with arbitrarily low error. For an information source with entropy greater than the channel capacity, no coding could transmit all the information that the source provides.
Therefore, the capacity is a very useful quantity, because it provides with a hard upper bound to how accurately data can be transmitted through a channel.

2.2 Genetic regulatory networks

Single-cell organisms respond to multiple environmental changes and cells in multicellular organisms communicate with each other to achieve collective responses. How fast cells can respond to external stimuli is limited by the speed at which internal mechanisms transmit and process the information. In fact, there is a wide range of response

timescales, from milliseconds in synaptic signaling to hours or days in embryonic development. This reveals the large variety of mechanisms used to transmit and process information and generate responses.
Cells have developed an elaborate machinery that allows them to communicate with the environment as well as each other and it involves nearly half of the protein abundances across organisms. In human (HeLa) cells around 40% of the proteome is involved in genetic and environmental information processing, while for *E. coli* the value exceeds 50% when cells grow in minimal medium [Milo and Philips, 2016]. In this section, we will provide a biological description of signaling and gene regulatory networks. These primary systems are in charge of transmitting and processing information that leads to cell response.

2.2.1 Cell signaling

The main communication channel used by cells to integrate environmental information is represented by signaling networks. These networks comprise sequences of chemical reactions and allosteric modifications that start at the cell surface. When an extracellular signal (ligand) binds typically to a transmembrane receptor, it changes its state from susceptible to active, which in turn activates one or more intracellular signaling pathways (see Fig 2.3). The pathways are composed of intracellular signaling proteins, which process the signal and distribute it to the appropriate targets. The targets are generally called effector proteins and implement the change in cell behaviour. Normally they are transcriptional regulators (also called transcription factors), ion channels, components of a metabolic pathway, or parts of the cytoskeleton.
The ligands bind to specific receptor proteins and in humans there are 1500 genes encoding different types of receptors. However, cells continuously develop mechanisms to improve specificity, e.g. the signaling complexes formed at activated receptors. Such complexes mainly recruit signaling molecules that will be part of the signaling pathway thus improving the accuracy of the signal delivery.

Through a series of downstream phosphorylation and dephosphorylation processes, the signaling proteins induce changes inside the cell: they can amplify signals when

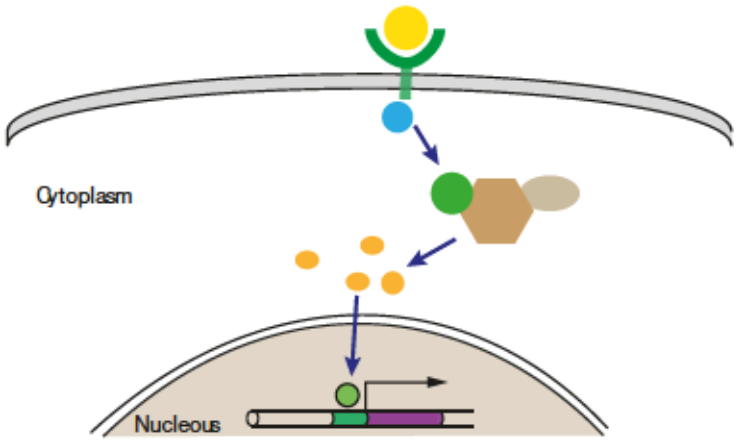

Figure 2.3: **Cell signaling.** The scheme represents the information flow from environmental signals, via the interaction of ligands (yellow) with the receptor (in green), activating intracellular signaling cascades in the cytoplasm and latter modifying gene regulation in the nucleus.

they are poor and frequently integrate certain extracellular signals for an appropriate response. For example, cell proliferation often depends on a combination of signals that promote cell division, survival, and growth [Alberts *et al.*, 2015].
For cell responses that require only covalent modifications of pre-existing downstream proteins, like the effector proteins interacting with the cytoskeleton, the speed of response is on the order of seconds to minutes, allowing a rapid spatial and temporal control of cell behaviour.
Ultimately, the most common destination of effector proteins is the nucleus, where they influence the activity of transcription factors. Consequently, these factors modify their binding properties to specific regulatory regions of the DNA leading to changes in gene regulatory networks which eventually are responsible for a behavioural response (see Fig 2.3).

Mathematical models used to describe cell signaling are various and diverse in complexity. The basic model for chemical kinetics are Rate Equations (REs), a set of ordinary differential equations, whose solution for a given set of initial concentrations report the concentration levels of the relevant molecules through time. The limitation of this modeling approach is the assumption that diffusion rates are close to infinity, and that the copy number of molecules is very large, so that concentrations can be treated

as continuous. However, that is not always the case in cell signaling. For instance, the budding yeast in the MAPK phosphatase pathway, contains approx. 40 copies of the Msg5 signaling protein while the most abundant proteins have up to 2×10^4 copies. The low copy numbers reveal the stochasticity of the process as mentioned in the introduction. For this reason, the appropriate formalism to describe signaling networks uses probability distributions instead of concentrations. One such method is the Chemical Master Equation (CME) [Van Kampen, 2007], while other methods, like the Reaction-Diffusion Master Equation or the Brownian Dynamics, use the spatial distribution of the molecules [Smith and Grima, 2018].

2.2.2 Gene regulation

The genetic code contains hereditary information that represents the blueprint for the organism as a whole. It includes certain species characteristics as well as particular traits inherited from its ancestors. The DNA molecule consists of two polynucleotide chains made up of 4 nucleotide subunits (bases). The bases attached to the sugar-phosphate backbone can alternate between A, C, G, T, corresponding to adenine, cytosine, guanine and thymine respectively. The cell can use the information encoded in the DNA through gene transcription and subsequent translation of protein coding genes. For transcription, the RNA polymerase can copy a section of the DNA into a single-stranded nucleotide RNA sequence. The chemical structure of the RNAs is similar to the DNA, with the difference that it contains uracil (U) instead of thymine, and the sugar that forms the backbone is ribose instead of deoxyribose.
A small fraction of RNAs are non-coding and are not translated into proteins since they have specific functions, like being part of the ribosomes. On the other hand, the majority of RNAs, callled messenger RNAs (mRNAs), are translated into proteins. The translation process is orchestrated by the ribosomes which make use of transporter RNAs (tRNAs) in order to match a nucleotide triplet on the mRNA (i.e. codon) with a corresponding amino acid, thus forming a polypeptide chain that will fold into a functional protein.

Multicellular organisms have a vast variety of cell types (200-400 in a healthy hu-

man). Nevertheless, every single cell shares the same genome. What is then different between cells types if their genetic code is identical? In humans, the DNA is about 10^9 nucleotides long, and it encodes $\sim 10^4$ proteins, but no single cell needs to synthesize all these proteins at once. Actually, only $30 - 60\%$ of coding genes are transcribed into RNAs in a typical human cell. Hence, an essential difference between cell types is the amount of expressed protein concentrations that they posses of each type, in particular, which genes are silenced (i.e., not transcribed at all).

The process of extracting information from the DNA to synthesize proteins is called gene expression, while the cellular processes controlling which genes are expressed is called *gene regulation*.

The central dogma proposed by Francis Crick, [Crick, 1970] (see Fig 2.4 A) argues

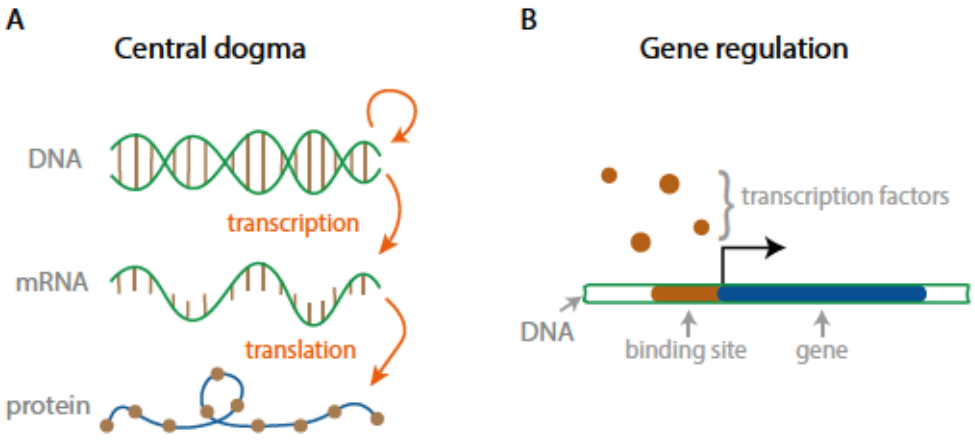

Figure 2.4: **Central dogma and gene regulation.** **A** The central dogma scheme as introduced in the original paper of Francis Crick. **B** A simplified representation of the elements involved in transcriptional regulation. The binding site is a short DNA sequence where the transcription factor binds and make it more or less likely that the RNA polymerase initiates transcription of the gene.

that the information flow to generate proteins follows the path from DNA to RNA to protein. In his original work, he suggests that in extraordinary circumstances the information might flow from DNA to proteins and from mRNA into DNA, but never from proteins to DNA, mRNA or back to proteins. Nevertheless, in 1961 Monod and Jacob [Jacob and Monod, 1961] , showed experimentally the existence of regulatory proteins that interacted with the DNA to induce or repress gene expression. Thus, the central dogma is restricted to the information flow regarding the constitutive elements of the

proteins, while the specific proteins called transcription factors (TFs) convey information to the DNA and activate the gene expression machinery. The TFs control which proteins should be synthesized, when are they required and how many of them are necessary.

In principle, regulation of gene expression can occur at several stages of protein synthesis: selecting when and which genes are transcribed by the RNA polymerase during transcription, choosing which mRNA molecules are translated by the ribosome during translation or after translation, controlling splicing and splice variants or the proper folding of the proteins and their degradation. Nonetheless, for most genes, transcriptional regulation is the most important aspect of regulation, and for most organisms, approximately 10% of the protein-coding genes correspond to transcription factor proteins.

Transcriptional regulation begins with transcription factors identifying specific sequences along the DNA-those sequences are called cis-regulatory regions or binding sites (see Fig 2.4 B). The recognition of the regulatory domains is possible mainly because of the complementary surface of the protein that matches the cis segment in the DNA. Additionally, TFs can operate as repressors or activators; activators promote transcription of a gene while repressors prevent the process. For example, in *E. coli*, the tryptophan repressor switches the gene off, while the cAMP binds to CRP, an activator that promotes the usage of alternative carbon sources when glucose is not available. But for eukaryotes the regulation process is much more complicated. Namely, compared to prokaryotes, the binding sites are shorter, the DNA is longer, there are several types of regulators, the transcription factor-binding site specificity is substantially decreased and the the binding affinities are also reduced. Typical binding sites are $\sim 10-15$ base pairs (bp) long for prokaryotes, and $\sim 5-10$ base pairs long for eukaryotes. The DNA of *E. coli* is in the order of 10^6 bp and in humans on the order of 10^9 bp. Therefore, it is more likely that a sequence of 6 bp appears randomly in a human genome. Similarly, a large variety of TFs in eukaryotes lead to lower specificity because of binding site similarities due to decreased length. Hence, transcriptional regulation in eukaryotes involves several transcription factors, the specific interactions between them and various cis-regulatory regions.

Turning back to the description of cell signaling and gene regulation we thus rec-

ognize that the processes can be complex and specific to each system, making the networks very diverse. At the same time, the formalism described at the beginning of this chapter is only concerned with the input and output probability distributions. Despite the details of the network the correlation between the two random variables can be quantified and measured in bits. Consequently, we can frame the function of biochemical networks in those terms. For cell signaling, the concentration of ligands can be considered the input while any observable change in cell behavior the output (e.g. concentration of TFs, cell growth, motion speed of motion, etc.). Similarly, for gene regulation, the concentration of TFs can be regarded as input while the mRNA or protein levels the output. The initial studies of biological networks following this framework started in the field of neuroscience during the early 50's. After the seminal papers of [MacKay and McCulloch, 1952] many studies have followed, both experimentally and theoretically. In other areas of biology, however, it's been only since the late 2000's that this framework became a popular formalism to study biological systems. We list relevant studies where mutual information has been used to investigate the properties and functions of biological systems in table 2.2.2. The relevant reviews are considered first, followed by the studies which experimentally measured information transmission in specific biological systems and eventually the theoretical approaches.

Information theory applied to biology		
Reference	**system**	**details**
Perspectives on theory at the interface of physics and biology [Bialek, 2018]	-	**(Review)** Provides prospects towards a unified theoretical physics of biological systems.
Information processing in living systems [Tkačik and Bialek, 2016]	-	**(Review)** Presents an exhaustive analysis of information theory-based studies, both: data-based and theory studies.
Environmental sensing, information transfer, and cellular decision-making [Bowsher and Swain, 2014]	-	**(Review)** Establishes the achievements and limitation of information theory applications into biology.
Biophysics: searching for principles [Bialek, 2012]	-	Book.
Information transmission in genetic regulatory networks: a review [Tkačik and Walczak, 2011]	-	**(Review)** Summarizes the applications of information theory to gene regulatory networks.
The application of information theory to biochemical signaling systems [Rhee *et al.*, 2012]	-	**(Review)** Describes in detail the bases of information theory and three successful application in biology.
How information theory handles cell signaling and uncertainty [Brennan *et al.*, 2012]	-	**(Review)** Motivates the use of information theory in cell signaling.
Information theory and neural coding [Borst and Theunissen, 1999]	neural system	**(Review)** Shows the variety of applications that information theory has in neuroscience.

Information flow and optimization in transcriptional regulation [Tkačik *et al.*, 2008]	*Drosophila* embryo	Estimates, for the first time, the mutual information in gene expression, where $\hat{I} = 1.5 \pm 0.15$ [bits] are found to be $\sim 90\%$ of the optimal theoretical value: $I_{opt} = 1.7$ [bits].
Estimating mutual information and multi-information in large networks [Slonim *et al.*, 2005]	yeast *Saccharomyces cerevisiae*	Estimates the information relations that characterize large networks and find more than 1 [bit] of mutual information between pairs of gene expression levels.
Information transduction capacity of noisy biochemical signaling networks [Cheong *et al.*, 2011]	*mouse* (Tumor Necrosis Factor)	Estimates $I = 1.8$ [bits] of mutual information, from a clusters of 14 cells, suggesting that collective cellular response can increase information transfer.
Positional information, in bits [Dubuis *et al.*, 2013b]	*Drosophila* embryo	Estimates $I = 4.14$ [bits] in the four gap genes, which define the position of a cell's position with $\sim 1\%$ error.
Robustness and compensation of information transmission of signaling pathways [Uda *et al.*, 2013]	PC12 Cells	Calculates ~ 1 [bit] of mutual information between growth factor and gene expression.
Limits on information transduction through amplitude and frequency regulation of transcription factor activity [Hansen and O'Shea, 2015]	yeast *Saccharomyces cerevisiae*	Finds that two genes can transduce more information than a single one and that the amplitude transmits more information than the frequency in the system.
Information transfer by leaky, heterogeneous, protein kinase signaling systems [Voliotis *et al.*, 2014]	Human (*HeLa cells*)	Estimates about 1 [bit] of information on the ERK signaling pathway, using the knn method.
Accurate information transmission through dynamic biochemical signaling networks [Selimkhanov *et al.*, 2014]	ERK, calcium and NF-κB	Measures ~ 1.5 [bits] of mutual information encoded in dynamical signals, using the knn method.
Information Transfer in Gonadotropin-releasing Hormone (GnRH) Signaling [Garner *et al.*, 2016; Garner *et al.*, 2017; Voliotis *et al.*, 2018]	LβT2 and HeLa cells	Finds that out of 3 bits of input entropy, less than 1 [bit] of information is transmitted through both systems, estimations used $> 10,000$ individual cells and the knn method.
Dynamic sampling and information encoding in biochemical networks [Potter *et al.*, 2017]	Fibroblast cells	Uses more than 10.000 cells to estimate ~ 1 [bit] of information encoded in ATP-induced calcium response, using the knn method.
Information processing in the NF-κB pathway [Tudelska *et al.*, 2017]	MEF cells *mouse*	Estimates less than 1 [bit] of information at several time points, using the knn method.
Distributed and dynamic intracellular organization of extracellular information [Granados *et al.*, 2018]	*Saccharomyces cerevisiae*	Estimates up to ~ 2.5 [bits] of information from dynamic responses of multiple transcription factors.
Information capacity of genetic regulatory elements [Tkačik *et al.*, 2008a]	General method with examples on *yeast* and *Drosophila*	Calculates the channel capacity of simple gene regulation elements. Finding that for realistic noise levels more than 1 bit of information should be achievable.
Mutual information between in- and output trajectories of biochemical networks [Tostevin and ten Wolde, 2009]	*E. coli*	Assumes a Gaussian model to estimate the MI between temporal signals, with the application to the chemotaxis network of bacteria.
Optimizing information flow in small genetic networks. I, II, III, IV [Tkačik *et al.*, 2009a; Walczak *et al.*, 2010b; Tkačik *et al.*, 2012; Sokolowski and Tkačik, 2015a]	Develops general methods with some applications on *Drosophila*	A series of optimal information transmission studies in small genetic networks, considering: a single transcription factor controlling one or more genes, the role of feed forward interactions between genes, a single, self-interacting gene and spatially coupled gene regulatory networks at the steady state.

Identifying sources of variation and the flow of information in biochemical networks [Bowsher and Swain, 2012]	General theory with application in yeast	Introduces a variance decomposition method and shows that 80% of the response in the osmosensing system in yeast is induced by the input stress.
Time-dependent information transmission in a model regulatory circuit [Mancini *et al*., 2013]	–	Studies the architecture of biochemical two-state model networks that maximize the information transmission, when the time dependent response is delayed respect to the dynamical input.
Efficiency of cellular information processing [Barato *et al*., 2014]	-	Proposes an entropic rate that is bounded by the thermodynamic entropy production and characterizes how much the internal process learns about the external process.
Optimal prediction by cellular signaling networks [Becker *et al*., 2015]	*E. coli*	Studies how accurately linear signaling networks can predict future signals.
The Impact of Different Sources of Fluctuations on Mutual Information in Biochemical Networks [Chevalier *et al*., 2015]	synthetic circuit	Suggests that single cell circuits can transmit higher information than the computed by the population. Where cell to cell variability can lower the information transmission estimations.
Information processing by simple molecular motifs and susceptibility to noise [Mc Mahon *et al*., 2015]	-	Quantifies how extrinsic and intrinsic noise affects the transmission of simple signals along simple motifs of molecular interaction networks via estimating information transmission.
Thermodynamics of computational copying in biochemical systems [Ouldridge *et al*., 2017]	-	Studies the optimality and efficiency of a canonical biochemical readout network.
Multidimensional biochemical information processing of dynamical patterns [Hasegawa, 2018]	-	Studies how biochemical systems can process multidimensional information embedded in dynamical patterns.
Statistics of optimal information flow in ensembles of regulatory motifs [Crisanti *et al*., 2018]	-	Computes the statistics of the maximal mutual information transmitted in an ensemble of regulatory motifs.
Information content of downwelling skylight for non-imaging visual systems [Thiermann *et al*., 2018]	Opsin	Quantifies circalunar and circadian regularities in the spectrum of downwelling radiance salient to non-imaging opsins.

3 Estimating information in time-varying signals

The work presented in this chapter was conducted jointly with and Gašper Tkačik and Jakob Russ, it corresponds to the theoretical section of the paper submitted for publication to PLOS Biology (see pre-print in [Cepeda-Humerez *et al.*, 2019]) and is reproduced here with minimal changes.

Abstract

Across diverse biological systems—ranging from neural networks to intracellular signaling and genetic regulatory networks—the information about changes in the environment is frequently encoded in the full temporal dynamics of the network nodes. A pressing data-analysis challenge has thus been to efficiently estimate the amount of information that these dynamics convey from experimental data. Here we develop and evaluate decoding-based estimation methods to lower bound the mutual information about a finite set of inputs, encoded in single-cell high-dimensional time series data. For biological reaction networks governed by the chemical Master equation, we derive model-based information approximations and analytical upper bounds, against which we benchmark our proposed model-free decoding estimators. In contrast to the frequently-used k-nearest-neighbor estimator, decoding-based estimators robustly extract a large fraction of the available information from high-dimensional trajectories with a realistic number of data samples. We apply these estimators to previously published data on Erk and Ca^{2+} signaling in mammalian cells and to yeast stress-response, and find that substantial amount of information about environmen-

tal state can be encoded by non-trivial response statistics even in stationary signals. We argue that these single-cell, decoding-based information estimates, rather than the commonly-used tests for significant differences between selected population response statistics, provide a proper and unbiased measure for the performance of biological signaling networks.

3.1 Introduction

For their survival, reproduction, and differentiation, cells depend on their ability to respond and adapt to continually changing environmental conditions. Environmental information must be sensed and often transduced to the nucleus, where an appropriate response is initiated, usually by selectively up- or down-regulating the expression levels of target genes. This information flow is mediated by biochemical reaction networks, in which concentrations of various signaling molecules encode for different environmental states or different response programs. This map between environmental input or response output and the internal chemical state is, however, highly stochastic, because typical networks operate with small absolute copy numbers of signaling molecules [Eldar and Elowitz, 2010]; moreover, different environments can be encoded by the same signaling molecule, by differentially regulating the dynamics of its concentration [Purvis and Lahav, 2013]. This raises two fundamental questions: first, how much information the cells could, even in principle, encode in the combinatorial and possibly time-varying concentrations of multiple signaling molecules and how such information could be plausibly read out during "downstream" processing; and second, how can we quantify, in an unbiased and model-free fashion, the amount of information available to the cells from limited experimental data.

Information theory provides a framework within which the theoretical study of limits to communication as well as the empirical study of actual information flows can be addressed [Shannon and Weaver, 1949]. Applications of information theory to questions in biology and, in particular, neuroscience started already in the 1950s and continue to this day, with the main focus to understand how—and with what accuracy—neural activity encodes information about the environment [Paninski, 2003; Strong *et al.*, 1998; Quiroga and Panzeri, 2009]. Applications of analogous techniques to biochemical sig-

naling only started recently and represent an active area of research at the interface of physics, biology, statistics, and engineering [Bowsher and Swain, 2014; Bialek, 2012; Tkačik and Bialek, 2016; Tkačik and Walczak, 2011].

Recent theoretical work analyzed the reliability of information transmission through specific reaction systems in the presence of molecular noise, e.g., during ligand binding [Thomas and Eckford, 2016], in chemotaxis [Tostevin and ten Wolde, 2009], gene regulation [Tkačik *et al.*, 2008a; Sokolowski and Tkačik, 2015b; Sokolowski *et al.*, 2016; Tkacik *et al.*, 2012; Walczak *et al.*, 2010a; Tkačik *et al.*, 2009b; Rieckh and Tkačik, 2014], biochemical signaling networks [Cheong *et al.*, 2011], etc., and asked how such transmission can be maximized by tuning the reaction rates. Generally, these studies focused on steady state, by considering the information encoded in a single temporal snapshot of the reaction network at equilibrium given the input signals. Rigorous extensions to dynamical signals have been either rare and only possible for simple cases, like the BIND channel [Thomas and Eckford, 2016], or required specific operating regimes that permitted linearization and Gaussianity assumptions [Tostevin and ten Wolde, 2009; Tostevin and Ten Wolde, 2010; de Ronde *et al.*, 2011]. At its core, the analysis of signal transduction through nonlinear noisy chemical systems requires one to have control over the distribution of concentration trajectories given the (possibly) time-varying inputs; even if it were possible to calculate this distribution in principle, the curse of dimensionality puts strong limits to the manipulations required to compute the information transmission. Consequently, problems of this kind are currently considered intractable in their full generality.

Empirical estimates of information transmission in biochemical networks similarly focused on the steady state [Dubuis *et al.*, 2013b; Voliotis *et al.*, 2014], or considered only specific, hand-picked dynamical features, such as the amplitude or the frequency of the response, as information carriers [Hansen and O'Shea, 2015]. Recent developments of fluorescent reporters and microfluidics have enabled the characterization of dynamical responses at a single cell resolution using large ($> 10^4$) numbers of sampled response trajectories, thereby permitting direct information estimates using generic estimators like the k-nearest-neighbors (knn) [Selimkhanov *et al.*, 2014]. Existing approaches, however, suffer from severe limitations: they still require a prohibitive number of samples, especially when the response is distributed over multiple

chemical species; or they necessitate uncontrolled assumptions about trajectory features that are supposed to be "relevant". We recently proposed and applied decoding-based information estimators [Granados *et al.*, 2018] as an alternative that draws on the past experiences in neuroscience [Borst and Theunissen, 1999; Marre *et al.*, 2015; Rieke *et al.*, 1993] to dissect the yeast stress-response network. In this study we provide a detailed account of the new methodology, show that it alleviates the most pressing problems of existing approaches, and benchmark it against synthetic and real data.

3.2 Models and Methods

3.2.1 Biochemical reaction networks

At their core, cellular processes consist of networks of chemical reactions. A chemical reaction network consists of a set of m molecular species $\{\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_m\}$ that interact through K coupled reactions of the form:

$$\nu'_{1k}\tilde{X}_1 + \ldots + \nu'_{mk}\tilde{X}_m \xrightarrow{\theta_k} \nu''_{1k}\tilde{X}_1 + \ldots + \nu''_{mk}\tilde{X}_m, \quad k = 1, \ldots, K \tag{3.1}$$

where $\nu'_{1k}, \ldots, \nu'_{mk}$ and $\nu''_{1k}, \ldots, \nu''_{mk}$ are coefficients that determine how many molecules of each species are consumed and produced in the k-th reaction. $\theta_1 \ldots \theta_k \in \mathbb{R}^+$ determine the rates at which the reactions occur and depend on binding affinities of chemical species, temperature and possibly the external conditions.

If we assume that the system is well-stirred, in thermal equilibrium and the reaction volume is constant, it can be shown that the probability that a reaction of type k takes place in an infinitesimal time interval $[t, t+dt]$ can be written as $a_k(\tilde{x}, \theta_k)dt = \theta_k g_k(\tilde{x})dt$, where $\tilde{x} = [\tilde{x}_1, \ldots, \tilde{x}_m]^T \in \mathbb{N}_0^m$ contains the amounts of molecules of the m species that are present in the system at time t, and $g_k(\tilde{x}) = \prod_{i=1}^{m} \binom{\tilde{x}_i}{\nu'_{ik}}$ counts all possibilities of choosing the required reaction molecules out of all available molecules [Gillespie, 1992; Van Kampen, 2007]. θ_k is a constant that depends on the physical characteristics of the cell but also on the environmental conditions.

Let us denote the probability that $\tilde{x}$ molecules of the m species are present in the system at time $t \in \mathbb{R}^+$ by $p(\tilde{x}, t)$ and define the stoichiometric change vectors $\nu_k =$

$[\nu_{1k}, \ldots, \nu_{mk}]^T \in \mathbb{Z}^m,\ k = 1, \ldots, K$, as the net changes in the amount of molecules in the reactions, i.e. $\nu_{ik} = \nu''_{ik} - \nu'_{ik},\ i = 1, \ldots, m,\ k = 1, \ldots, K$. Then it can be shown [Van Kampen, 2007] that the chemical master equation (CME) can be written as:

$$\dot{p}(\tilde{x}, t) = -p(\tilde{x}, t) \sum_{k=1}^{K} a_k(\tilde{x}, \theta_k) + \sum_{k=1}^{K} p(\tilde{x} - \nu_k, t) a_k(\tilde{x} - \nu_k, \theta_k), \tag{3.2}$$

or in a more compact form [Van Kampen, 2007]

$$\dot{\mathbf{p}}(t) = \mathbf{M}\mathbf{p}(t), \tag{3.3}$$

where $\mathbf{p}(t)$ is a vector with components $p(\tilde{x}, t)$, which is, in principle, infinite dimensional, and $\mathbf{M}$ contains the transition rates between all possible states, e.g. the transition rate from state $\tilde{x}'_k = \tilde{x} - \nu_k$ to state $\tilde{x}$ is given by

$$M_{\tilde{x}, \tilde{x}'_k} = a_k(\tilde{x}'_k, \theta_k) - \delta_{\tilde{x}, \tilde{x}'_k} \sum_q a_q(\tilde{x}, \theta_q), \tag{3.4}$$

where δ is the Kronecker delta, which is 1 when $\tilde{x} = \tilde{x}'_k$ and 0 otherwise.

The CME given in Eq (3.3) is an instance of a continuous-time discrete-state-space Markov Chain for a random process X that can be solved exactly only for a few simple cases. It is nevertheless possible to efficiently generate samples x of the random process X, which we will refer to as "trajectories" or "paths", for a selected time interval, $t \in [0, T]$, according to the correct probability distribution p, by the Stochastic Simulation Algorithm (SSA, or the Gillespie algorithm) [Gillespie, 1977].

To study information transmission through the biochemical networks described by the CME, we need to define the input and output signals. In the simplest setup considered here, the input U is a discrete random variable that can take on one of the $q \geq 2$ possible values, $U \in \{u^{(1)}, u^{(2)}, \ldots, u^{(q)}\}$. Each input in general corresponds to a distinct set of reaction rate constants θ, but in models of real biological networks, changing input often modulates only one or a few rates in the system, e.g., by representing the change in a key external ligand concentration, receptor activity, etc. Changes in the input are reflected in changes in the resulting trajectories of chemical species amounts, x. Typically, only a subset of chemical species could be considered as biologically-relevant "outputs" that encode the information about the environmental change: this would correspond to marginalizing p in Eq (3.3) over the unobserved (non-output)

chemical species for the purposes of information transmission. While this is an interesting theoretical problem in its own right, here we work with simple toy examples where the output will be the trajectory, x, over the complete state space, i.e., we assume that all chemical species in the reaction network can be fully and perfectly observed. As we explain below, this allows us to define and compute the mutual information between a discrete input, U, and the output random process X given by the CME in a straightforward fashion. We later show that this computation can be carried out also when the continuous-time process X is sampled at uniform discrete times, as would be the case with experimental measurements.

3.2.2 Mutual information between discrete inputs and response trajectories

Information theory introduces the mutual information as the measure of fidelity by which changes in one random variable, e.g., the input U, can effect changes in another random variable, e.g., X. In this sense, mutual information is simply a measure of statistical dependency (i.e., any correlation, be it linear or not) between U and X, and can thus be written as a functional of the joint probability density function $p(x,u)$:

$$I(X;U) = \int_X \int_U p(x,u) \log_2 \left(\frac{p(x,u)}{p_X(x) p_U(u)} \right) \, du \, dx \tag{3.5}$$

where p_U and p_X are the marginal density functions for U and X, respectively, and we have generically written u and x as continuous variables; if they are discrete, integral signs are replaced by summations over the support for the corresponding probability distributions, as appropriate.

Mutual information is a non-negative symmetric quantity that is measured in bits, and is zero only if X and U are statistically independent. When studying information transmission through a channel $U \to X$ specified by $p(x|u)$, for which U serve as inputs drawn from an input distribution $p_U(u)$, it is common to rewrite Eq. (3.5) as

$$I(X;U) = H(U) - H(U|X) = H(X) - H(X|U), \tag{3.6}$$

where $H(X)$ is the differential entropy of X (and analogously for $H(U)$), defined as

$$H(X) = -\int_X p_X(x) \log_2 p_X(x) \, dx, \tag{3.7}$$

and the conditional entropy, $H(X|U)$, is

$$H(X|U) = \int_U H(X|u)p_U(u)\,du = -\int_U \int_X p_U(u)p(x|u)\log_2 p(x|u)\,dx\,du. \tag{3.8}$$

Equations (3.6) can be interpreted in two ways: information is either the difference between the total variability in the repertoire of responses X that the biochemical network can generate (measured by the *response entropy*, $H(X)$) and the average variability at fixed input that is due to noise in the network (measured by the *noise entropy*, $H(X|U)$); alternatively, information is also the entropy of the inputs, $H(U)$, minus equivocation $H(U|X)$, or the average uncertainty in what input was sent given that a particular response was observed. These interpretations make explicit the dependence of information both on the properties of the channel (the biochemical reaction network), as well as on the distribution of signals p_U that the network receives. In this work, we will consider discrete inputs and will assume uniform p_U. It is, however, also possible to compute the *channel capacity* C by maximizing the information flow at given $p(x|u)$ over all possible input distributions,

$$C = \max_{p_U} I(X;U); \tag{3.9}$$

Shannon's classic work then proves that error-free transmission at rates higher than those given by capacity is impossible, while error-free transmission at rates below capacity can be achieved with the optimal use of the channel. Contrary to engineering, where the focus is on finding encoding and decoding schemes that best utilize a given channel, in biophysics and systems biology mutual information is used as a tool to quantify the limits to biological signal processing due to noise without needing to make assumptions about possible biochemical encoding and decoding mechanisms.

The setup we consider here is one in which inputs U are iid drawn from a uniform distribution and change rarely, i.e., at a rate that is much lower than the (inverse) timescale on which the reaction network in Eq (3.1) relaxes to its steady state. We assume that after an input change, we observe a fixed-time segment of the complete network dynamics, x, which is a sample path in m-dimensional discrete space, making direct calculation of information, $I(X;U)$, by integrating / summing over all possible trajectories as implied by Eq (3.5) intractable. We will nevertheless show that estimates of exact information are possible if the reaction network is known, by explicitly using

the transition matrix $\mathbf{M}$ of the Markov Chain from Eq (3.3) and generating exact sample paths, that is, realizations of X, using SSA. We call this model-based approach *exact Monte Carlo approximation* and contrast it to uncontrolled model-free estimations such as those obtained by using Gaussian approximations or k-nearest-neighbors methodology. We then introduce various decoding estimators and establish a hierarchy through which these estimates upper and lower-bound the true information, as shown in Fig 3.1.

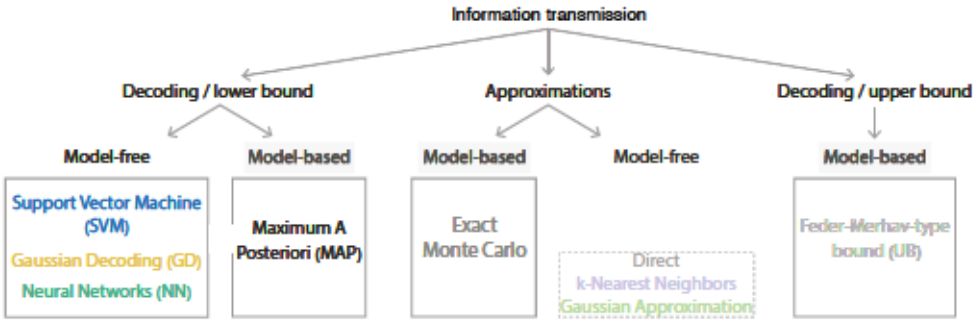

Figure 3.1: **Information transmission between discrete inputs and response trajectories in biochemical networks.** For fully-observed reaction networks whose dynamics are governed by a known chemical Master equation, information can be approximated to an arbitrary accuracy via Monte Carlo integration for either continuous-time or discrete-time response trajectories (model-based *exact Monte Carlo*, Section 3.2.3). Since full knowledge of the reaction system is used, these approximations are tractable deep in the regimes where model-free estimations break down with uncontrolled errors (Section 3.2.4). True information estimates are lower-bounded by model-based maximum a posteriori (MAP) or Bayes optimal decoding (Section 3.2.5). This decoding gives the lowest average probability of error and the corresponding information lower bound can be used as a benchmark for information estimates derived from other model-free decoding approaches (that have at least the error probability of the MAP decoder); in Section 3.2.6 we compare Support Vector Machine (SVM), Gaussian Decoding (GD) and Neural Network (NN) decoding approaches. Upper bounds like the Feder-Merhav bound [Feder and Merhav, 1994] and our improvement on it [Hledík *et al.*, 2018] complete the picture by estimating the gap between optimal decoding-derived and exact information values (Section 3.2.5).

3.2.3 Exact information calculations for fully observed reaction networks

Responses in continuous time. Given the specification of the biochemical reaction network in Eq (3.1), we sample N trajectories, x, using the Gillespie (SSA) algorithm. Each trajectory x can be represented as the sequence of consecutive states representing molecular species counts, $\mathbf{s} = [s_1, s_2, \ldots, s_r]$, where $s_1 = \tilde{x}(t = 0)$, etc., and $s_r = \tilde{x}(t = \sum_{i=1}^{r-1} t_i))$, and the sequence of time intervals spent in each state, $\mathbf{t} = [t_1, \ldots, t_r],\ 0 < t_i < T,\ i = 1, \ldots, r$ and $T = \sum_i^r t_i$. Then the likelihood of x for a given input u is:

$$p(x|u) = p(s_1) \exp(M_{s_1 s_1} t_1) \prod_{i=2}^{r} M_{s_i s_{i-1}} \exp(M_{s_i s_i} t_i) \tag{3.10}$$

where $p(s_1)$ is given by the initial conditions of the process, and the transition matrix $\mathbf{M}$ depends on the input u. To get the marginal distribution, $p_X(x)$, we sum over all possible input values:

$$p_X(x) = \sum_{i=1}^{q} p(x|u^{(i)}) p_U(u^{(i)}). \tag{3.11}$$

Since we are able to compute the exact likelihood for each path generated by the stochastic process, entropic quantities can be approximated without significant biases using Monte Carlo integration, where the integral over states in Eq (3.7) is replaced by an average over N sampled trajectories:

$$\tilde{H}(X) = -\frac{1}{N} \sum_{i=1}^{N} \log_2 p_X(x_i). \tag{3.12}$$

Similarly, we can approximate $H(X|U)$:

$$\tilde{H}(X|U) = -\sum_{j=1}^{q} p_U(u^{(j)}) \left(\frac{1}{N} \sum_{i=1}^{N} \log_2 p(x_i|u^{(j)}) \right). \tag{3.13}$$

The exact Monte Carlo information approximation is finally obtained using Eq (3.6):

$$I^*_{\text{exact}} = \tilde{H}(X) - \tilde{H}(X|U), \tag{3.14}$$

where $*$ reminds us that the paths are represented in continuous time.

Responses in discrete time. We can resample the continuous trajectories X on a grid of uniformly spaced time points to obtain a new discrete random variable, $\mathbf{X} =$

$[X(t = \Delta t), \ldots, X(t = i\Delta t) \ldots, X(t = d\Delta t)] \in \mathbb{N}_0^{m \times d}$, where Δt is the discretization step, $d = T/\Delta t$ is the length of $\mathbf{X}$. For convenient notation we denote this random variable as $\mathbf{X} = [X^0, \ldots, X^d]$, and its realizations, the discrete trajectories, as $\mathbf{x}$.

In the discrete case, the likelihood of $\mathbf{x}$ for a given input u can be computed using the general solution to Eq (3.3):

$$\mathbf{p}(t) = e^{\mathbf{M}t}\mathbf{p}(0), \tag{3.15}$$

where $\mathbf{p}(t)$ is the probability distribution of states after time t, with the initial probability distribution $\mathbf{p}(t = 0) = \mathbf{p}(0)$. Using this formal solution we compute the transition matrix between discrete timesteps separated by Δt to get:

$$\mathbf{W} = e^{\mathbf{M}\Delta t}, \tag{3.16}$$

where $\mathbf{M}$ and thus $\mathbf{W}$ again depend on u. The likelihood of any discrete path can then be obtained by multiplying the transition probabilities between all the d consecutive states in the path for a given input u:

$$p(\mathbf{x}|u) = p(\mathbf{x}^0) \prod_{i=2}^{d} \mathbf{W}_{\mathbf{x}^i \mathbf{x}^{i-1}}. \tag{3.17}$$

We can now approximate the information between input U and a discretely sampled trajectory $\mathbf{X}$, I_{exact}, as in the continuous case: we get the marginal $p_X(\mathbf{x})$ with Eq (3.11) and use Eqs (3.12, 3.13) in Eq (3.14). In general, temporal discretization loses information relative to the full (continuous-time) trajectory, where reaction events in the trajectory x are recorded with infinite temporal precision, so the information in discretely-sampled trajectories, I, must be bounded from above by the information in continuous-time trajectories, I^*:

$$I_{\text{exact}}(\mathbf{X}; U) \leq I^*_{\text{exact}}(X; U), \tag{3.18}$$

where equality is approached in the limit of ever finer temporal discretization, $\Delta t \to 0$.

3.2.4 Model-free information estimators

In the absence of a full stochastic model for the biochemical reaction network, mutual information estimation is tractable only if we make assumptions about the distribution of response trajectories given the input. We briefly summarize two approaches below:

in the first, k-nearest-neighbor procedure, the space in which the response trajectories are embedded is assumed to have a particular metric; in the second, Gaussian approximation, we assume a particularly tractable functional form for the channel, $p(\mathbf{x}|u)$.

K-nearest-neighbors (knn) estimator. The idea of using the nearest neighbour statistics to estimates entropies is at least 70 years old [Dobrushin, 1958; Vasicek, 1973], while estimators for mutual information have been developed during the early 2000s [Kaiser and Schreiber, 2002; Kraskov *et al.*, 2004]. The cornerstone of the approach is to compute the estimate from the distances of d-dimensional real valued data points to their k-th nearest neighbour. Hence, the estimator depends on the metric chosen to define this distance. Furthermore, its performance is known to depend on the value of k (number of nearest neighbours), where small k increase the variance and decrease the bias [Khan *et al.*, 2007]. This method has been used in several studies that estimated mutual information from single cell time series [Potter *et al.*, 2017; Selimkhanov *et al.*, 2014; Voliotis *et al.*, 2014]. These studies used large numbers of response trajectories to provide the first evidence that the information available from the full timeseries of the response could be substantially higher than the information available from any response snapshot.

Gaussian approximation. A simplifying assumption in the Gaussian approximation is that the distribution of trajectories sampled at discrete times given input is approximately Gaussian, with the mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ that may both depend on the input, u:

$$p(\mathbf{x}|u) = \mathcal{N}(\mathbf{x}; \mu(u), \Sigma(u)) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right). \tag{3.19}$$

The entropy of the multivariate distribution in Eq (3.19) has an analytical expression that only depends on Σ:

$$H_{\mathrm{G}}(\mathbf{X}|u) = \frac{1}{2}\log(\det(2\pi e\Sigma(u))), \tag{3.20}$$

which can be averaged over $p_U(u)$ to get the conditional entropy, $H_G(\mathbf{X}|U)$. To estimate the information, we further need $H(\mathbf{X})$ from Eq (3.6). This entropy of a Gaussian mixture has no closed form solution, but can be computed by Monte Carlo integration as

in the previous section, following discrete analogs of Eqs (3.11,3.12): we draw random samples from each of the q different multivariate Gaussian distributions, Eq (3.19), one for each possible input u, and assign the marginal probabilities to each sample x as

$$p_X(\mathrm{x}) = \sum_{i=1}^{q} p_U(u^{(i)}) \mathcal{N}(\mathrm{x}; \mu(u^{(i)}), \Sigma(u^{(i)})), \tag{3.21}$$

permitting us to use Eq (3.12) to approximate the total entropy of output trajectories in the Gaussian approximation, $H_G(\mathbf{X})$, and thus to obtain the Gaussian estimate for the information, $I_{\mathrm{G}}(\mathbf{X};U) = H_{\mathrm{G}}(\mathbf{X}) - H_{\mathrm{G}}(\mathbf{X}|U)$.

To apply this estimator, one must use real (or simulated) data to estimate the conditional mean, $\mu(u)$, and conditional covariance, $\Sigma(u)$ for every possible u, from a limited number of samples. While general caveats for such estimations have been detailed in many textbooks [Anderson *et al.*, 1958], we emphasize that information estimation is particularly sensitive due to the computation of the determinant in Eq (3.20) which can easily lead to ill-posed numerics when the number of samples is small. This can be mitigated by various regularization methods (one of which, the diagonal regularization, we demonstrate later) that impose a prior structure on the estimated covariance. Yet even in the case of significant oversampling that we can explore using simulated data, the Gaussian approximation introduced here—in contrast to Gaussian decoding estimator introduced in the next section—can provide information values that deviate significantly from the true value and are not guaranteed to bound the true value from either above or below. This is because the true solutions of the CME live in the positive quadrant of the discrete space, and are thus essentially different from the Gaussian distributions assumed here. We nevertheless present this estimator because (i) it forms the basis for the Gaussian decoding estimator, introduced below, and (ii) real data itself often deviates from stochastic trajectories sampled from the CME in that it is continuous (since we measure, e.g., fluorescence proxy for a concentration of a protein of interest) and contains extra noise, making Gaussian approximation potentially applicable.

3.2.5 Decoding-based information bounds

Here and in the next section we introduce a class of decoding-based calculations that lower-bound the exact information, $I(\mathbf{X};U)$, and can tractably be used as information

estimators over realistically-sized data sets. Let $\mathcal{D}$ consist of a set of N labeled paths, typically represented in discretely sampled time, $\mathcal{D} = \{(u_1, \mathbf{x}_1), (u_2, \mathbf{x}_2), \ldots, (u_\mathrm{N}, \mathbf{x}_\mathrm{N})\}$, where u_i and $\mathbf{x_i}$, for $i = 1, \ldots, N$, are realizations of the random variables $U \in \{u^{(1)}, \ldots, u^{(q)}\}$ and $\mathbf{X} \in \mathbb{R}^{m \times d}$, respectively. Here, $\mathcal{D}$ can represent either real data (typically containing $N \sim 10^2 - 10^3$ trajectories) in case of model-free information estimates, or trajectories generated by exact simulation algorithms (in which case the sample size, N, is not limiting) from the full specification of the biochemical reaction network in case of model-based approximations.

The procedure of estimating the input $\hat{u}$ from $\mathbf{x}$, such that the estimated $\hat{u}$ is "as close as possible" to true u for a given trajectory $\mathbf{x}$, is known as decoding in information theory and neuroscience, and can equivalently be viewed as a classification task in machine learning or as an inference task in statistics. This procedure is implemented by a decoding function,

$$\hat{u} = F_\omega(\mathbf{x}); \tag{3.22}$$

F is typically parametrized by parameters ω that need to be learned from data for model-free approaches, or derived from biochemical reaction network specification in case of model-based approaches. F assigns to every $\mathbf{x}_i$ in the dataset a corresponding "decode" $\hat{u}_i$ from the same space over which the random variable U is defined; formally, these decodes are instances of a new random variable $\hat{U}$. The key idea of using decoding for information estimation starts with the observation that random variables

$$U \to X \xrightarrow{\mathrm{T_d}} \mathbf{X} \xrightarrow{F_\omega} \hat{U}, \tag{3.23}$$

where T_d represents time discretization, form a Markov chain. In other words, the distribution of $\hat{U}$ is conditionally independent of U and only depends on $\mathbf{X}$, $p(\hat{u}|\mathbf{x}, u) = p(\hat{u}|\mathbf{x})$, and so

$$p(\hat{u}, \mathbf{x}, u) = p_U(u)p(\mathbf{x}|u)p(\hat{u}|\mathbf{x}). \tag{3.24}$$

The data processing inequality [Cover and Thomas, 2005] can be used to further extend the bounds in Eq (3.18):

$$I(U; \hat{U}) \leq I_{\mathrm{exact}}(U; \mathbf{X}) \leq I^*_{\mathrm{exact}}(U; X), \tag{3.25}$$

where equality between the fist two terms holds only if $I(U; \mathbf{X}|\hat{U}) = 0$. Consequently, $I(U; \hat{U})$ is a lower bound to the information between trajectories X and the input

U [Brunel and Nadal, 1998]. Note that analogous reasoning holds for decoding directly from continuous-time trajectories X. Better decoders which increase the correspondence between the true inputs and the corresponding decoded inputs will typically provide a tighter lower bound on the information.

To compute the information lower bound, we apply the decoding function to each trajectory in $\mathcal{D}$ in model-based approximations or to each trajectory in the testing dataset for model-free estimators that need to be learned over training data first. We subsequently construct a $q \times q$ confusion matrix, also known as an error matrix, where each element ϵ_{ij} counts the fraction of realizations of x generated by an input $u = u^{(i)}$ that decode into $\hat{u} = u^{(j)}$. This matrix provides an empirical estimate of the probability distribution $p(\hat{u}, u)$, which can thus be used to compute the information estimate:

$$I(\hat{U};U) = \sum_{u,\hat{u}} p(\hat{u},u)\log_2 \frac{p(\hat{u},u)}{p_U(u)p_{\hat{U}}(\hat{u})} \approx \sum_{i=1}^{q}\sum_{j=1}^{q} \epsilon_{ij}\log_2 \frac{\epsilon_{ij}}{(\sum_k \epsilon_{kj})\,(\sum_l \epsilon_{il})}, \tag{3.26}$$

Crucially, in this estimation $O(N)$ data points are used to empirically estimate the elements of a $q \times q$ matrix ϵ, and information estimation involves a tractable summation over these matrix elements; in contrast, direct estimates of $I(U;\mathrm{X})$ would involve an intractable summation over (vastly undersampled) space for X. For typical applications where q is small, decoding thus provides an essential dimensionality reduction prior to information estimation: in a simple but biologically relevant case of two distinct stimuli ($q = 2$), information estimation only requires us to empirically construct a 2×2 confusion matrix. If required, one can apply well-known debiasing techniques for larger q [Strong *et al.*, 1998].

Maximum a posteriori (MAP) lower bound. In MAP lower bound, the decoding function F_ω is given by Bayesian inference of the most likely input u given that a response trajectory x was observed, under the exact probabilistic model for the biochemical reaction network. MAP decoder is optimal in that it provides the lowest average probability of error, $\Pr(\hat{U} \neq U)$, among all decoders. Typically, this will lead to a high mutual information value $I(\hat{U};U)$ compared to other (sub-optimal) decoders whose probability of error will likely be higher, making the information lower bound from MAP decoder a good benchmark for other decoder-based information estimates. We remind the reader, however, that even though MAP decoder achieves minimal error and typically

high $I(\hat{U};U)$ values, this does not mathematically guarantee that its information will *always* be higher or equal to the information of any other possible decoder, a fact that can be demonstrated explicitly using toy examples.

The MAP inference consists of finding the input that maximizes the posterior distribution [Murphy, 2012]

$$p(u|\mathrm{x}) = \frac{p(\mathrm{x},u)}{p_{\mathbf{X}}(\mathrm{x})} = \frac{p(\mathrm{x}|u)p_U(u)}{p_{\mathbf{X}}(\mathrm{x})}. \tag{3.27}$$

This corresponds to the following decoding function:

$$\hat{u} = F_\omega(\mathrm{x}) = \mathrm{argmax}_u \left[\log p(\mathrm{x}|u) + \log p_U(u)\right], \tag{3.28}$$

where ω represents the specification of the biochemical reaction network which determines $p(\mathrm{x}|u)$. Here, $p_U(u)$ is assumed to be known, and the likelihood $p(\mathrm{x}|u)$ can be calculated using Eqs. (3.10) or (3.17), for the continuous or discrete time representations, respectively.

One can apply the MAP-decoding based calculation of $I_{\mathrm{MAP}}(\hat{U};U)$ in two ways. First, when applied over real data $\mathcal{D}$, one can think of the procedure as a proper statistical estimation assuming that the biochemical network model is the correct generative model of the data (with estimation bias arising if it is not). Second, when applied, as we will do in the Results section, over trajectories $\mathcal{D}$ generated using exact stochastic simulation from the biochemical network model in the large N limit, this procedure is a Monte Carlo approximation to the information lower bound.

Note that even though the MAP decoder is optimal, it does not follow that $I_{\mathrm{MAP}}(\hat{U};U) = I(\mathbf{X};U)$. This is because optimal channel use that realizes $I(\mathbf{X};U)$ may need to employ block codes, where a *sequence* of inputs is encoded jointly into a *sequence* of trajectories, which is later also jointly decoded. In contrast, the decoding bound $I_{\mathrm{MAP}}(\hat{U};U)$ relies on one-shot use of the channel: a single input u is transduced into x which can immediately be decoded back into the estimate of the input, $\hat{u}$, on the basis of which the cell might make a decision. For many biological situations, this decoding setup should be more appropriate than the exact information calculation, as cells often need to react to stimuli as rapidly as possible in order to gain a selective advantage. Furthermore, it is difficult to conceive of biologically realistic encoders that would transform inputs into a block code in order to use the biochemical network channels optimally.

Maximum a posteriori upper bound (UB). Given that the optimal MAP decoding does not necessarily reach the exact mutual information, it is reasonable to ask how large the gap is between these two quantities. For discrete inputs, classic work in information theory proved a number of upper bounds on this gap when the channel is known [Samengo, 2002], with the Feder-Merhav bound perhaps being the most well known [Feder and Merhav, 1994]; Feder-Merhav provides an upper bound on the channel capacity given the overall probability of error in MAP decoding. In a separate work [Hledík *et al.*, 2018], we computed a new upper bound on information $I_{\mathrm{UB}}(U;\mathbf{X})$ that is consistent with not just the overall probability of error as in Feder-Merhav bound, but with the full confusion matrix ϵ obtained from optimal MAP decoding, and showed that the new bound is tight.

Our self-contained derivation [Hledík *et al.*, 2018] gives the following result

$$I(U;\mathbf{X}) \leq I_{\mathrm{UB}} = H(U) - \sum_{\hat{u}} p_{\hat{U}}(\hat{u})\phi(\pi_{\hat{u}}), \tag{3.29}$$

where $\pi_{\hat{u}} = \Pr(U \neq \hat{U}|\hat{u}) = 1 - \Pr(U = \hat{u}|\hat{U} = \hat{u})$ and functions ϕ and α can be expressed with the help of the floor and ceiling functions as:

$$\phi(\pi) = \alpha(\pi)\log_2\left\lfloor\frac{1}{1-\pi}\right\rfloor + (1-\alpha(\pi))\log_2\left\lceil\frac{1}{1-\pi}\right\rceil \tag{3.30}$$

$$\alpha(\pi) = \left\lfloor\frac{1}{1-\pi}\right\rfloor\left((1-\pi)\left\lceil\frac{1}{1-\pi}\right\rceil - 1\right). \tag{3.31}$$

This bound applies irrespectively of how the response trajectory space is represented (continuous or discrete, possibly of dimensionality much larger than that of the random variable U), since it is stated solely in terms of the input variable U and its MAP decode, $\hat{U}$.

3.2.6 Decoding-based information estimators

Support Vector Machine (SVM) lower bound estimator. The first model-free decoding approach we consider is based on classifiers called Support Vector Machines (SVMs). To begin we consider two possible inputs, $q = 2$. We define a decoding func-

tion F_ω by means of a helper function $f_\omega(\mathbf{x})$, such that $F_\omega(\mathbf{x}) = u^{(1)}$ if $\text{sign} f_\omega(\mathbf{x}) = -1$ and $F_\omega(\mathbf{x}) = u^{(2)}$ otherwise. Here,

$$f_\omega(\mathbf{x}) = \sum_{i=1}^{N_t} \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b \tag{3.32}$$

where $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the so-called "kernel function" to be defined below, b is the bias constant, N_t is the number of samples in $\mathcal{D}_{\text{train}}$ and $\alpha_1, \ldots, \alpha_{N_t}$ are obtained by solving standard SVM equations:

$$\min_{\substack{\alpha_1,\ldots,\alpha_{N_t}\in\mathbb{R},\\ \xi_1,\ldots,\xi_{N_t}\in\mathbb{R}^+}} \sum_{i,j=1}^{N_t} \alpha_i \alpha_j k(x_i, x_j) + \frac{C}{N_t} \sum_{i=1}^{N_t} \xi_i \tag{3.33}$$

subject to

$$y_i \sum_{j=1}^{N_t} \alpha_i k(\mathbf{x}_j, \mathbf{x}_i) \geq 1 - \xi_i, \qquad \text{for } i = 1, \ldots, N_t. \tag{3.34}$$

$y_i = -1$ whenever the input corresponding to the i-th trajectory in the training set, $\mathbf{x}_i$, is $u^{(1)}$, i.e., $u_i = u^{(1)}$; similarly $y_i = +1$ whenever the corresponding input is $u^{(2)}$, i.e., $u_i = u^{(2)}$. C is a positive regularization constant. Together, the parameters of the decoding function are $\omega = \{b, \alpha, \xi, C\}$.

To prevent overfitting and set the regularization parameter C using cross-validation, we split the full dataset $\mathcal{D}$ into training data, $\mathcal{D}_{\text{train}}$, that consists of N_t (here $\sim 70\%$ of the total, N) of labeled sample trajectories, chosen randomly but balanced across different inputs u; the remaining 30% of the data constitutes the testing data, $\mathcal{D}_{\text{test}}$. Parameters ω are estimated only over $\mathcal{D}_{\text{train}}$, after which the error matrix ϵ and the corresponding information estimate $I_{\text{SVM}}(\hat{U}; U)$ of Eq (3.26) are evaluated solely over $\mathcal{D}_{\text{test}}$. The test/train split procedure can be repeated multiple times to compute the mean and the bootstrapped error bar estimate for the information estimator, I_{SVM} [Granados *et al.*, 2018].

When we apply SVM decoding, we are still free to choose the kernel function. Here, we focus on two possibilities:

- **Linear kernel,** $k(x, x') = x^T x'$. The information estimate is based on a linear classifier that can learn to distinguish responses that differ in their conditional means, $\mu(u)$, but will result in close to chance performance if they don't. This is the simplest model-free decoding estimator and is thus a useful benchmark for more complex, non-linear decoders.

- **Radial basis functions kernel,** $k(x, x') = \exp(-\frac{||x-x'||^2}{2\sigma^2})$. This model-free decoder can be sensitive both to difference in the conditional means as well as higher-order statistics, e.g., the covariance matrix. Parameter σ is set via cross-validation to maximize the performance.

For multiclass classification we use a decision-tree SVM classification method [Benabdeslem, Khalid and Bennani, 2006], also called Dendrogram-SVM (DSVM) [Lajnef *et al.*, 2015]. To translate the multi-class classification into the canonical binary classification problem, this method uses hierarchical bottom-up clustering to define the structure of the graph, on which a binary classification is performed using SVMs at each graph node.

Gaussian decoder (GD) lower bound estimator. In this model-free estimation, we revisit the assumption that the (discretely sampled) output trajectories x given input u can be approximated with a multivariate Gaussian distribution, Eq. (3.19). The decoding function is then

$$\hat{u} = F_\omega(\mathrm{x}) = \mathrm{argmax}_u \left[\log \mathcal{N}(\mathrm{x}; \mu(u), \Sigma(u)) + \log p_U(u)\right]. \tag{3.35}$$

Here, parameters ω consist of conditional means and (possibly regularized) covariance matrices of the Gaussian distributions that need to be estimated from data, following the test/train procedure analogous to SVM decoding.

This method can be used with different parametric multivariate probability density functions replacing the multivariate Gaussian in Eq (3.35), with choices that approximate the statistics of the data (and thus the CME-derived distribution) better providing tighter lower bound estimates, $I_{\mathrm{GD}}(\hat{U}; U)$, of the exact information. By analogy with the exact MAP decoding using CME-derived response distribution, this method can also be understood as maximum a posteriori decoder but using approximate response distributions that need to be estimated from data. Here we decided to use the Gaussian distributions because they are the most unstructured (random) distributions based on measured first- and second-order statistics of the data. GD decoder thus should be able to discriminate various inputs if their responses differs either in the response mean or response covariance.

Neural Network (NN) lower bound estimator.

Artificial neural networks, first introduced by the neurophysiologist Waren McCulloch and the mathematician Walter Pitts in 1943 [McCulloch and Pitts, 1943], are nowadays the method of choice for classification that generally outperforms alternative machine learning techniques on very large and complex problems. Here we use one of the simplest neural networks, called the multi-layer perceptron (MLP). MLP is composed of layers of linear-threshold units (or LTUs), where each LTU computes a weighted sum of its inputs $z = \omega^T \mathrm{x}$, then applies an activation function to that sum and outputs the result $y = h(z) = h(\omega^T \mathrm{x} + \omega_0)$. Using a single LTU amounts to training a binary linear classifier by learning the weights ω. As with linear SVM, such classifier only has a limited expressive power [Rosenblatt, 1957], which can, however, be extended by stacking layers of LTUs so that outputs of the first layer are inputs to the second layer etc.

For illustrative purposes we choose for our decoding function $F_\omega(\mathrm{x})$ a fully connected neural network with two hidden layers (with 300 and 200 LTUs, respectively) that uses the exponential activation function with $\alpha = 1$:

$$h_\alpha(z) = \begin{cases} \alpha(\exp(z) - 1) & \text{if } z \leq 0 \\ z & \text{if } z > 0 \end{cases}.$$

For training, we used He-initialization, which initializes the weights with a random number from a normal distribution with zero mean and standard deviation $\sigma = 2/\sqrt{n_{in}}$, where n_{in} is the number of inputs to units in a particular layer [Géron, 2017], and Adam optimization with batch normalization and drop-out regularization [Abadi *et al.*, 2015; Géron, 2017]. As before, we trained the neural network on $\mathcal{D}_{\text{train}}$, followed by the evaluation of the error matrix ϵ and of the corresponding information estimate, $I_{NN}(\hat{U};U)$, from Eq (3.26), over $\mathcal{D}_{\text{test}}$. We emphasize that the detailed architecture of the neural network we selected here is not relevant for other estimation cases; in general, the architecture is completely adjustable to the problem at hand and should be selected depending on the size of the training dataset. The only selection criterion is the network performance on test data, with better performing networks for a given dataset typically providing tighter information estimates.

3.3 Results

3.3.1 Model-based estimation on simulated data

We start by considering three simple chemical reaction networks for which we can obtain exact information values using the model-based approach outlined in Methods Section 3.2.3. This will allow us to precisely assess the performance of decoding-based model-free estimates, and systematically study the effects of time discretization, the number of sample trajectories, and the number of distinct discrete inputs, q.

The three examples are all instances of a simple molecular birth-death process, where molecules of $\tilde{X}$ are created and destroyed with rates α and β, respectively:

$$\xrightarrow{\alpha(U)} \tilde{X} \xrightarrow{\beta(U)} \varnothing. \tag{3.36}$$

The reaction rates, α and β, will depend in various ways on the input, U, and possibly time, as specified below. Given an initial condition, $x(t=0)$, the production and degradation reactions generate continuous-time stochastic trajectories, $x(t)$, recording the number of molecules of $\tilde{X}$ at every time $t \in [0, T]$, according to the Chemical Master Equation (3.3). These trajectories, or their discretized representations, are considered as the "outputs" of the example reaction networks, defining the mutual information $I(X;U)$ that we wish to compute. In all three examples we start with the simplest case, where the random variable U can only take on two possible values, $u^{(1)}$ and $u^{(2)}$, with equal probability, $p_U(u^{(1)}) = p_U(u^{(2)}) = 0.5$.

- **Example 1.** In this case, $x(t=0) = 0$, $\beta = 0.01$, independent of the input U, and the production rate depends on the input as $\alpha(u^{(1)}) = 0.1$, $\alpha(u^{(2)}) = 0.07$. Here, the steady state is given by Poisson distribution with mean number of molecules $\langle x(t \to \infty) \rangle = \alpha/\beta$. Steady-state is approached exponentially with the timescale that is the inverse of the degradation rate, β^{-1}. These dynamics stylize a class of frequently observed biochemical responses where the steady-state mean expression level encodes the relevant input value. Even if the stochastic trajectories for the two possible inputs are noisy as shown in Fig 3.2A, we expect that the mutual information will climb quickly with the duration of the trajectory, T, since (especially in steady state) more samples provide direct evidence about the relevant input already at the level of the mean trajectories.

- **Example 2.** In this case, $x(t=0)=0$, $\beta=0.01$, independent of the input U, and the production rate depends on the input as $\alpha(u^{(1)},t)=0.1$, $\alpha(u^{(2)},t)=0.05$ for all $t<1000$, while for $t\geq 1000$ the production rate is very small and independent of input, $\alpha(u,t)=5\cdot 10^{-4}$. In the early period, this network approaches input-dependent steady state with means whose differences are larger than in Example 1, but the difference decays away for $t>1000$ as the network settles towards vanishingly small activity for both inputs, as shown in Fig 3.2B. These dynamics stylize a class of transient biochemical responses that are adapted away even if the input state persists. In this case, lengthening the observation window T will not provide significant increases in information.

- **Example 3.** In this case, $x(t=0)=10$. All reaction rates depend on the input, $\alpha(u^{(1)})=0.1, \alpha(u^{(2)})=0.05, \beta(u^{(1)})=0.01, \beta(u^{(2)})=0.005$, and are chosen so that the mean $\langle x(t)\rangle=10$ is constant across time and equal for both conditions, as shown in Fig 3.2C. In this difficult case, inputs cannot be decoded at the level of mean responses but require sensitivity to at least second-order statistics of the trajectories. Specifically, signatures of the input are present in the autocorrelation function for x: the timescale of fluctuations and mean-reversion is two-fold faster for u^1 than u^2. While this case is not frequently observed in biological systems, it represents a scenario where, by construction, no information about the input is present at the level of single concentration values and having access to the trajectories is essential. Because there is no difference in the mean response, we expect linear decoding methods to provide zero bits of information about the input. This case is also interesting because of the recent focus on pulsatile stationary-state dynamics in biochemical networks [Dalal *et al.*, 2014]. These pulses, reported for transcription factors such as Msn2, NF-κB, p53, etc., occur stochastically and, when averaged over a population of desynchronized cells, can yield a flat and featureless mean response. Information about the stimulus could, nevertheless, be encoded in either the frequency, amplitude, or other shape parameters of the pulses. While a generative description of such pulsatile dynamics goes beyond a birth-death process considered here, from the viewpoint of decoding, both pulsatile signaling and our example present an analogous problem, where the mean response is not informative about the applied input.

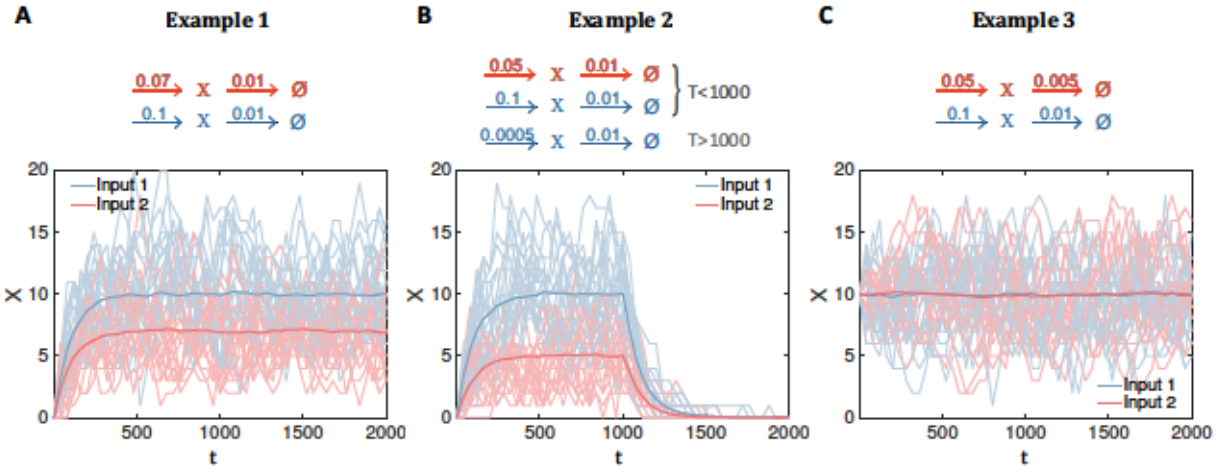

Figure 3.2: **Example biochemical reaction networks and their behavior.** Three example birth-death processes, specified by the reactions in the top row for each of the two possible inputs ($u^{(1)}$ in blue, $u^{(2)}$ in red), stylize simple behaviors of biochemical signaling networks. **(A)** Input is encoded in both the transient approach to steady state and the steady state value. **(B)** Input is encoded in the magnitude of the transient response which is subsequently adapted away. **(C)** Input is encoded only at the level of temporal correlations of the response trajectory. Bottom row shows example trajectories generated using the Stochastic Simulation Algorithm for the copy number of $\tilde{X}$ molecules, $t \in [0, 2000]$, for each network and the two possible inputs (light blue, light red); while plotted as a connected line for clarity, each trajectory represents molecular counts and is thus a step-wise function taking on only integer or zero values. Dark blue, red lines show the conditional means over $N = 1000$ trajectory realizations.

Exact information approximations and bounds for continuous and discrete trajectories. Armed with the full stochastic model for the three example reaction networks, we can compute the mutual information, $I^*_{\text{exact}}(X, U)$, between the continuous-time stochastic trajectories and the (binary) input variable U, following Eq (3.14). This result depends essentially on the length of the observed trajectory, $t \in [0, T]$, since T controls the number of observed reaction events and thus the accumulation of evidence for one or the other alternative input. As the approximation is implemented by Monte-Carlo averaging of exact log probabilities for the response trajectories, its variance will depend on the number of sample trajectories generated by the SSA. Because these information values will represent the "gold truth" against which to evaluate

subsequent estimators, we choose a large number of $N = 1000$ trajectory realizations per input condition, and verify the tightness of the exact Monte Carlo approximation by computing the standard deviation over 20 independent re-runs of the approximation procedure.

Fig 3.3 shows how the exact Monte Carlo information computation depends on the trajectory duration, T, for each of the three example cases. As expected, the information increases monotonically with T towards the theoretically maximal value of 1 bit, corresponding to perfect information about two *a priori* equally likely input conditions. The exact shape of the information curve depends on the shape of the mean trajectory, as well as on its variance and higher-order statistics: for example, even though the two inputs for Example 1 are most distinct at the level of mean responses for later T values, the noise is higher compared to Example 2, such that at $T = 2000$ there is more total information in trajectories of Example 2 than Example 1. Conversely, even though the trajectories in Example 3 do not differ at the level of the mean at all, they still carry all information about the relevant input once sufficiently long trajectories can be observed (and assuming full knowledge of the reaction network is available).

One can similarly compute the Bayes-optimal or MAP decoding bound using Eq (3.28) for continuous trajectories. This quantifies the ultimate accuracy limit with which each single observed trajectory can be decoded into the input that gave rise to it. As demonstrated in Fig 3.3 in dashed black line and consistent with the Data Processing Inequality requirements outlined in the Methods, $I^*_{\mathrm{MAP}}(\hat{U};U) \leq I^*_{\mathrm{exact}}(X;U)$. Equality is not reached because the optimal use of the channel requires block coding schemes, in contrast to our setting where different inputs are sequentially sent through the biochemical network and immediately decoded. The observed gap between the MAP optimal decoding estimate and the true information appears to be small in each of the three cases; one can upper-bound the gap itself by an improvement over the standard Feder-Merhav calculation following Methods Section 3.2.5. While the resulting upper bound on information, I_{UB}, is not tight in this case, it nevertheless provides a control of how far optimal decoding could be from the true information estimate, a question that has repeatedly worried the neuroscience community facing similar problems [Borst and Theunissen, 1999]. It is worth noting that if MAP decoder can tractably be computed, so can the upper bound, irrespective of the dimensionality of the space of responses,

X.

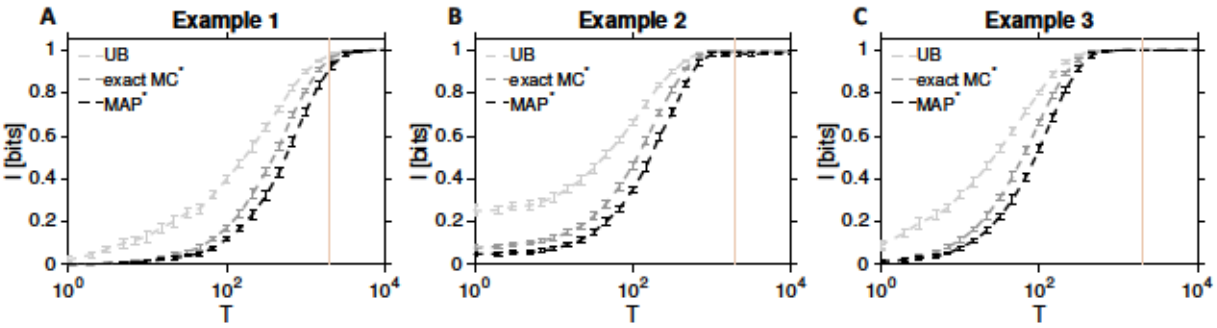

Figure 3.3: **Information about inputs encoded by complete response trajectories of the example biochemical reaction networks.** Exact Monte Carlo approximation for the information, $I^*_{\text{exact}}(X;U)$, is shown for Example 1 **(A)**, Example 2 **(B)**, and Example 3 **(C)** from Fig 3.2 in dashed dark gray line; error bars are standard deviations across 20 replicate estimations, each computed over $N = 1000$ independently generated sample trajectories per input condition. Information is plotted as a function of the trajectory duration, T; yellow vertical line indicates $T = 2000$ as a representative duration used in further analyses below, at which most of the information about input is in principle available from the response trajectories of our systems. $I^*_{\text{MAP}}(\hat{U};U)$ (dashed black line) is the optimal decoding lower bound, and I^*_{UB} (dashed light gray) is the upper bound on the information, computed by applying Eq (3.29).

Fig 3.3 summarizes the absolute limits on information transmission and optimally decodable information, for each of our three example networks. These values are limits inasmuch as they assume that every reaction event can be observed and recorded with infinite temporal precision, and that the encoding stochastic process is perfectly known. While it is interesting to contemplate whether biological systems themselves could compute with or act on singular, precisely-timed reaction events and thus make optimal use of the resulting channel capacity (mimicking the debate between spike timing code and spike rate code in neuroscience), our primary focus here is to estimate information flows from experimental data. Typically, experiments record the state of the system—e.g., concentration of signaling molecules—in discretely sampled time. To explore the effects of time discretization, we first fix the observation length for our trajectories to $T = 2000$, sufficiently long that the trajectories in principle contain more

than 90% of the theoretically maximal information for each of the three example cases. We then resample the trajectories on a grid of d equally spaced time points, as illustrated in Fig 3.4A.

Figs 3.4B-D compare the exact Monte Carlo information approximation for discrete trajectories, $I_{\text{exact}}(\mathbf{X};U)$, MAP lower bound for discrete trajectories, $I_{\text{MAP}}(\hat{U};U)$, and the corresponding upper bound, I_{UB}, to the theoretical limits from Fig 3.3 obtained using continuous trajectories. In line with the chain of inequalities in Eq (3.25), information in discretely resampled trajectories is lower than the true information in continuous trajectories, but converges to the true value as $d \to \infty$. In particular, once the discretization timestep T/d is much lower than the inverse of the fastest reaction rate in the system, discretization should incur no significant loss of information. In practice, however, high sampling rate (large d) limit has significant drawbacks: first, it is technically difficult to take snapshots of the system at such high rates (e.g., due to fluorophore bleaching); second, the fast dynamics of the reaction network may be low-pass filtered by the readout process (e.g., due to fluorophore maturation time, or slower downstream reaction kinetics); and third, for model-free approaches high d implies that decoders need to be learned over input spaces of high dimensionality, which could be infeasible given a limited number of experimentally recorded response trajectories. In previous work [Hansen and O'Shea, 2015; Hafner *et al.*, 2017], trajectories were typically represented as $d \approx 1 \sim 100$ dimensional vectors, which in our examples would capture $\sim 80\%$ or more of the theoretically available information. It is likely that this can be improved further with smart positioning of the sampling points and that not all theoretically available information could actually be accessed by the organism itself, suggesting that typically used discretization approaches have the potential to capture most of the relevant information in the responses. What is important for the analysis at hand is that given the dimensionality d of the discretized response trajectories, MAP decoder is guaranteed to reach the minimal decoding error among all possible decoders, and will turn out to be a relevant benchmark, by yielding the highest information, $I_{\text{MAP}}(\hat{U};U)$, in Figs 3.4B-D among all decoders considered. In what follows, we will examine how various model-free decoding estimators approach this limit, as a function of d and the number of sample trajectories, N.

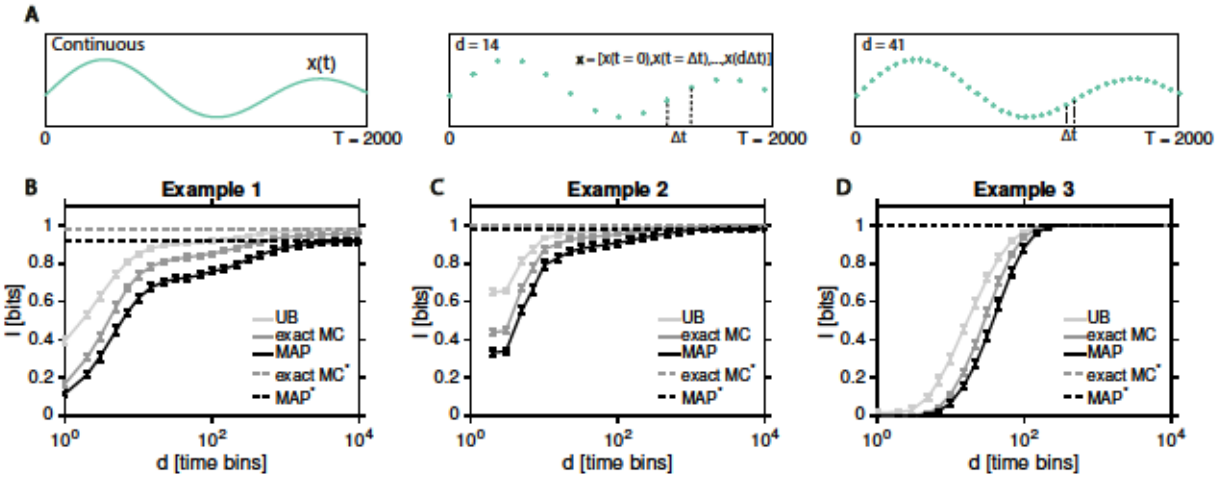

Figure 3.4: **Information loss due to temporal sampling. (A)** Schematic representation of the resampling of a continuous-time response trajectory (left) at $d = 14$ (middle) or $d = 41$ (right) equally spaced time points. Resampled response trajectories are represented as d-dimensional real vectors, $\mathbf{X} \in \mathbb{R}^d$, for the case of a single output chemical species. **(B–D)** Exact Monte Carlo information approximations for discrete trajectories, $I_{\text{exact}}(\mathbf{X};U)$ (dark solid gray), optimal decoding lower bound, $I_{\text{MAP}}(\hat{U};U)$ (dark solid black), and the upper bound, I_{UB} (light solid gray) are plotted as a function of d. Continuous-time limits from Fig 3.3 are shown as horizontal lines: $I^*_{\text{exact}}(X;U)$ (dashed dark gray), $I^*_{\text{MAP}}(\hat{U};U)$ (dashed black). Error bars as in Fig 3.3.

Performance of decoding-based estimators. After establishing our model-based "gold standard" for decoding-based estimators acting on trajectories represented in discretized time, $I_{\text{MAP}}(\hat{U};U)$, we turn our attention to the performance comparison between various model-free algorithms. The results are summarized in Fig 3.5, which shows how estimator accuracy depends on the dimensionality of the problem, d, given a fixed number, $N = 1000$, of sample trajectories per input condition. In contrast, Fig 3.6 assumes a fixed dimensionality d of trajectory vectors and explores how the estimator performance depends on the number of samples, N.

Figs 3.5 and 3.6 lead us to the following conclusions:

- **Nonlinear SVM** using the radial basis functions (rbf) kernel performs best for Examples 1 and 2. Regardless of the number of samples, N, or the number of time bins, d, its estimates are very close to I_{MAP}, especially for the relevant regime

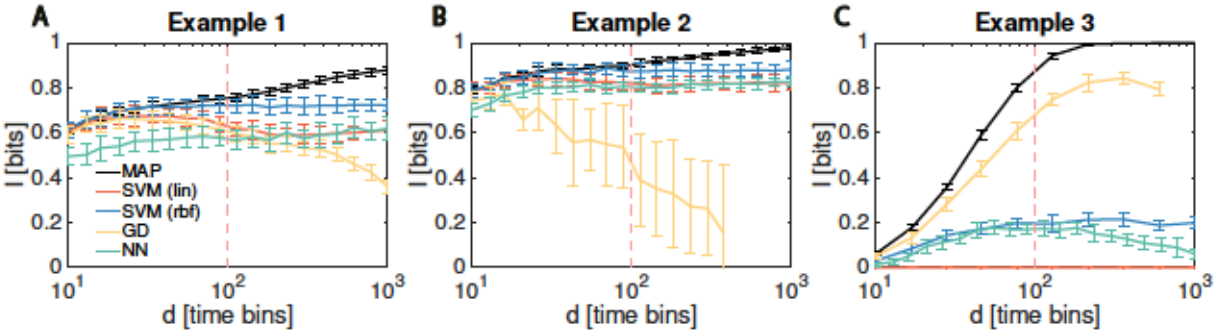

Figure 3.5: **Performance of decoding-based estimators depends on the dimensionality of the response trajectories.** Performance of various model-free decoding estimators (colored lines) for Examples 1 **(A)**, 2 **(B)**, 3 **(C)**, respectively, compared to the MAP bound, I_{MAP} (black line), as a function of input trajectory dimension, d. In all cases, the number of sample trajectories per input condition is $N = 1000$, error bars are std over 20 replicate estimations. Decoding estimators: linear SVM, $I_{\mathrm{SVM(lin)}}$ (orange); radial basis functions SVM, $I_{\mathrm{SVM(rbf)}}$ (blue); the Gaussian decoder with diagonal regularization (see Fig 3.9), I_{GD} (yellow); multi-layer perceptron neural network, I_{NN} (green). Dashed vertical orange line marks the $d \leq 100$ regime typical of current experiments.

$d \sim 10 - 100$. Even for higher d, the estimator shows hardly any overfitting and thus stable performance, a feature we have observed commonly in our numerical explorations. The estimator is sample efficient, typically providing estimates with smallest error bars.

- **Linear SVM** slightly underperforms kernelized SVM on Examples 1 and 2, and—as expected—completely fails on the linearly inseparable Example 3. Interestingly, even though more expressive, kernelized SVM seems to incur no generalization cost relative to linear SVM even at low number of samples. For all examples we tested, kernelized SVM thus appears to be a method of choice; linear SVM, however, is still useful as a benchmark to measure what fraction of the information is linearly decodable from the signal.

- **The Gaussian decoder** has the best performance on Example 3, is competitive for low d for Example 1, and doesn't perform satisfactory for Example 2. As

shown in Fig 3.9, regularized estimation of covariance matrix appears crucial for good performance, but smoothing of the originally discrete trajectory does not help. Even with regularization, this estimator is not sample efficient for Example 1: the trajectories are linearly separable without a full estimate of the covariance (as evidenced by the success of the linear SVM), yet the Gaussian decoder requires one to two orders of magnitude more samples to match the linear decoder performance. This drawback turns into a benefit for Example 3: the Gaussian assumption can be viewed as a prior that second-order statistics are important for decoding (which is correct in this case). Kernelized SVM and the neural network, while more general, need to learn from many more training samples to zero in on these features, and fail to reach the Gaussian decoder performance even for $N = 10^4$. We hypothesized that the failure of the Gaussian decoder on Example 2 is due to the difficulty of the Gaussian approximation to capture the period $T > 1000$ when the mean number of $\tilde{X}$ is close to zero: here, first, the Gaussian assumption must be strongly violated, and, second, the estimation of (co)variance from finite number of samples is close to singular due to the small number of reaction events in this period. Even though the $T > 1000$ epoch is not informative about the input, a badly conditioned decoder for this epoch can actually adversely affect performance. We confirmed this hypothesis by building the Gaussian decoder restricted to $T < 1000$ that reliably extracted ≥ 0.8 bits of information in Example 2, close to the MAP decoding bound and the performance of SVM-based estimators.

- **Neural network decoder** reaches a comparable performance on Examples 2 and 3 to the SVMs, but fails to be competitive for the simple Example 1. This is most likely because this estimator is sample inefficient, as implied by its continual increase in performance with N that did not saturate at highest N we tried. Given their expressive power, neural network decoders should be viewed as the opposite benchmark to the linear decoders: they have the ability to pick up complex statistical structures but only with a sufficient number of samples. Indeed, as we will see subsequently for applications to real data, neural networks can match and exceed the performance of SVMs. We emphasize that we used a neural network with a fixed architecture for all three examples on purpose, to make results com-

parable across examples; the performance can likely be improved by optimizing the architecture separately for each estimation problem.

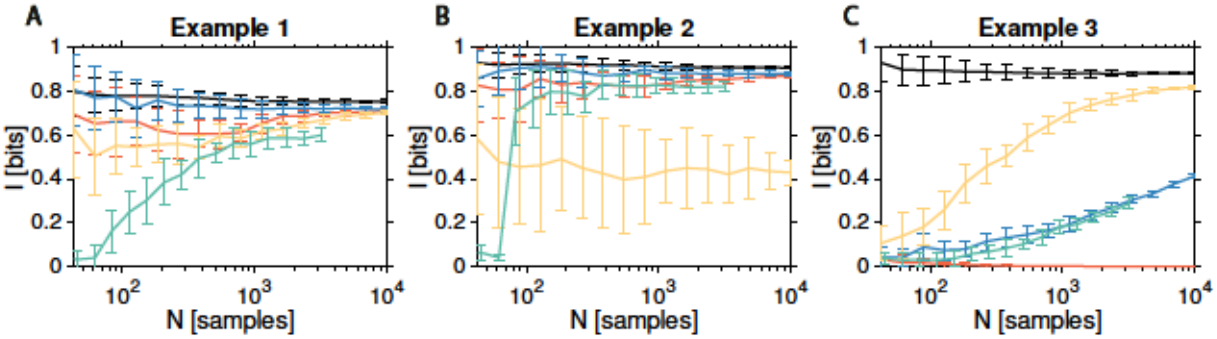

Figure 3.6: **Convergence of decoding-based estimator performance with increasing number of response trajectory samples.** Performance of various model-free decoding estimators (colored lines) for Examples 1 **(A)**, 2 **(B)**, 3 **(C)**, respectively, compared to the MAP bound, I_{MAP} (black line), as a function of the number of samples, N, per input condition. Response trajectories are represented as $d = 100$ dimensional vectors. Plotting conventions as in Fig 3.5.

Multilevel information estimation. We next asked whether our conclusions hold also when the space of possible inputs is expanded beyond binary, assuming that U can take on q distinct values with equal probability, i.e., $p_U(u) = 1/q$. We focused on Example 2, and constructed cases for $q = 2, \ldots, 5$ such that the production rate α for $0 < T < 1000$ takes on q uniformly spaced values between 0 and the maximal rate equal to $\alpha = 0.1$ used in Fig 3.2B. In effect, this "tiles" the original, two-state-input dynamic range uniformly with q input states, as illustrated in Fig 3.7A.

Our expectation is that with increasing q, the information should increase, but slowly saturate as reliable distinctions between nearby input levels can no longer be made due to the intrinsic biochemical stochasticity. This is indeed what we see in Fig 3.7B, which shows the exact information, the MAP lower bound and the upper bound. Consistent with our findings for two-input case, SVM using radial basis functions remains the estimator of choice for all values of q, followed by the linear SVM and then the neural network decoder, as shown in Fig 3.7C.

Performance comparisons with model-free information approximations. There exist many algorithms for estimating information directly, without making use of the de-

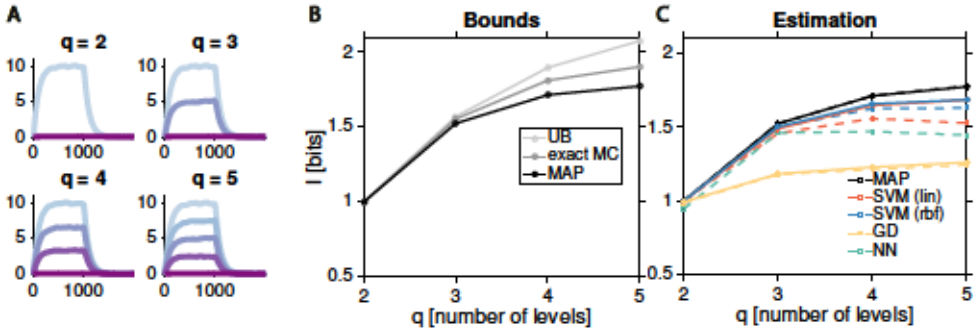

Figure 3.7: **Information estimation for multilevel inputs.** **(A)** Extension of Example 2 from Fig 3.2B to $q = 2, \ldots, 5$ discrete inputs. We chose the inputs such that the response for the system at $T < 1000$ converges towards q equally spaced levels with the same dynamic range as the original example; dynamics at $T \geq 1000$ remain unchanged from the original Example 2. **(B)** Model-based information bounds as a function of the number of input levels for trajectories represented as $d = 100$ dimensional vectors: exact Monte Carlo calculation (dark gray), MAP decoding bound (black), upper bound (light gray). **(C)** Performance of model-free estimators, as indicated in the panel, compared to the MAP bound (black). Dashed lines show estimations using $N = 10^3$ sample trajectories per condition, solid lines using $N = 10^4$ samples per condition; in both cases, we show an average over 20 independent replicates, error bars are suppressed for readability.

coding lower bound. The best known estimator for continuous signals is perhaps the k-nearest-neighbor (knn) estimator [Kraskov *et al.*, 2004]. We have also introduced estimators based on parametric assumptions about the response distribution, such as the Gaussian approximation (Methods Section 3.2.4); both belong in the family of binless approximations, which act directly on real-valued response vectors. In contrast, binning approximations first discretize the responses $\mathbf{X}$. The simplest such approach is perhaps the direct estimator of information or entropy [Strong *et al.*, 1998], and a good review is provided in Ref [Paninski, 2003]. We evaluated the performance of the Gaussian approximation to find that it can systematically overshoot the true information with a bias that is difficult to assess (Fig 3.10); this appears to happen also in the regime where the biochemical noise should be small (relative to the mean), and the stochastic dynamics should be describable in terms of Langevin approximations with the resulting Gaussian response distributions. These approximations converge to the

true solution in terms of their first and second moments, yet do not seem to lead to unbiased estimate for the entropies and thus the mutual information. In contrast to the Gaussian decoder, Gaussian approximation should not be used without a better understanding of its bias and applicability.

We therefore decided to focus on the comparison of decoding estimators with knn, which has been used previously on data from biochemical signaling networks [Selimkhanov *et al.*, 2014]. The results are shown in Fig 3.8. K-nearest-neighbors performs well on the easy Example 1, and suffers drastic performance drop for Example 2, while crashing catastrophically by reporting negative values in Example 3. We reasoned that part of the difficulty may be the fact that synthetic trajectories for our Examples are defined over non-negative whole numbers only, whereas the knn assumes real valued vectors. This is confirmed by Fig 3.11 which shows that the knn performance can be substantially improved by adding a small amount of Gaussian iid noise to every component of the response trajectory vectors, $\mathbf{X}$. This restores the knn performance in Example 2 close to that of the SVM-based estimators, but still produces close-to-zero bits of information for Example 3.

3.4 Conclusions

Here we show a tractable Monte Carlo scheme to estimate the information transmission with arbitrary precision when the complete state of reaction network is observed and the inputs linearly alter a set of reaction rates. We compute the exact information for three simple biochemical reaction networks examples which provide a reference to evaluate the performance of the family of decoding-based model-free estimators. While in the low data regime the linear or kernelized-based estimator can closely approach the optimal decoder performance in examples 1 and 2, in the large N regime the neural-network-based schemes can also provide good estimates.

In contrast to information approximations for which it is often impossible to assess its precision, bias or even its sign when the dimension, d, of the trajectories is large, the decoding approach yields a conservative estimate of the true information. Furthermore, it performs close to optimum when multiple input values are decoded. In the final part, while comparing the decoding-based estimations with the commonly used

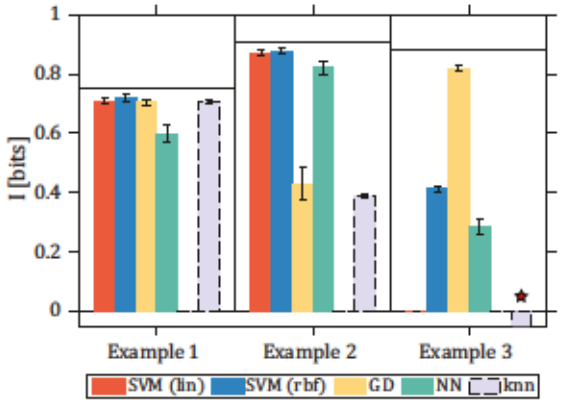

Figure 3.8: **Comparison of decoding-based and knn information estimators.** Information estimates for decoding-based (color bars) and knn (gray bar) algorithms (here we set $k = 1$, for further details of knn estimation, see Fig 3.11). Note that knn is not a decoding estimator and thus could exceed $I_{\mathrm{MAP}}(\hat{U}; U)$ (shown as a horizontal black line for each of the three example cases) to approach the exact $I_{\mathrm{exact}}(\mathbf{X}; U)$. Here we use trajectories discretized over $d = 100$ time bins, and $N = 10^4$ trajectory samples per input. The performance of knn can be substantially improved by adding a small amount of gaussian noise to the trajectory samples; its resulting performance as a function of N and d is shown in Fig 3.11. Red star denotes the failure of knn on Example 3 where substantially negative information values are returned (exact value not plotted).

knn estimator we show that in many cases they perform better, especially in Example 3, where the noise levels are high.

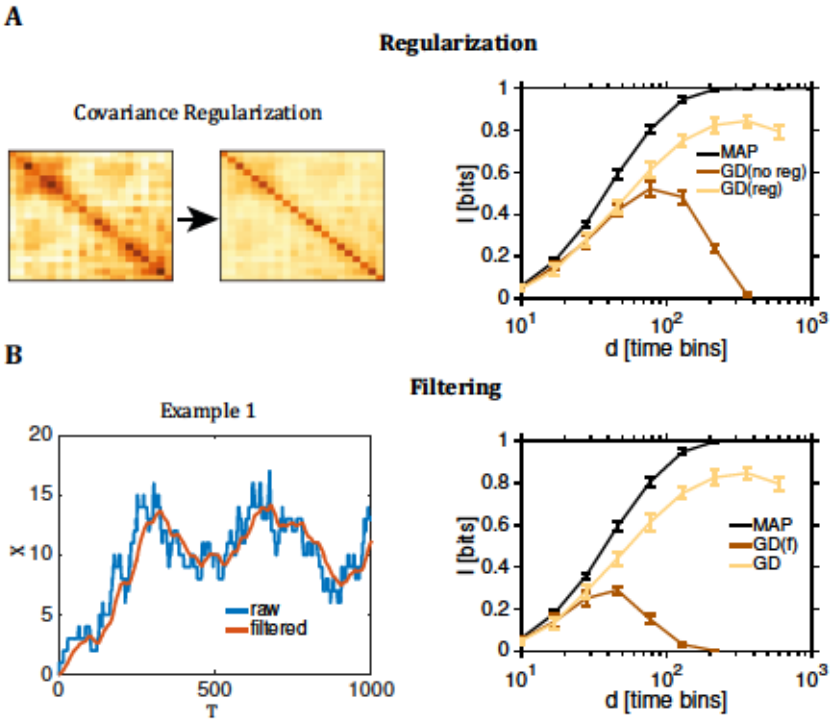

Figure 3.9: **Effects of covariance matrix regularization and signal smoothing on Gaussian-decoder-based estimation. (A)** At left. Diagonal covariance regularization following Ref [Yatsenko *et al.*, 2015]. Briefly, λ times the identity matrix is added to the empirical covariance matrix with the hyperparameter λ set so that the likelihood on test data is maximized. Shown is the empirical (left) and regularized (right) covariance matrix for Example 3, using $d = 20$ and $N = 30$ sample trajectories. At right. Information estimates for Example 3: I_{MAP} decoding bound (black), Gaussian decoder estimate, $I_{\mathrm{GD(reg)}}$, with optimal diagonal regularization for each d (yellow, as in Fig 3.5C), Gaussian decoder estimate, $I_{\mathrm{GD(noreg)}}$ (brown). Without regularization, the estimate suffers an abrupt drop as d increases and the empirically estimated covariance matrix becomes close to singular. N and plotting conventions are as in Fig 3.5. **(B)** The effects of trajectory filtering on information estimates. At left. A raw integer-valued stochastic trajectory for $\tilde{X}$ (blue) can be filtered by a low-pass exponential decay filter with adjustable timescale, $\tau = 1 - 10^3$, here $\tau = 50$ (red) to yield real-valued trajectory. At right. Regularized Gaussian-decoder information estimates with (brown) and without (yellow) filtering. Filtering does not improve but can decrease the estimation performance, even when the filtering timescale is adjusted.

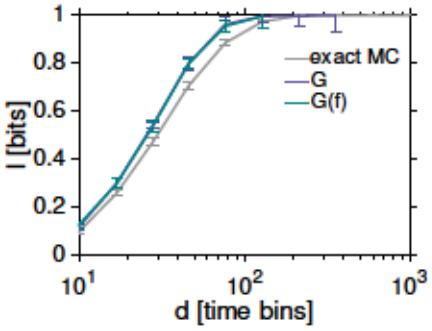

Figure 3.10: **Gaussian approximation to the information can lead to an uncontrolled overestimation of the true information.** Gaussian approximation is evaluated for Example 3 in Fig 3.5C, using $N = 1000$ per condition. Exact Monte Carlo approximation of the information, $I_{\mathrm{exact}}(\mathbf{X}; U)$, is shown in dark gray. Information estimates following Methods Section 3.2.4 are shown in violet (Gaussian approximation for raw, integer-valued response trajectories) or in cyan (Gaussian approximation for filtered trajectories, as in Fig 3.9). In both cases the Gaussian approximation overshoots the true information value. Further numerical analyses (not shown) indicate that the difference is hard to predict and that it persists even when the reaction rates are chosen such that the mean expression level is ten-fold higher (and the intrinsic stochasticity correspondingly lower). This makes direct Gaussian approximation risky to use, in contrast to the Gaussian-decoder based estimate, which is guaranteed to stay below I_{exact}.

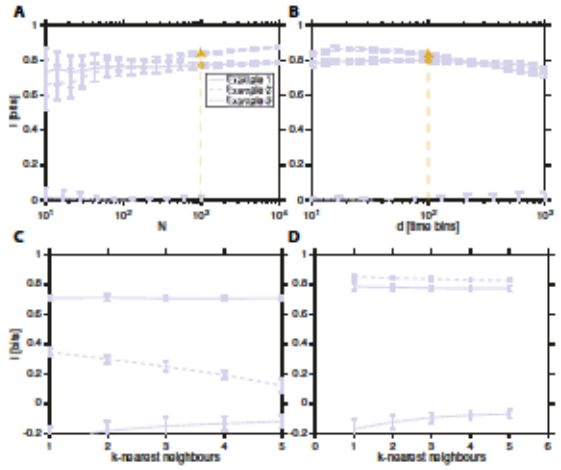

Figure 3.11: **Behavior of the knn information estimator.** Compared to knn results in Fig 3.8, the results in A, B and D are estimated following the same procedure, while adding a small amount of IID zero-mean Gaussian noise to each response trajectory at every time bin; the noise variance must be $\ll 1$ but otherwise does not affect the results much. This results in good estimates even at low sample number, N, and provides nearly stable estimation as a function of the trajectory dimension, d, for Example 1 and Example 2. It, however, does not resolve the estimator failure for Example 3. **(A)** Dependence of the knn estimator performance on the number of samples. Yellow plot symbols indicate the number of samples per condition, $N = 10^3$, used in Fig 3.8. **(B)** Dependence of the knn estimator performance on the trajectory dimension. Yellow plot symbols indicate the dimension, $d = 10^2$, used in Fig 3.8. **(C, D)** Dependence of the knn estimates on the number of nearest neighbors, k, at $N = 10^3$ and $d = 10^2$, without the addition of noise (C) or with the addition of noise (D).

4 Application of decoding-based information estimates to single cell dynamical data

The work presented in this chapter contains in section 4.2 a study performed in collaboration with Gašper Tkačik, Alejandro Granados, Peter Swain, Julian Pietsch and Iseabail Farquhar, and was published in PNAS (see [Granados *et al.*, 2018]). The section 4.3, was carried out by Sarah Cepeda, analyzing unpublished data provided by Alejandro Granados, it corresponds to the experimental section of the paper submitted for publication (see pre-print in [Cepeda-Humerez *et al.*, 2019]) and is reproduced here with minimal changes.

Abstract

In this section, we show the application of i) the linear SVM decoding estimator to the responses of 10 transcription factors in yeast and ii) the methods proposed in chapter 3, to previously published data on ERK and Ca^{+2} signaling, and yeast stress response. From i) we learn about the internal organization of yeast stress response signals, which we find happens through two logical channels: the generalist and specialist channel. For each channel, we identify characteristic features, for instance the timing of the responses. Specialists respond to specific stresses, are faster and sensitive to a wide range of changes in stress levels, whereas generalists respond to multiple stresses only if the stress is high. In ii) we estimate for the first time the information encoded in random pulses of nuclear-translocating transcription factors; specifically, we establish that environment-related information is present in the higher-order (beyond mean response) statistical features of the steady state response. Furthermore, we report the performance of non-model based estimators and show several results that are qualita-

tively consistent with the estimations in synthetic signals (examples 1-3 in chapter 3).

4.1 Dynamical signals in single-cells

The development of fluorescent sensors that permit high-resolution time-lapse imaging in living cells allows researchers to accurately track the dynamic behaviour of specific proteins in single cells. These time sequences, describe changes in concentration and protein localization over time [Purvis and Lahav, 2013]. However, the observables vary over a broad range of timescales, thus it is crucial to consider both: the right duration of the observations T and the appropriate sampling frequency T/d (the relevant variables considered in chapter 3). Inappropriate values for these quantities can lead to misinterpretations or inaccurate descriptions, for instance, when the observed levels of the phosphorylated kinase ATM (ATM-P) were measured after DNA damage, the initial observations done within one hour at high sampling rate, reported fast responses that reached the maximum after 5 minutes [Cuadrado *et al.*, 2006]. However, subsequent measurements done within 10 hours at 1 hour sampling rate exhibited an oscillatory response following DNA damage [Batchelor *et al.*, 2008].

In theory, these oscillations or other dynamical patterns of the signal may contain

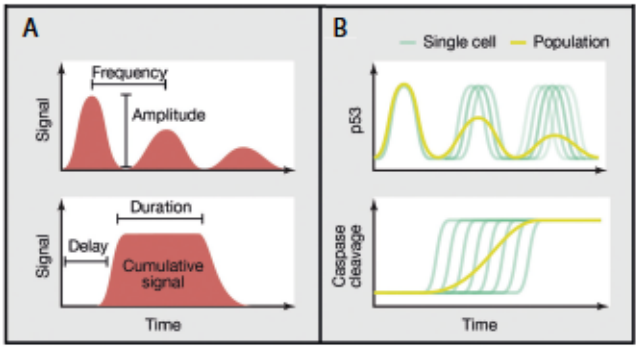

Figure 4.1: **Dynamic responses in single cells. (A)** Shows dynamical signals and common features considered as information carriers. **(B)** Single cell response compared to the response of the population in two examples. This figure is Modified from [Purvis and Lahav, 2013].

biologically relevant information about environmental changes. Current analysis approaches hand-pick certain signal features, for instance frequency or amplitude [Behar

and Hoffmann, 2010] (see Fig 4.1 A), without conclusive evidence that these features best represent environmental signals. Nevertheless, information might also be encoded in complex characteristics of the curve that are hard or practically impossible to distinguish by eye (this is conceptually similar to Example 3 in chapter 3).
For non-linear dynamical encoding it is essential to observe single-cell responses in order to identify subtle differences given that previous observations of individual cells revealed that the average dynamic response of a population often represents a distorted version of individual patterns. For example, p53 pulses in response to DNA damage have similar height and duration, but the loss of synchrony among individual cells gives the appearance of damped oscillations in the mean collective response [Lahav *et al.*, 2004]. Similarly, the "switch-like" responses of individual cells, namely cleavage of caspase substrates during apoptosis, seem to occur gradually in a population of cells [Tyas *et al.*, 2000] (see Fig 4.1 B).

In principle, the fact that the signals change over time does not necessarily mean that they encode information in their dynamics and trigger distinct reactions, but practical observations of single signaling factors show that in fact they do. For example, in the extracellular signal-regulated kinase (Erk) pathway in rat neural precursors, the nerve growth factor (NGF) induced sustained Erk activation while the epidermal growth factor (EGF) triggers a transient response (see Fig 4.2). Furthermore, it was also known that

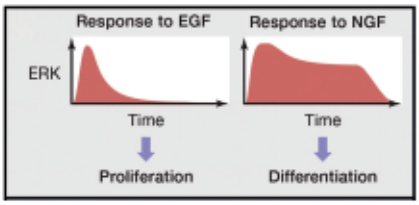

Figure 4.2: **Dynamic signaling encodes input identity and leads to different cell response**. The extracellular signal-regulated kinase (ERK) pathway encodes in its dynamics the identity of the growth factor and triggers different cell fates of rat neural precursors. This figure is modified from [Purvis and Lahav, 2013].

NGF induces differentiation, whereas EGF induces proliferation [Gotoh *et al.*, 1990; Traverse *et al.*, 1992; Nguyens *et al.*, 1993]. Similar examples involve other signaling molecules p53, NF-κB, Ca^{2+}, Msn2 [Hafner *et al.*, 2017; Hao and Shea, 2012] and a

recent study showing that pulses encode the identity of the stimuli in the Notch pathway [Nandagopal *et al.*, 2018].

4.1.1 Nuclear translocation

The spatiotemporal observation of regulatory proteins, especially during the 1990's [Kuge *et al.*, 1997], revealed that many signaling responses affect the nuclear localization of transcription factors and of kinases [Cyert, 2001; Hao and Shea, 2012]. Recently with the help of time-lapse microscopy and microfluidic systems it is possible to observe dynamical changes in single cells by trapping them, then removing daughter cell with fluid flow and allowing accurate control of the environment with the laminar flow of the media [Crane *et al.*, 2014]. An example of nuclear localization of Msn2 in yeast, under carbon stress, is shown in Fig 4.3. Depending on whether TFs act as repressors or

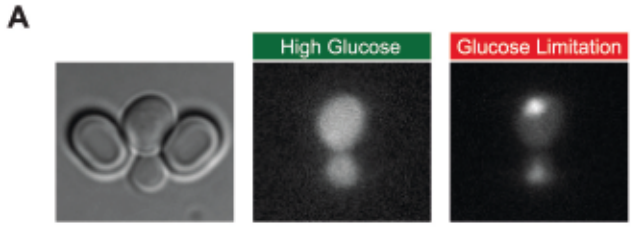

Figure 4.3: **Nuclear localization.** An image of a yeast cell in a microfluidic system. The signal (Msn2) at high glucose is distributed in the cytoplasm and during glucose limitation it is localized in the nucleus. This figure is modified from [Crane *et al.*, 2014].

activators, they have a signature dynamic response. Some characteristics of these response include the time spent inside the nucleus or the translocation speed. In certain instances even though the stimulus is at a constant level the pattern of the dynamic response could still encode information about the environment.

Pulsatile dynamics

A particular dynamical pattern observed in key transcription regulatory factors is represented by a series of on and off pulses that often appear stochastically, even when cells are maintained at constant conditions [Cai *et al.*, 2008]. In many cases the pulses

are asynchronous between cells, as a result the average dynamical behaviour of the population appears nearly constant, highlighting the importance of analyzing individual cell's dynamic responses [Dalal *et al.*, 2014]. Interestingly these pulsatile phenomena are pervasive across organisms, from bacteria to mammalian cells (see Fig 4.4).

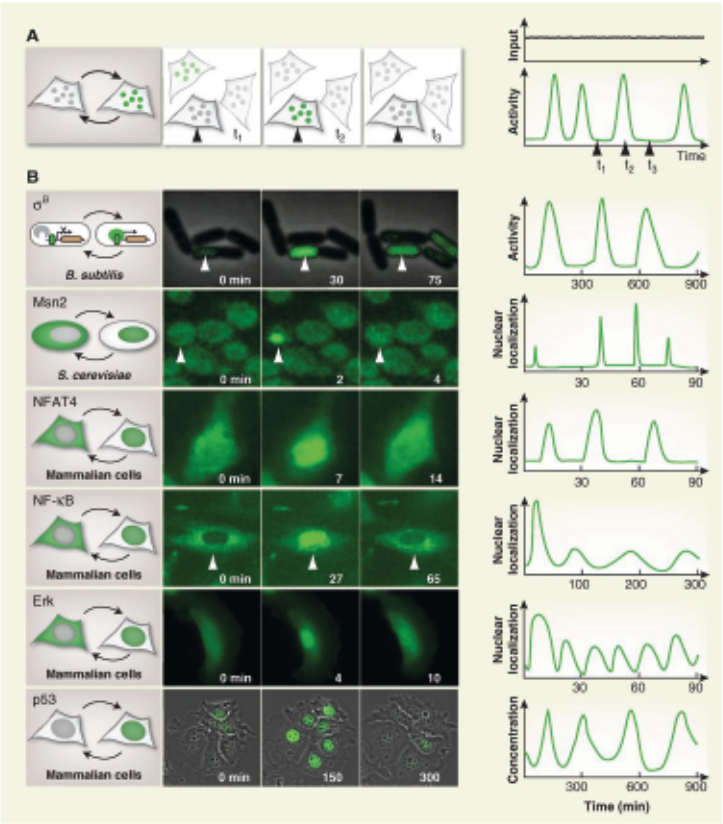

Figure 4.4: **Prevalence of pulsatile regulatory dynamics across species**. **A** Pulsing pattern involves transient simultaneous activation of molecules even under constant input. **B** Pulsing has been observed in many types of proteins and it involves several time scales. For each example, a schematic representation of the regulation is shown on the left, a typical microscopy image in the middle and a representation of the typical dynamical response on the right. This figure is reproduced from [Levine *et al.*, 2013].

Furthermore, in certain cases the pulse frequency encodes environmental states that consequently influence gene expression [Albeck *et al.*, 2013; Hafner *et al.*, 2017; Cai *et al.*, 2008]. Moreover it has been hypothesized that the pulses display certain advantages like randomizing sequences of cellular states, which in the context of bet-hedging may be helpful because it allows cells to dynamically control the distribution of states within the cell population [Levine *et al.*, 2013].

So far we have presented examples of dynamical patterns that contain information and how they influence cell response. In the next section we will quantify the information contained by these patterns in order to investigate the cell's internal representation of its environment.

4.2 Internal representation of environmental signals in yeast

This study was performed in collaboration with Gašper Tkačik, Alejandro Granados, Peter Swain, Julian Pietsch and Iseabail Farquhar, and was published in PNAS (see [Granados *et al.*, 2018]).

In this section we look at the transient responses of 10 yeast TFs to 3 stress conditions, all of which reduce growth (see Fig 4.5 A), and analyze their temporal pattern of nuclear translocation (see Fig 4.5 B). Based on the responses one can identify two groups: the generalists that respond to each stress (Dot6, Tod6, Msn2/4, Maf1 and Sfp1) and the specialists that respond either to one or two stress conditions (Mig1/2 - glucose depletion, Yap1 - oxidative stress and Hog1 - osmotic stress). By observing the temporal responses we can deduce some of their characteristics, for example, the delayed response and changed amplitude in the case of Dot6 could encode certain information about stress, nevertheless we can not specifically determine the amount of encoded information in a given condition. In order to understand which TFs are more informative of a given stressor, we estimate the mutual information between the presence or absence of a stressor and the temporal responses of the cells.

In each experiment are recorded a few hundreds of single cell responses $(100-300)$ of 20 dimensions, corresponding to observations of 50 min. For the information estimations in this study we use the linear SVM decoding method introduced in chapter 3. This involves a training process where we use 70% of the data to train a classifier,

which is used on the remaining testing data set to classify the time series into two groups (absence or presence of the stressor). The confusion matrix obtained from the testing set is then used to estimate the mutual information, although formally this is a lower bound to the real value, it is an estimate of the information that the cell could recover from observing a single time series. Additionally, by varying the duration of the responses (20-dimensional vectors) we can get estimates as a function of time and identify how quickly cells accumulate information.

The transition between the rich medium into a stress condition defines, in this case, two environmental states, therefore, the mutual information that we compute ranges between 0 and 1 bit[1], here 1 bit means that cells could perfectly distinguish the environmental state from the transcription factor dynamics and 0 bits correspond to the time series being indistinguishable among conditions. Using this measure we find that the glucose specialists Mig1/2 perform almost optimally in carbon stress (see Fig 4.5 C) and that the specialist Mig1 is the TF that accumulates information faster (~ 5 min). Moreover, we can rank the TFs according to their information content, interestingly the most accurate TFs are generally the fastest.

In general, high information content does not imply fast encoding, thus the delay in information encoding of certain TFs may arise as a consequence of the intracellular wiring, indicating in this case that, in carbon stress, the specialist Mig1/2 are activated through a different biochemical pathway than the other TFs.
If we define the encoding delay as the time in which half of the maximum information estimated is reached for each TF, we observe, that the specialists are the ones encoding the highest amount of information in the fastest manner (highlighted in blue in Fig 4.5 D, E, F), followed by the environmental stress response Msn2/4 (in pink), which in turn are followed by the others. This structure repeats throughout all stress conditions but the details are stress-specific suggesting that not only the presence but the identity of the stress is encoded by some TFs dynamic response.

[1]The maximum value (1 bit) corresponds to the entropy of the source, that is the logarithm of the number of environmental states $\log_2 2$.

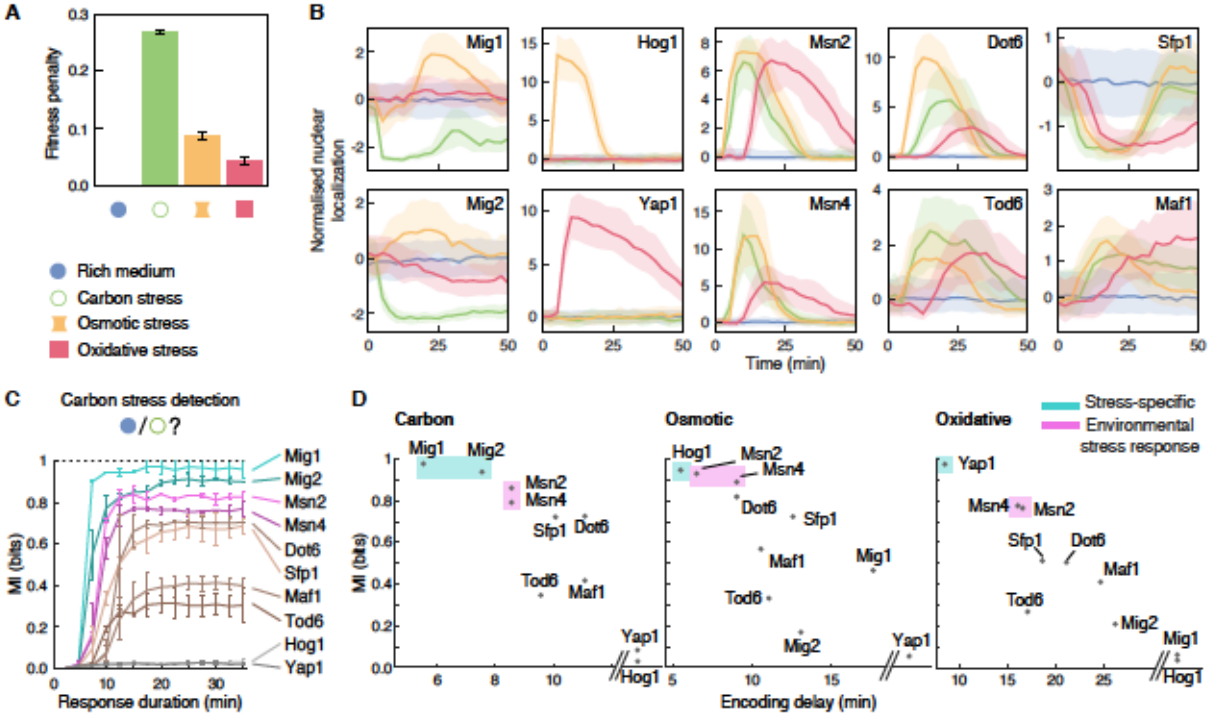

Figure 4.5: **Mutual information encoded in the nuclear translocation dynamics of 10 yeast transcription factors in four environmental conditions.** **A** The osmotic (0.4 M NaCl), carbon (0.1% glucose) and oxidative (0.5 mM H_2O_2) stressors reduce cell growth compared to the rich media (2% glucose). **B** Transient dynamic trajectories of ten TFs in response to a step like change from the rich media to the stressor at time $t = 0$. The solid lines correspond to the median normalized nuclear localization for each type of stress and shaded areas are the interquartile range over few hundreds of cells. In each panel the name of the TF is shown at the top corner and the color of the traces corresponds to each condition: rich medium in blue, carbon stress in green, osmotic stress in yellow and oxidative stress in red. **C** In response to carbon stress information estimates as a function of time, the errorbars show the standard deviation over two biological replicates. **D** TFs hierarchy regarding the information content and the encoding delay show that specialists encode more information and faster (in blue), they are followed by Msn2/4, the environmental stress response (in pink), and then followed by the other TFs. This figure is reproduced from [Granados *et al.*, 2018].

Increasing the complexity of the task, we consider four conditions instead of two (no stress, carbon, oxidative and osmotic stress) and estimate how much information

can provide each TF. In this case, the maximum possible mutual information is 2 bits; although no single TF can encode that much information, we find that the time series of Msn2/4 and Dot6 provide ~ 1.4 bits of information which is at least 30% higher than the best estimations from a single time point. Conversely, for the specialists Hog1 and Yap1, at least for such large stress levels, the difference between estimates from the full series and a single time point is practically negligible [Granados *et al.*, 2018]. However, the relevance of the specialists' dynamical signals is exhibited when a similar analysis is performed with 5 levels of stress intensity [Granados *et al.*, 2018].

So far we have extended and diversified the extracellular conditions and studied the information encoded by single TFs; however, single cells have available the information of multiple TFs at once. We now check whether transcription factors encode stress conditions collectively, thus before the estimation we concatenate the signals to be considered together and use those new vectors for estimation. For instance, in the case of two TFs, by concatenating the time series we obtain 40-dimensional vectors and estimate the information from the combined sequence. Furthermore, we estimate how much information is shared between pairs of TFs by comparing the mutual information encoded in a pair of TFs vs. each individual TF in the pair and define a measure of information redundancy $r = 1 - I_{1,2}/(I_1 + I_2)$, where I_i is the mutual information calculated individually and $I_{1,2}$ is the mutual information calculated from the pair (the concatenated data).
The information redundancy matrix between all pairs, where we consider 4 extracellular conditions (no stress and the 3 stresssors), allows us to build a network, where we expect specialists to share scarce information between them and generalists to contain higher levels of information redundancy. This representation helps understanding the upstream regulation of TFs, where indeed the generalists display higher redundancy among themselves compared to the specialists, indicating that they share a common upstream signaling (Fig 4.6 B). Additionally, this pattern overlaps with the network built using data from kinase substrates [Sharifpoor *et al.*, 2011] in Fig 4.6 A.

Out of the 45 possible pairs, we observe that the ones formed by a specialist and a generalist are mainly more informative than the other pairs in encoding the identity

of a stressor (see Fig 4.6 C). However, nothing prevents the cell to use more than two signals to encode complex environmental conditions. In fact, we observe that as the environment increases in complexity, more TFs are needed (Fig 4.6 D), which means that the collective dynamical response carries stress-specific and detailed information.

In this section, we estimate information with the linear SVM decoding method, because for transient responses after the switch between conditions it performs equally good as the SVM kernelized methods [Granados *et al.*, 2018]. However, nonlinear dynamical encoding, like in pulsatile dynamics, may be estimated better with the methods that capture higher order statistical features of dynamical signals. In section 4.3, we will show how the family of decoding based methods perform on transient and stationary responses and in the regime of small number of samples.

4.3 Decoding-based information estimators on experimental data

In this section, we focus on the evaluation of the methods introduced in chapter 3. While in section 4.2 we use the method based on SVM linear classification to ask biologically relevant questions about the internal representation of extracellular signals, here we study in detail the performance of the methods using experimental data.

To illustrate the use of our estimators in a realistic context, we analyzed data from two previously published papers. The first paper focused on the representation of environmental stress in the nuclear localization dynamics of several transcription factors (here we focus on data for Msn2, Dot6, and Sfp1; known to exhibit pulsatile dynamics even in the absence of stress) in budding yeast [Granados *et al.*, 2018]. The second paper studied information transmission in biochemical signaling networks in mammalian cells (here we focus on data for ERK and Ca^{2+}) [Selimkhanov *et al.*, 2014]. In both cases, single-cell trajectory data were collected in hundreds or thousands of single cells sampled at sufficient resolution to represent the trajectories discretized at tens to hundreds of timepoints. Similarly, both approaches estimate the information transmis-

sion in trajectories about a discrete number of environmental conditions: Ref [Granados *et al.*, 2018] uses the linear SVM approach presented here, while Ref [Selimkhanov *et al.*, 2014] uses the knn estimator. This makes the two datasets perfectly suited for estimator comparisons. We further note that in both datasets the trajectories can be divided into two time epochs: the early "transient" period when the external condition changes, and the late "near steady-state" period. Typically, the transient exhibits very clear differences in the trajectory means between various conditions, reminiscent of our Example 1 or early Example 2 (in section 3.3.1); in contrast, in the late period the response may have been adapted away, or the stimulus could be encoded only in higher-order statistics of the traces, reminiscent of the late period in Example 2 or Example 3 in section 3.3.1.

Fig 4.7 shows the raw data and summarizes our estimation results for the early and late epochs for the three translocating factors in yeast that report on the change from 2% glucose rich medium to 0.1% glucose poor stress medium. Fig 4.8 similarly shows the raw data and estimation results for the early and late epochs for the signaling molecules in mammalian cells responding to multilevel inputs.

Consistent with the published report [Granados *et al.*, 2018], transient response in yeast nuclear localization signal can be decoded well with the linear SVM estimator that yields about 0.6 bits of information per gene about the external condition. Kernelized SVM outperforms the linear method slightly by extracting an extra 0.1-0.2 bits of information, while knn underperforms the linear method significantly for Msn2 and Dot6 (but not for Sfp1). Gaussian decoder estimate shows a mixed performance and the neural network estimate is the worst performer, most likely because the number of samples here is only $N = 100$ per input condition and neural network training is significantly impacted.

It is interesting to look at the stationary responses in yeast which haven't previously been analyzed in detail. First, very low estimates provided by linear SVM for Msn2 and Dot6 imply that information in the stationary regime, if present, cannot be extracted by the linear classifier. Second, Gaussian decoder also performs poorly in the stationary regime, potentially indicating that the relevant features are encoded in higher-than-pairwise order statistics of the response (e.g., pulses could be "sparse" features as in sparse coding [Olshausen and Field, 2004]); it is, however, hard to exclude small num-

ber of training samples as the explanation for the poor performance of the Gaussian decoder. Third, K-nearest-neighbor estimator also yields low estimates, either due to small sample number or low signal-to-noise ratio, the regime for which knn method has been observed to show reduced performance [Khan *et al.*, 2007]. A particularly worrying feature of the knn estimates is their non-robust dependence on the length of the trajectory T. As Fig 4.9 shows, the performance of knn peaks at $T \approx 50$ min and then drops, even well into unrealistic negative estimates for $T \approx 400$ min (corresponding to the highest dimensionality $d = 170$ of discrete trajectories). While it is possible to make an *ad hoc* choice to always select trajectory duration at which the estimate peaks, the performance of kernelized SVM is, in comparison, extremely well behaved and increases monotonically with T, as theoretically expected. Finally, nonlinear SVM estimator extracts up to 0.4 bits of information about condition per gene, more than half of the information in the early transient period. This is even though on average the response trajectories for the two conditions for Msn2 and Dot6 are nearly identical. For Sfp1 there is a notable difference in the mean response, which the linear estimator can use to provide a ~ 0.15 bits of information, yet still significantly below ~ 0.4 bits extracted by the nonlinear SVM. For both transient and stationary responses in yeast, our results are qualitatively in line with the expectations from the synthetic example cases in section 3.3.1, Fig 3.2 given the small number of trajectories, tightest and most robust estimates are provided by the decoding information estimator based on nonlinear (kernelized) SVM. Regardless of the decoding methodology and even without small sample corrections at $N = 100$ trajectories per input, our estimates are not significantly impacted by the well-known information estimation biases thanks to the dimensionality reduction that decoding provides by mapping high dimensional trajectories $\mathbf{X}$ back into the space for inputs U which is low dimensional; this is verified in Fig 4.10 by estimating the (zero) information in trajectories whose input labels have been randomly assigned.

Random pulses that encode stationary environmental signals have been observed for least in 10 transcription factors in yeast [Dalal *et al.*, 2014] and for tens of transcription factors in mammalian cells [Levine *et al.*, 2013]. Recent studies investigated the role of the pulsatile dynamics in cellular decision-making [Albeck *et al.*, 2013; Hafner *et al.*, 2017]. Nevertheless, methods for quantifying the information encoded in stochastic pulses are still in their infancy. Our nonlinear SVM decoding estimates

convincingly show that there is information to be learned at the single cell level from the stationary stochastic pulsing. An interesting direction for future work is to ask whether hand-crafted features of the response trajectories (pulse frequency, amplitude, shape, etc) can extract as much information from the trajectories as the generic SVM classifier: for that, one would construct for each response trajectory a "feature vector" by hand, compute the linear SVM decoding bound information estimate from the feature vectors, and compare that to the kernelized SVM estimate over the original trajectories. This approach is a generic and operationally-defined path for finding "sufficient statistics" of the response trajectories—or a compression of the original signal to the relevant set of features—in the information-theoretic sense.

A different picture emerges from the mammalian signaling network data shown in Fig 4.8. The key difference here is the order of magnitude larger number of sample trajectories per condition compared to yeast data. Most of the information seems linearly separable in both the early and late response periods, as evidenced by the success of the linear SVM based estimator whose performance is not improved over by the kernelized SVM (indeed, for early ERK epoch linear SVM gives a slightly higher estimate than the nonlinear version). The big winner on this dataset is the neural-network-based estimator that yields best performance in all conditions among the decoding-based estimators, likely owning to sufficient training data. As before, gaussian decoder shows mixed performance which can get competitive with the best estimators under some conditions. Lastly, knn appears to do very well except on the late Ca^{2+} data (perhaps due to low signal-to-noise ratio). It also shows counter-intuitive non-monotonic behavior with trajectory duration T in Fig 4.11. Once again it is worth keeping in mind that knn is estimating the full mutual information which could be higher than the information decodable from individual responses.

4.4 Conclusions

In the analysis of the transient responses of transcription factors to extacellular step-like changes, we show that the information can be linearly decodable from the transcription factors' nuclear translocation dynamics to almost perfectly distinguish high levels of stress from the rich media. Multiple stressors can be identified with some degree

of error by the transcription factors identified as generalists, and complementarily, the specialists can better distinguish the intensity of the corresponding stress and do it so faster. Nevertheless, no single transcription factor can accurately encode both the identity and strength of the stressors. Furthermore, our results show that multiple dynamical signals of transcription factors can encode better complex stimuli.

Additionally, the study in section 4.3 shows that decoding-based estimators in many cases perform better than the knn estimator, especially with typical problem dimensions ($d \sim 1 - 100$) and typical number of sample trajectories ($N \sim 10^2 - 10^3$). This is especially true when we ask about the combinatorial representation of the environmental state in the time trajectories of several jointly observed chemical species, as in our initial analysis [Granados *et al.*, 2018], where alternative information estimation methods usually completely fail due to the high dimensionality of the input space. For problems in the low data regime (small N), linear or kernelized SVM approaches appear very powerful, while at larger N neural-network-based schemes can provide a better performance and thus a tighter information lower bound. Our results show that the pulsatile dynamics of Msn2 and Dot6 contain information about the environmental state, this is only decoded with the kernelized SVM methods, despite the limited number of samples (~ 100) and the high similarity between the mean population responses between conditions.

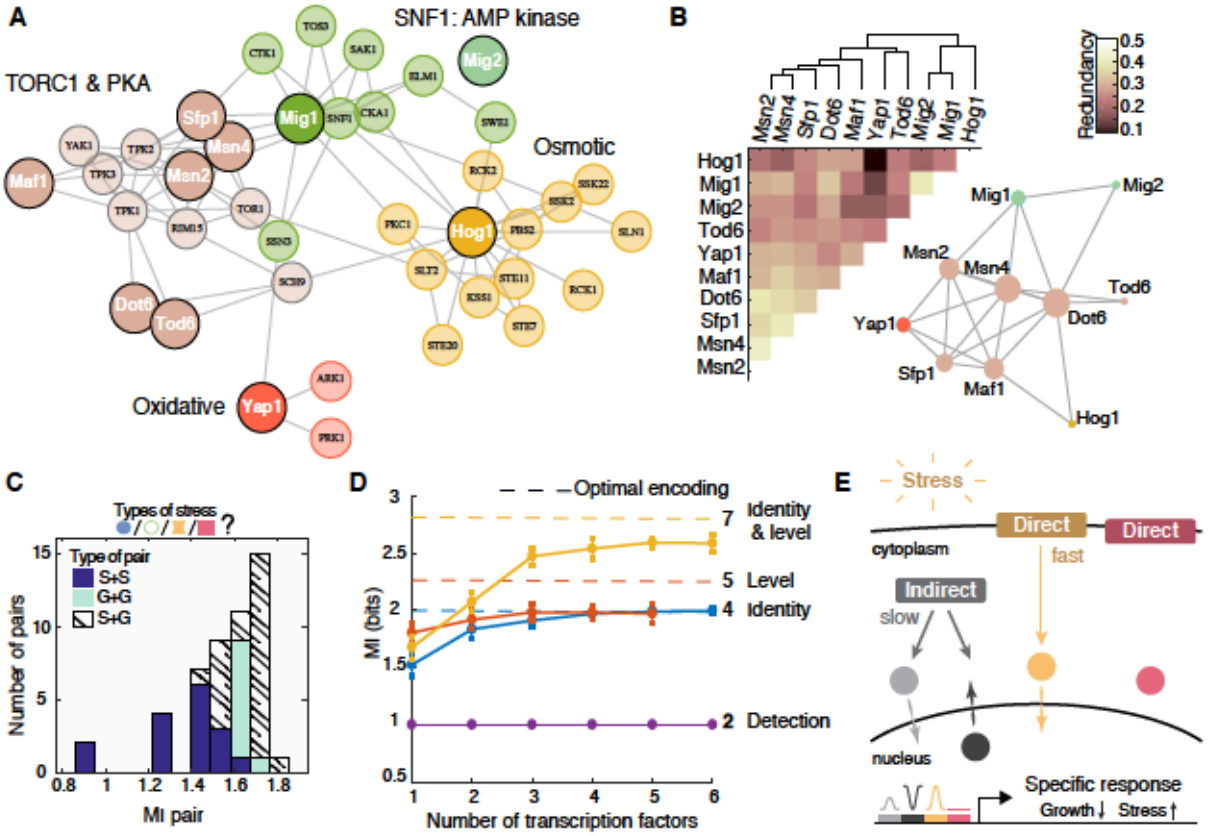

Figure 4.6: **Complex environments can be encoded collectively by several transcription factors. A** A representation of the intracellular signaling network, where the edges between kinase and substrate are proportional to the evidence for that interaction [Sharifpoor *et al.*, 2011]. **B** The information redundancy matrix for all pairs of transcription factors; displayed as a network reflect intracellular signaling features, the edges thickness are proportional to a pair's redundancy and the size of each node increases with the number of edges. **C** Pairs formed by a specialist and a generalist (S+G) typically encode more information with respect to all possible pairs. **D** Complex environmental conditions can be encoded by several TFs. From all possible combinations of concatenated signals the solid lines report the maximum estimate for a given number of TFs. The colors illustrate the number of states: purple, two states (*detection* of the presence or absence of the stress); blue, four states (*identification* of the stressors among 4 conditions); red, five states (distinction among 5 *levels* of the same stress) and orange, seven states (differentiation of the identity and levels of the stress conditions). **E** Cells transduce information through two types of channels: the generalists and the specialists. This figure is reproduced from [Granados *et al.*, 2018].

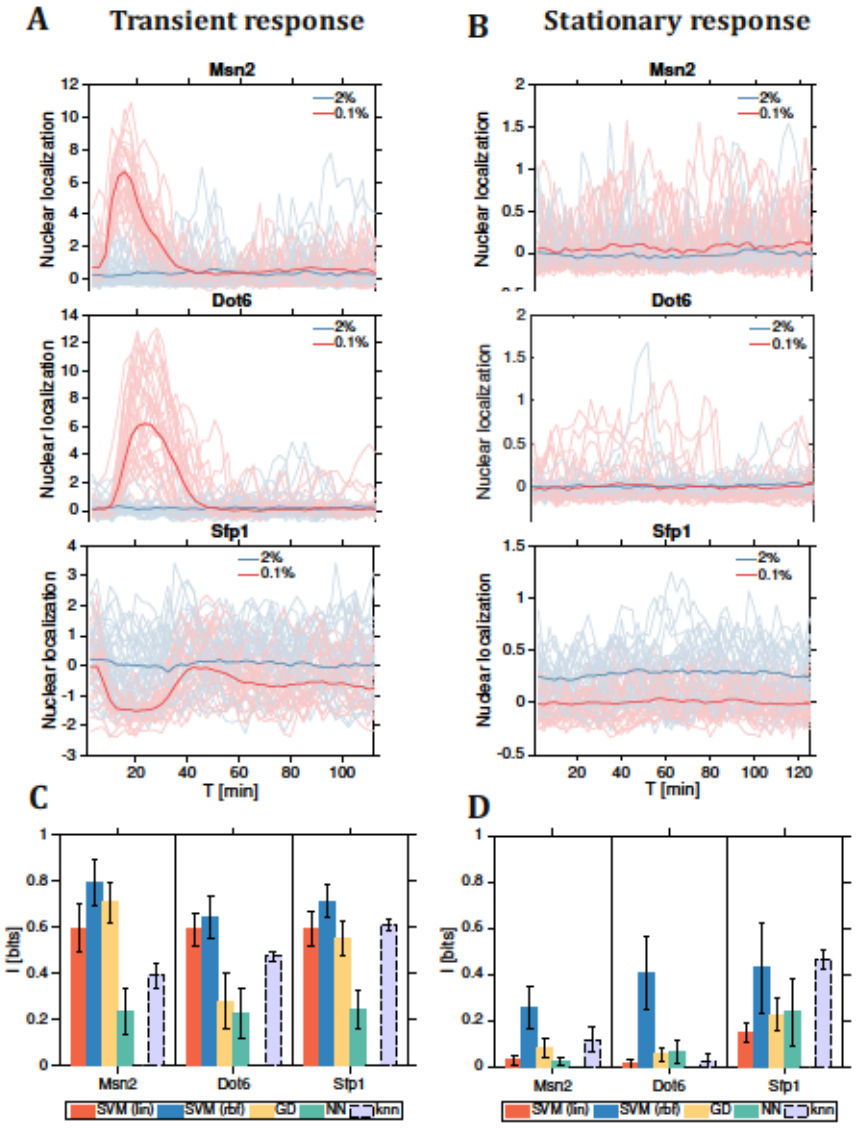

Figure 4.7: **Two-level mutual information estimates from single-cell time-series data for nuclear translocation of yeast transcription factors. (A, B)** Data replotted from Ref [Granados *et al.*, 2018] for Msn2 (top row), Dot6 (middle row), and Sfp1 (bottom row); early transient responses (A) after nutrient shift at $t = 0$ min from glucose rich (2%, blue traces) to glucose poor (0.1%, red traces) medium are shown in the left column, stationary responses (B) are collected after cells are fully adapted to the new medium. Sampling frequency is 2.5 min, $d = 45$, and the number of sample trajectories per nutrient condition is $N = 100$. Thin lines are individual single cell traces, solid lines are population averages. **(C, D)** Information estimates for the transient (left, C) and stationary (right, D) response periods. Colored bars use model-free decoding-based estimators as indicated in the legend, gray bar is the knn estimate; error bars computed from estimation bootstraps by randomly splitting the data into testing and training sets.

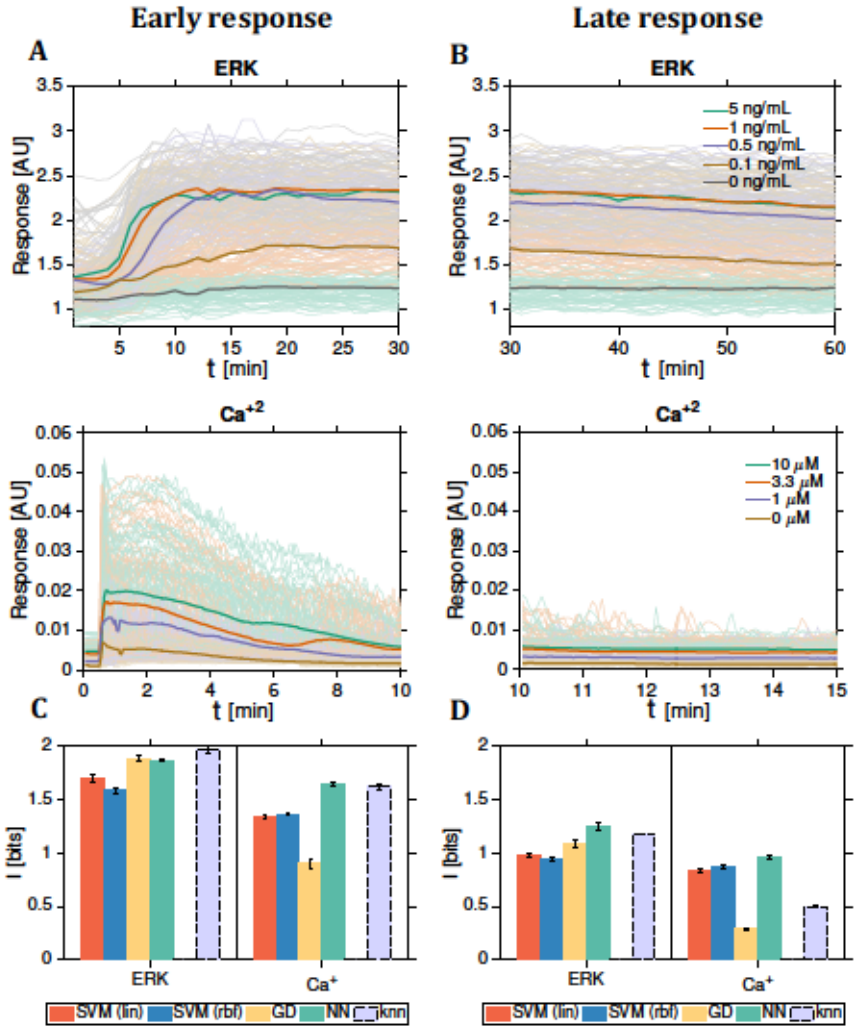

Figure 4.8: **Multilevel mutual information estimates from single-cell time-series data for mammalian intracellular signaling.** Data replotted from Ref [Selimkhanov *et al.*, 2014] for ERK (top row) and Ca^+ (bottom row). **(A)** Early transient responses after addition of 5 different levels of EGF for ERK (or 4 different levels of ATP for Ca^+, respectively) at $t = 0$ min, as indicated in the legend. **(B)** In the late response most, but not all, of the transients have decayed. Data for ERK: $N = 1678$ per condition, $T = 30$ min ($d = 30$) for early response and $T = 30$ min ($d = 30$) for late response. Data for Ca^+: $N = 2995$ per condition, $T = 10$ min ($d = 200$) for early response and $T = 5$ min ($d = 100$) for late response. Plotting conventions as in Fig 4.7.

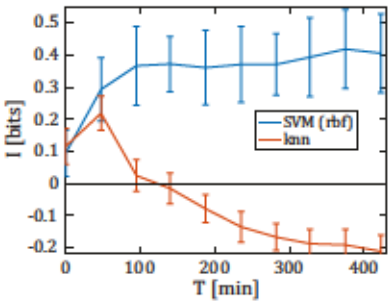

Figure 4.9: **Estimator behavior for longer trajectory data for Dot6.** When the samples are limited, here to $N = 100$ samples per input glucose level condition as in Fig 4.7A (middle), radial-basis-function SVM estimate (blue) is well-behaved with no observable overfitting and consequent drop in information estimate as the trajectory duration, T, is increased (maximal T corresponds to $d = 170$ dimensional trajectory vectors). In contrast, knn estimate (brown) shows a collapse in the estimation performance, even yielding strongly negative numbers, as the dimensionality of input vectors is increased at fixed number of trajectory samples.

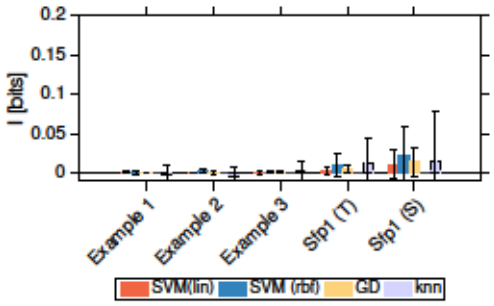

Figure 4.10: **Assessing information estimation bias due to small sample size.** By randomly shuffling the binary labels assigned to different response trajectories, we break all response-input correlations leading to zero information. Here we test whether our estimators correctly report zero information within error bars given a finite number of samples, or are subject to positive information estimation bias. Decoding-based estimates (linear SVM, red; kernelized SVM, blue; Gaussian decoder, yellow) and knn (gray). First three sets of bars correspond to synthetic examples of Fig 3.3; estimations are done with $d = 100$ and $N = 1000$ per input condition as in Figs 3.5 and 3.6, following the same plotting conventions. Last two sets of bars are estimated with $N = 100$ per input condition using real data for Sfp1 yeast TF from Fig 4.7A. In all cases, even without explicit small-sample debiasing for Eq (3.26) (which may be required for multilevel estimation), the estimates are consistent with zero.

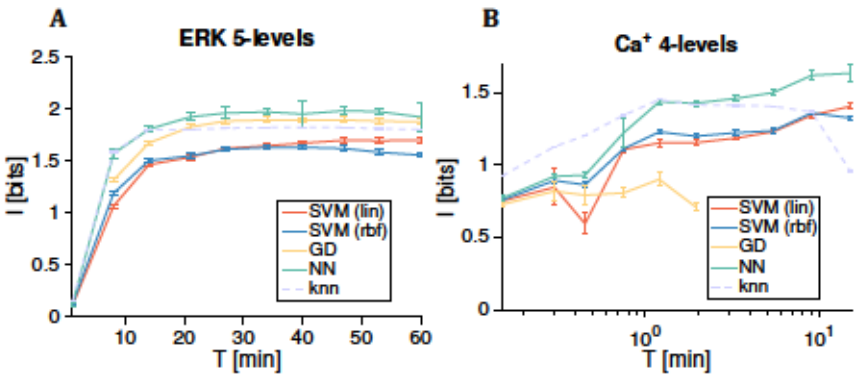

Figure 4.11: **Information estimates for mammalian signaling networks as a function of the trajectory duration.** Shown are information estimates as a function of the total trajectory duration, T, for the early response period for ERK (left) and Ca^+ (right). Plotting conventions, procedures, and data set sizes same as in Fig 4.8.

5 Crosstalk and kinetic proofreading in transcriptional regulation

The work presented in this chapter was conducted jointly with Gašper Tkačik and Georg Rieckh. It has been published in the Physical Review Letters (see [Cepeda-Humerez *et al.*, 2015]) and is reproduced here with minimal changes.

Abstract

Gene expression is controlled primarily by interactions between transcription factor proteins (TFs) and the regulatory DNA sequence, a process that can be captured well by thermodynamic models of regulation. These models, however, neglect regulatory crosstalk: the possibility that non-cognate TFs could initiate transcription, with potentially disastrous effects for the cell. Here we estimate the importance of crosstalk, suggest that its avoidance strongly constrains equilibrium models of TF binding, and propose an alternative non-equilibrium scheme that implements kinetic proofreading to suppress erroneous initiation. This proposal is consistent with the observed covalent modifications of the transcriptional apparatus and predicts increased noise in gene expression as a trade-off for improved specificity. Using information theory, we quantify this trade-off to find when optimal proofreading architectures are favored over their equilibrium counterparts. Such architectures exhibit significant super-Poisson noise at low expression in steady state.

5.1 Introduction

While noise in gene expression has been studied extensively, a question that has received considerably less attention is that of crosstalk in gene regulation. By *crosstalk* we mean the possibility that the gene may be induced (or repressed) by the erroneous binding of a non-cognate transcription factor, and the interference of such factors with the transcriptional machinery (e.g., RNA polymerase). This problem may be particularly acute in eukaryotes. Recent studies on this topic quantify the limits that crosstalk places on the regulatory system [Friedlander *et al.*, 2016; Carballo-Pacheco *et al.*, 2018]. Here, we will consider molecular mechanisms that can cope with crosstalk. One such class of plausible mechanisms that we will propose and analyze here involves coupling the initiation of transcriptional regulation to an energy source to keep the regulatory system out-of-equilibrium and thus able to reject erroneous attempts at initiating gene transcription, beyond what is possible by any molecular machine operating at thermodynamic equilibrium.

Transcriptional regulation in eukaryotic cells is a complex procedure involving a multitude of molecular steps, the details of which are currently an active area of research [Phatnani and Greenleaf, 2006; Saunders *et al.*, 2006]. Binding sites of TFs in eukaryotic cells are considerably shorter than in prokaryotes (~ 10 base pairs compared to typical prokaryotic lengths of $15-20$ basepairs, with mismatch penalties similar in both kingdoms, given by the typical scale of hydrogen bonds of $2-3\ k_BT$), implying that the equilibrium occupancy ratio of a specific site for a typical TF in an eukaryote is favored by a factor of $10^3 - 10^4$ relative to a non-specific site. On the other hand, the number of genomic locations where TFs can bind in eukaryotes exceeds that of bacteria by roughly 10^3. In addition, eukaryotes have roughly 10 times more different types of TFs, many of which are descended from common ancestor proteins that share DNA recognition sequences [Milo and Philips, 2016]. This rises the question of how eukaryotic cells achieve high specificity [Todeschini *et al.*, 2014]: the ability to correctly activate the desired genes by low concentrations of the cognate TFs, and simultaneously the ability to not activate the genes that are supposed to remain silent.

The large number of sequences nearly identical to the consensus sequence, some of which occur in promoters or enhancers of non-cognate genes, implies that the proba-

bility of crosstalk should be large. At first glance the mechanism of cooperative binding would appear to solve the problem but recent studies show that variants of cooperativity and combinatorial regulation are not enough to mitigate the crosstalk problem [Friedlander *et al*., 2016]. While this might help where cognate and non-cognate TFs compete for the binding site, it does not seem to help when the cognate TFs are absent, for instance, when the gene should be left un-induced; without the cognate TFs around, the gene could easily be erroneously activated by crosstalking TFs.

A similar question has been addressed in a very different biological setting by John Hopfield [Hopfield, 1974] and by Jacques Ninio in 1975 [Ninio, 1975]. They tried to reconcile the observed high specificity of certain biosynthetic reactions, such as DNA replication, protein synthesis etc, with the bounds that equilibrium thermodynamics places on such specificity, given the known recognition energies between the cognate vs non-cognate reactants. Their results suggested that the observations can only be reconciled with theory if the recognition process is out-of-equilibrium. Thus they proposed concrete reaction schemes that could achieve such a high specificity; these are collectively known as *kinetic proofreading models* (KPMs). The basic predictions from these schemes were later experimentally confirmed. Early proofreading models were devised purely to maximize specificity, without considering other trade-offs that the system might face, such as noise or reaction speed. Subsequent research has addressed some of these trade-offs [Savir and Tlusty, 2013] but not within the context of information theory. In particular, this has not been attempted in transcriptional regulation, which is often cited as a textbook example of an equilibrium regulatory process that exhibits high specificity.

5.1.1 Kinetic proofreading in transcriptional regulation and the two state model

The simplest model for transcriptional gene regulation is the so-called two state model. The promoter of a given gene can be in an OFF state (not transcribing the gene), or in an ON state (transcribing the gene into messenger RNA). The promoter switches stochastically between the two states and the rate of switching between the two states is determined by the concentration of the relevant transcription factor. Molecularly,

these states could be represented by a promoter with a bound activating transcription factor and an empty promoter, although other molecular arrangements mapping into the same ON/OFF architecture exist.

To extend such a model to handle the crosstalk due to binding of noncognate TFs, we need to add at least one more state – that of a promoter with the noncognate TF bound, which is consequently transcribing the gene erroneously. This yields what we call the three-state model (TSM). In equilibrium, the two-state and the three state models must obey detailed balance.

Finally, to extend these molecular schemes to out-of-equilibrium situations capable of kinetic proofreading, we propose the 5-state transition diagram depicted in Fig. 5.1**B**. This is the simplest model useful for theoretical study and not intended to mimic any particular known gene regulatory element in detail; it is useful to think of this model in terms of a promoter with a single binding site for an activating TF.

State (0) is the *empty* state of the promoter, (1_c) and (1_{nc}) are intermediate *occupied* states (with the cognate and non-cognate TF bound, respectively) and $(2_{c,nc})$ are *active* states, from which transcription of mRNA proceeds at rate r. Molecules of mRNA are decaying with rate d, the slowest timescale in the problem, and $1/d$ sets the unit in which we will express all other timescales. The cognate TFs are present at concentration c_c, have a binding-rate (to the occupied state) k_+, a forward-rate (to the active state) $1/q$ and off-rate k_-^c. Acting as noncognate TFs, we assume that there are ν other species of TFs present in the system, each at concentration c_{nc}, with the rates $1/q$, k_+, being the same as for the cognate case, and a faster off-rate k_-^{nc}.

The number of noncognate TF species ν and the affinity ratio $\sigma = k_-^{nc}/k_-^c$ define the relevant parameter $\Lambda = \nu/\sigma$, which determines how important the crosstalk is: for $\Lambda \ll 1$, crosstalk is not important, for $\Lambda \sim 1$, the equilibrium model would have roughly equal occupancy of the binding site by the cognate and one of the noncognate TFs, and for $\Lambda \gg 1$, the crosstalk constraint is dominating the problem.

The specificity of cognate vs noncognate TFs to the promoter is carried solely by the difference in off-rates (unbinding rates), as in Hopfield's original proposal, consistent with the simple picture of diffusion limited binding reactions and different binding energy of cognate and noncognate TFs. The key characteristic of the model are the irreversible transitions from occupied to active states: with a properly chosen forward

transition rate, the equilibrium specificity σ of a two-state model can be squared in the proofreading scenario. To realize such irreversible transitions, thermodynamics requires the energy to be spent, typically by coupling the transition to another reaction proceeding down a strong thermodynamic gradient, e.g., hydrolysis of ATP. Concretely, the mechanism could hydrolize one molecule of ATP to put a phosphorylation tag on the appropriate state of the transcriptional machinery (e.g., on RNA Poly CTD tail when RNA Poly is simultaneously bound with the TF).

5.1.2 Erroneous induction as a novel noise source (semantic noise)

Traditionally, in proofreading scenarios, what has been studied is the "error fraction": the ratio between the number of incorporations of the erroneous substrate vs the number of incorporations of the cognate substrate. In our case, it is interesting to observe that the "error" product and the "correct" product are the same mRNA molecule – "error" and "correct" here rather refer to whether the same molecule has been produced in response to the correct signal or not, such that the error fraction η has the form:

$$\eta = \frac{mRNA(Erroneous\ Initiation)}{mRNA(Correct\ Initiation)}. \tag{5.1}$$

Because both the correct and erroneous pathways end with the same product in our case, crosstalk can, interestingly, be viewed as an additional noise source in transcriptional regulation: binding of non-cognate TFs will induce stochastic production of mRNA even in the absence of cognate TF signals. This is important, because to quantify the regulatory power of genetic regulatory elements using information theory, one needs to compute not only about their average response to the signal (which can be analyzed using deterministic rate equations, as in Hopfield's original work), but also their noise characteristics. Several sources of noise in gene regulation have been studied, the most relevant and well identified are the extrinsic and intrinsic noise sources [Swain *et al.*, 2002], or, in an alternative classification, the output, switching and diffusion noise [Tkačik *et al.*, 2008b] contributions. The reaction scheme of Fig 5.1 **B** allows us to compute the contributions from crosstalk, switching (due to stochastic transitions between promoter states), and output (due to birth-death expression of mRNA) noise, and ask how much information can be transmitted from the signal, i.e., the input cognate TF concentration c_c, to the mRNA expression level m.

An interesting aspect of this problem, as compared to traditional examples of kinetic proofreading, is that there is no need for a separate "utility" function: the erroneous induction manifests directly as increased noise and decreased dynamic range, and can therefore be captured directly by a decrease in information transmission, without any further assumptions.

5.2 Results

In prokaryotes, transcription factors recognize and bind specific DNA sequences $L = 10 - 20$ base-pairs (bp) in length, usually located in promoter regions upstream of the regulated genes [Ptashne and Gann, 2002]. Regulation by a single TF, or a small number of TFs interacting cooperatively, is sufficient to quantitatively account for the experimental measurements of gene expression [Kuhlman *et al.*, 2007], as well as to explain how any gene can be individually "addressed" and regulated only by its cognate TFs [Wunderlich and Mirny, 2009], without much danger of regulatory crosstalk. In eukaryotes, however, TFs seem to be much less specific ($L = 5 - 10$ bp, perhaps due to evolvability constraints [Tuğrul *et al.*, 2015]; but the total genome size is larger than in prokaryotes by $\sim 10^3$) [Wunderlich and Mirny, 2009; Sandelin *et al.*, 2004], binding promiscuously to many genomic locations [Li *et al.*, 2008], including to their non-cognate binding sites [Rockel *et al.*, 2013]. What are the implications of this reduced specificity for the precision of gene regulation?

Thermodynamic models of regulation postulate that the rate of target gene expression is given by the equilibrium occupancy of various TFs on the regulatory sequence [Shea and Ackers, 1985; Bintu *et al.*, 2005], and the success of this framework in prokaryotes [Kinney *et al.*, 2010] has prompted its application to eukaryotic, in particular, metazoan, enhancers [Janssens *et al.*, 2006; He *et al.*, 2010; Fakhouri *et al.*, 2010]. To illustrate the crosstalk problem in this setting, consider the ratio σ of the dissociation constants to a nonspecific and a specific site for an eukaryotic TF; typically, $\sigma \sim 10^3$ (corresponding to a difference in binding energy of $\sim 7\, k_B T$) [Maerkl and Quake, 2007; Rockel *et al.*, 2013]. Because there are $\nu \sim 10^2 - 10^3$ of different TF species in a cell, TFs nonspecific to a given site will greatly outnumber the specific ones. For an isolated binding site, this would imply roughly equal occupancy by cognate and noncognate

TFs, suggesting that crosstalk could be acute. For multiple sites, cooperative binding is known for its role in facilitating sharp and strong gene activation even with cognate TFs of intermediate specificity—but could the same mechanism also alleviate crosstalk? First, there exist well-studied TFs which do not bind cooperatively (e.g. [Giorgetti *et al.*, 2010]). Second, to reduce crosstalk, cooperativity needs to be strong and specific, stabilizing only the binding of *cognate* TFs [Friedlander *et al.*, 2016]; many proposed mechanisms lack such specificity (e.g. [Mirny, 2010; Todeschini *et al.*, 2014]). Third, even when cooperative interactions are specific, crosstalk can pose a serious constraint. Regulating a gene implies varying the cognate TF concentration throughout its dynamic range, and when this concentration is low and the target gene should be uninduced, cooperativity cannot prevent the erroneous induction by noncognate TFs. For that, the cell could either keep the genes inactive by binding of specific repressors, or by making the whole gene unavailable for transcription. The first strategy seems widely used in bacteria but less so in eukaryotes; the second strategy ("gene silencing") is widespread in eukaryotes, but only happens at a slow timescale and involves a complex series of nonequilibrium steps.

Here we propose a plausible and fast molecular mechanism which alleviates the effects of crosstalk; a detailed account of when crosstalk poses a severe constraint for gene regulation will be presented elsewhere [Friedlander *et al.*, 2016]. The proposed mechanism is consistent with the known tight control over which genes are expressed in different conditions or tissues (e.g., during development [McGinnis and Krumlauf, 1992]) on the one hand, and on the other, explains the high levels of measured noise in transcription initiation of active genes [Raj *et al.*, 2006; Little *et al.*, 2013].

The simplest proofreading architecture for transcriptional gene activation that can cope with erroneous binding is presented in Fig 5.1**A**,**B**, motivated by a scheme first proposed by Hopfield [Hopfield, 1974]. Specificity is only conveyed by differential rates of TF unbinding ("off-rates" $k_-^{\mathrm{c}}, k_-^{\mathrm{nc}}$, with $\sigma = k_-^{\mathrm{nc}}/k_-^{\mathrm{c}}$). There are ν noncognate TF species whose typical concentration we take to be $c_{\mathrm{nc}} = \frac{1}{2}\nu C$, and C is the maximal concentration for the cognate TFs c_{c}, $c_{\mathrm{c}} \in [0, C]$. The ratio $\Lambda = \nu/\sigma$ determines the severity of crosstalk, which is weak for $\Lambda \ll 1$ and strong for $\Lambda \gg 1$. The response of the promoter to the dimensionless input concentration c ($= k_+ c_{\mathrm{c}}/d$, see Fig 5.1**B**) of cognate TFs is captured by the steady state distribution of mRNA, $P(m|c)$; the spread

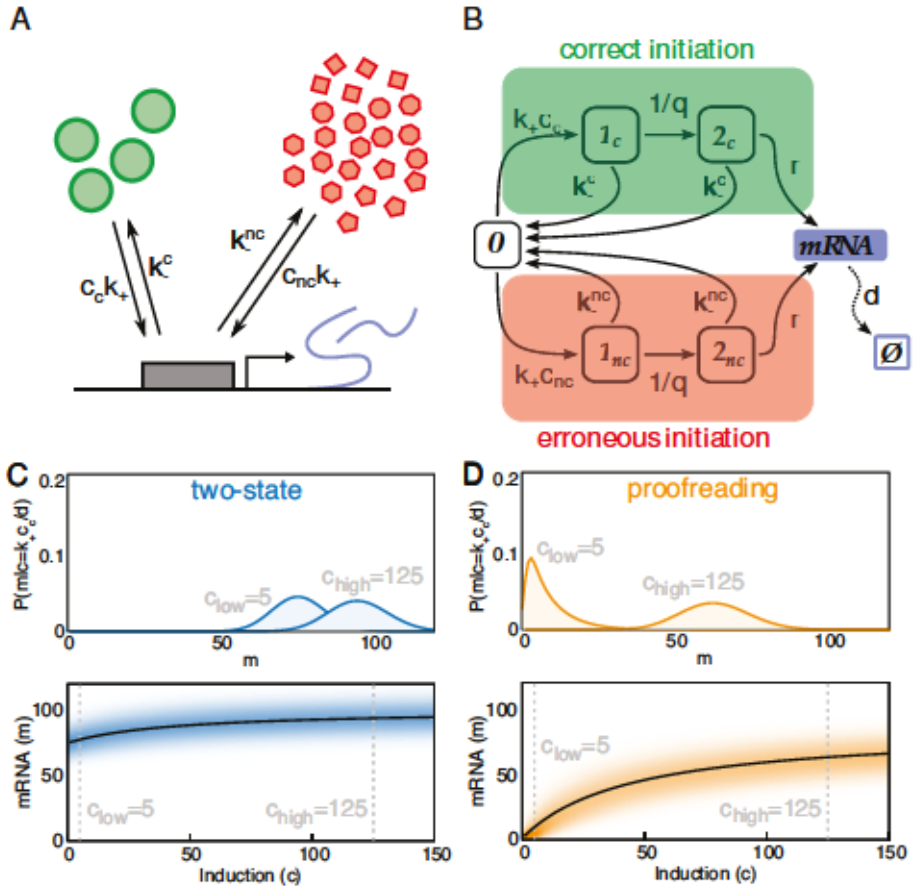

Figure 5.1: **A)** A schematic of cognate (green circles) and ν kinds of noncognate (various red shapes) TFs binding to a gene regulatory element on the DNA (gray box), to control the mRNA expression level. **B)** Transition state diagram for the proofreading gene regulation. The regulatory element can cycle between an empty state (0), state occupied by either cognate (1_c) or noncognate (1_{nc}) TF; to initiate gene expression, a further non-equilibrium transition into "2" states (with rate $1/q$) is required, driven by, e.g., hydrolysis of ATP. mRNA is expressed at rate r and degraded with rate d, the slowest process that sets our unit for time. In this figure we use $r/d = 100, k_-^{nc}/d = 2500, \sigma = 500, \nu = 50, \Lambda = \nu/\sigma = 0.1$; dimensionless concentration is $c = k_+ c_c/d$. **C,D)** Steady-state mRNA distributions for low and high concentrations of the cognate TF, c. As $qd \to 0$ (C), the proofreading model reduces to the two-state model of gene expression [Rieckh and Tkačik, 2014]; here, noncognate TFs initiate transcription at a high rate even when c is low, causing overlapping output distributions (blue; top) and small dynamic range (black line = $\langle m(c) \rangle$, blue shade = $\sigma_m(c)$; bottom). Proofreading (D) suppresses erroneous initiation, leading to separable output distributions (orange; top) and higher dynamic range (bottom).

of this distribution is due to the stochasticity in gene expression, which includes random switching between promoter states and the birth-death process of mRNA expression [Peccoud and Ycart, 1995]. If the reaction rates are known, $P(m|c)$ is computable from the Chemical Master Equation corresponding to the transition diagram in Fig 5.1**B**; using finite-state truncation, this becomes a linear problem that is numerically tractable.

Figs 5.1**C** and **D** each compare the steady state distributions of mRNA at low and high concentration of cognate TF, c. The behavior crucially depends on the out-of-equilibrium rate qd. When $qd \to 0$, the scheme of Fig 5.1**B** becomes a normal two-state promoter as the states 1_c and 2_c (likewise 1_{nc} and 2_{nc}) fuse into a single state. In this limit, the effect of crosstalk is highly detrimental already at $\Lambda = 0.1$ used in this example: at low c, the promoter repeatedly cycles through erroneous initiation and the gene is highly expressed both at low c as well as at high c (where most of the expression is indeed due to correct initiation); as a result, the distributions $P(m|c)$ show substantial overlap in the two input conditions shown in Fig 5.1**C**. In contrast, for a non-trivial choice of q ($k_-^c \ll 1/q \simeq k_-^{nc}$), the model can exhibit proofreading. Even at low cognate concentration c, the slow irreversible transition ensures that noncognate TFs unbind from the promoter and that erroneous initiation is consequently rare, which is manifested as a sharp peak of $P(m|c_{low})$ at small m in Fig 5.1**D**. The proofreading architecture generates a larger output dynamic range and consequently makes the responses distinguishable.

What are the costs to the cell of the proposed proofreading mechanism? First, the mechanism requires an energy source, e.g., ATP, to break detailed balance, but this metabolic burden seems negligible compared to the processive cost of transcription and translation. Second, however, is an indirect cost in terms of gene expression noise. While proofreading decreases erroneous induction, it takes longer to traverse the state transition diagram from empty state 0 to expressing state 2, and since the promoter can perform aborted erroneous initiation cycles, the fluctuations in the time-to-induction will also increase [Bel1 *et al.*, 2010]. This will result in additional variance in the mRNA copy number at steady state compared to the two-state ($qd \to 0$) scheme. While the speed/specificity trade-off in protein synthesis has been examined before using deterministic chemical kinetics [Savir and Tlusty, 2013], this stochastic formulation of proofreading has, to our knowledge, remained unexplored. Proofreading in gene

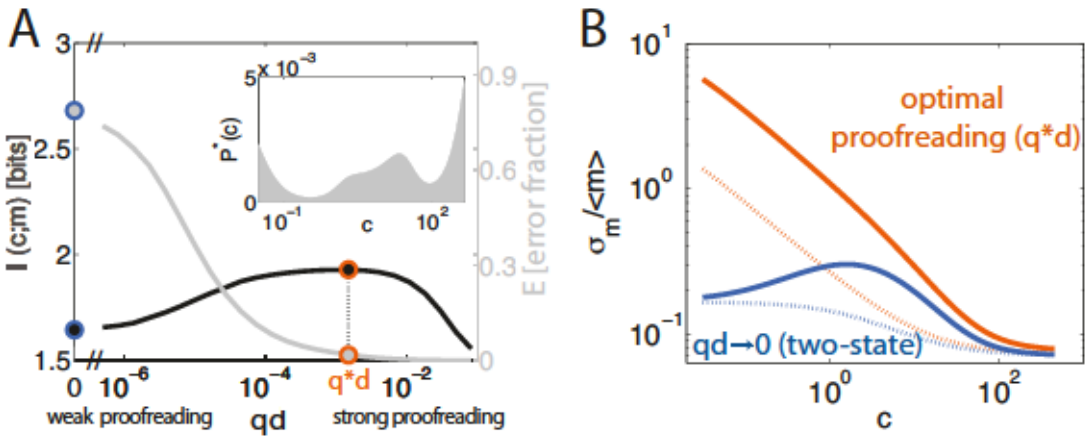

Figure 5.2: **A)** Maximal information transmission (left axis, black) and the error fraction (right axis, gray) as a function of the inverse irreversible reaction rate, qd. Increasing qd suppresses the error fraction, but only at the cost of increasing the gene expression noise, leading to a trade-off and an information-maximizing value of q^*d (orange). This maximum is reached robustly with input distributions that are close to optimal. **B)** Noise in gene expression, $\sigma_m/\langle m \rangle$, computed from the moments of $P(m|c)$, as a function of the dimensionless input concentration c, for the optimal proofreading (orange) and the two-state (blue) architectures. Dotted lines show the Poisson limit, $\sigma_m^2 = \langle m \rangle$, for comparison. In both cases, the average number of mRNA expressed if fixed to $\bar{m} = 100$.

regulation is thus expected to increase the output dynamic range, which is favorable for signaling, but also to increase the noise, which is detrimental.

How can we formalize the trade-off between noise and dynamic range for gene regulatory schemes and find when proofreading is beneficial? In existing analyses of proofreading the erroneous incorporation of the substrate leads to an error product that is *different* from the correct one [Hopfield, 1974; Savir and Tlusty, 2013]; in contrast, here the gene always expresses the *same* mRNA. What is important for signal transduction, however, is how well this expression correlates with the input signal, c. To quantify the regulatory power of the proofreading architecture, we computed the mutual information, $I(c;m)$ [Shannon and Weaver, 1949], between the signal c and the mRNA expression level m, following previous applications of information theory to gene regulation [Tkačik and Walczak, 2011; Rieckh and Tkačik, 2014]. The information depends not only on $P(m|c)$, which we compute from the Master equation, but also on the *a priori* unknown distribution of input concentrations, $P(c)$; we therefore determined the input distribution $P^*(c)$ that maximizes information transmission, subject to a constraint on the average number of expressed mRNA, $\bar{m} = \int dc P(c) \sum_m m P(m|c)$. This constraint on average number of mRNA was imposed to compare different regulatory architectures; otherwise, higher average expression could yield higher information transmission for trivial reasons. Such constrained information (capacity) maximization is a well-known problem in information theory that can be solved using the Blahut-Arimoto algorithm [Blahut, 1972].

Fig 5.2**A** shows how the information transmission $I(m;c)$ through the promoter depends on the (inverse) reaction rate qd. We start by looking at the classic measure of proofreading performance, the "error fraction," i.e., the ratio of the mRNA expressed from state 2_{nc} due to noncognate TFs, vs mRNA expressed from state 2_{c} due to cognate TFs. As qd is increased, the error fraction drops, with no clear optimum. In contrast, there exists an optimal q^*d at which the information is maximized—this is the point where proofreading is most effective, optimally trading off erroneous induction (here, suppressed by a factor of ~ 30 relative to no proofreading), noise in gene expression, and dynamic range at the output. In Fig 5.2**B** we plot the noise in gene expression, as a function of the input concentration c for the optimal proofreading architecture and the non-proofreading limit. In both cases the noise has super-Poisson components due

to the switching between promoter states, but this excess is substantially higher in the proofreading architecture, as expected.

While attractive, these results still depend on the particular rates chosen for the model in Fig 5.1**B**. Surprisingly, if we choose to compare the *optimal* proofreading scenario with the *optimal* non-proofreading one, the problem simplifies further. Given that the input TF concentration c varies over some limited dynamic range, $c \in [0, C_{\max} = k_+ C/d]$, there should exist also an optimal setting for k_-^c: set too high, the cognate TFs will be extremely unlikely to occupy the promoter for any significant fraction of the time and induce the gene; set too low, the switching contribution to noise in gene expression will blow up. With k_-^c and q in the "correct initiation" pathway of Fig 5.1**B** set by optimization, the remaining rates in the "erroneous initiation" pathway are fixed by the choice of crosstalk severity Λ. The remaining parameters regulating mRNA expression—the average mRNA count $\bar{m}$ and the rate r—do not change the results qualitatively. The mRNA expression rate r simply sets the maximal number of mRNA molecules at full expression in steady state (r/d); this influences the Poisson noise at the output, but does so equally for any regulatory architecture, proofreading or not. As long as r is large enough so that the average mRNA constraint $\bar{m}$ is achievable, the precise choice of these values is not crucial (we use $r/d = 200$, $\bar{m} = 100$, plausible for eukaryotic expression). In sum, we can compare how well the optimal proofreading architecture does compared to optimal non-proofreading architecture in terms of information transmission, as a function of two key parameters: the crosstalk severity, Λ, and the input dynamic range, $C_{\max}$.

Fig 5.3**A**shows the advantage, in bits, of the optimal proofreading architecture relative to the optimal non-proofreading one. This "information plane," $I_{q*}(m;c) - I_{q=0}(m;c)$, is plotted as a function of Λ and $C_{\max}$. In the limit $\Lambda \to 0$, the difference in performance goes to zero: there, optimization drives $q^* k_-^{nc,*} \gg 1$, but proofreading offers vanishing advantage over the optimal two-state promoter architecture when noncognate binding is negligible. As Λ increases, proofreading becomes beneficial over the two-state architecture, and more so for higher values of $C_{\max}$. Higher input concentrations $c \in [0, C_{\max}]$ permit faster on-rates, resulting in faster optimal off rates $k_-^{c,*}$ and faster optimal $1/q^*$. Generally, faster switching of promoter states in Fig 5.1**B** means that promoter switching noise will be lower and thus information higher (at fixed mean mRNA expression $\bar{m}$);

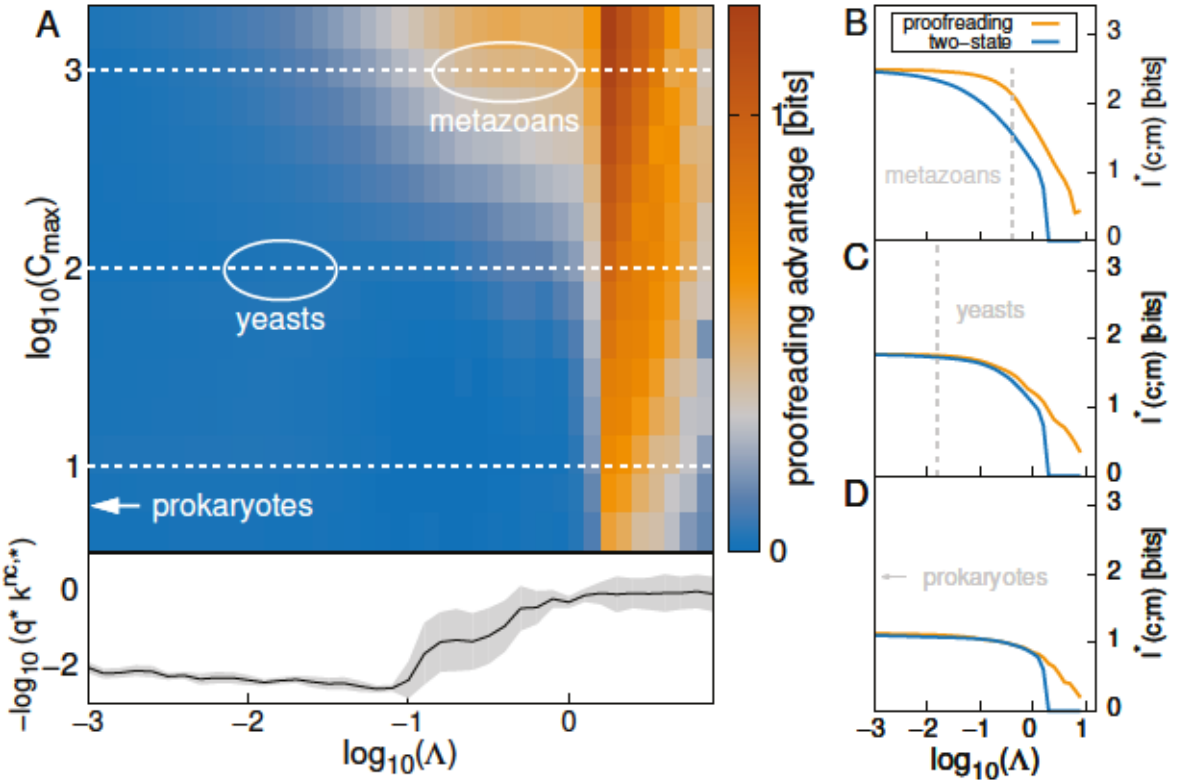

Figure 5.3: **A)** Information advantage (in bits, color scale) of optimal proofreading over optimal two-state architectures, as a function of crosstalk severity Λ and dynamic range of input TF concentration, $C_{\max}$. Typical values for prokaryotes, yeast, and metazoans are marked in white. Lower inset: optimal rates, $q^* k_-^{\mathrm{nc},*}$ (black line = average over $C_{\max}$, gray shade = std), indicate a switch to the proofreading strategy. **B, C, D)** Cuts through the information plane in (A) along white dashed lines showing the collapse of two-state performance as $\log_{10}(\Lambda) \to 0$ and a clear proofreading advantage for metazoan regulation.

in particular, optimization tends to minimize promoter switching noise by selecting the fastest $1/q$ that still admits error rejection, i.e., $q^* k_-^{\mathrm{nc},*} \sim 1$. At $\Lambda = \nu/\sigma \simeq 1$, the signaling capacity of the non-proofreading architecture collapses completely, with $I_{q=0}(c;m) \approx 0$ [1]. At this point optimal proofreading architectures are affected, but still generally maintain at least half of the capacity seen at $\Lambda = 0$; proofreading extends the performance of the gene regulation well into the $\Lambda > 0$ region, before finally succumbing to crosstalk.

Where do different organisms lie in the information plane? Prokaryotes have on the order of $\nu \sim 100$ types of transcription factors, whose binding site motifs typically contain around 23 bits of sequence information [Wunderlich and Mirny, 2009], or $16\, k_B T$ binding energy difference of between cognate and noncognate sites [Gerland *et al.*, 2002], corresponding to $\sigma \sim 10^7$. The resulting crosstalk severity is low, $\Lambda \sim 10^{-5}$. For yeast, the typical sequence information is 14 bits ($10\, k_B T$) [Wunderlich and Mirny, 2009], which gives $\Lambda \sim 0.01$ (for $\nu \sim 200$ [Jothi *et al.*, 2009]). For multicellular eukaryotes, the typical sequence information is 12 bits ($8\, k_B T$), and the number of TF species varies between $\nu \approx 10^3$ (*C. elegans*) to $\nu \approx 2 \cdot 10^3$ (human) [Milo *et al.*, 2010], putting Λ between 0.1 and 1. We can also estimate the dimensionless parameter $C_{\max} = k_+ C/d$. Assuming diffusion-limited binding of TFs to their binding sites, $k_+ C/d \approx 3DaN/R^3 d$, where $D \sim 1\mu\mathrm{m}^2/\mathrm{s}$ is the typical TF diffusion constant [Milo *et al.*, 2010], $a \sim 3$nm is the binding site size, $R = 3\mu$m (1μm) is the radius of an eukaryotic nucleus (prokaryotic cell), and N is the typical copy number of TFs per nucleus ($N \sim 10$ for prokaryotes, 10^3 for yeast, $10^3 - 10^5$ for eukaryotes). Typical mRNA lifetimes are $5 - 10$ min in prokaryotes, $20 - 30$ min in yeast, and > 1 hour in metazoans. This yields $C_{\max}$ of order 10 for prokaryotes, 10^2 for yeast cells, and $> 10^3$ for multicellular eukaryote cells. While these are very rough estimates, different kinds of cells clearly differ substantially in their location on the information plane of Fig 5.3A.

Taken together, these values suggest that crosstalk is acute for metazoans and that proofreading in gene regulation could provide a vast improvement over regulation at equilibrium, as in Fig 5.3**B**. In contrast, our proposal offers no advantage for prokaryotes, and remains agnostic about yeast (Figs 5.3**C**, **D**). While much remains

[1] This is independent of whether one modulates Λ by changing ν, as for Fig 5.3A, or by changing σ; although the optimal rates may take on different values, the information plane is essentially unchanged irrespective of how Λ is modulated.

unknown about the molecular machinery of eukaryotic gene regulation, it has been experimentally shown that transcriptional initiation (not just elongation) involves a series of out-of-equilibrium steps. Amongst those, perhaps the most intriguing are the covalent modifications on the eukaryotic RNA polymerase II CTD tail [Egloff and Murphy, 2008]. The tail contains tandem repeats of short peptides (from 26 repeats in yeast to 52 in mammals), which need to get phosphorylated in order to initiate transcription and subsequently cleared after completed transcription in order to reuse the polymerase; genetic interference with this tail seems to be lethal. One can contemplate a scenario where a sequence of such phosphorylation steps corresponds to the out-of-equilibrium reaction q of our simple proofreading scheme, "ticking away" time until the polymerase commits to initiation, with every tick giving the machinery another opportunity to check if cognate TFs are still bound and, if not, abort transcription. The existence of any such (or similar) proofreading scheme would be interesting, but is currently purely hypothetical. An alternative proofreading mechanism would make use of histone modifications: a TF could interact with histones to mark the +1 nucleosome and facilitate promoter escape for the RNA polymerase. More complex schemes could also exist, and might benefit from multiple out-of-equilibrium steps both to boost specificity and reduce promoter switching noise [Rieckh and Tkačik, 2014], which is an interesting topic for future research. How could these proofreading ideas be tested? Indirect evidence for kinetic schemes in regulation exists. Crystal structure of RNA polymerase II during early promoter clearance indicates that abortive initiation is a side-product of "promoter proofreading" [Liu *et al.*, 2011]. Experimentally documented interactions between histone tail modifiers, chromatin remodelers, and TFs appear consistent with kinetic proofreading [Blossey and Schiessel, 2011]. Kinetic studies of gene activation by TF binding are inconsistent with equilibrium models [Chen *et al.*, 2014]. Direct evidence showing that TF specificity is boosted by proofreading to reduce erroneous gene regulation is, however, lacking. Tests following [Hopfield *et al.*, 1976] to measure ATP consumption per mRNA upon initiation due to cognate vs noncognate TFs appear possible *in vitro* for RNA polymerase II CTD modification mechanism, but difficult for histone-based mechanisms, which might be better tested indirectly using genetic perturbations.

While we cannot exclude the existence of a complex equilibrium scheme that re-

duces crosstalk in gene regulation sufficiently, this and our related work [Friedlander *et al.*, 2016] suggest that equilibrium solutions, if they exist, are not simple. Here we advanced an alternative hypothetical mechanism, proofreading-based transcriptional regulation, to mitigate the crosstalk problem. Unlike most biophysical problems where we clearly appreciate their out-of-equilibrium nature, transcriptional regulation has remained a textbook example of a non-trivial *equilibrium* molecular recognition process, likely due to the success of the equilibrium assumption in prokaryotes. Crosstalk considerations should motivate us to reexamine this assumption in eukaryotic regulation.

6 Conclusions and future directions

Increasing availability of single-cell time-resolved data should allow us to address open questions regarding the amount of information encoded about the external world that is available in the time-varying concentrations, activation or localization patterns, and modification state of various biochemical molecules. Do full response trajectories provide more information than single temporal snapshots, as early studies suggest? Is this information gain purely due to noise averaging enabled by observing multiple snapshots, or—more interestingly—due to the ability of these intrinsically high-dimensional signals to provide a richer representation of the cellular environment? Can we isolate biologically relevant features of the response trajectories, e.g., amplitude, frequency, pulse shape, relative phase or timing, without *a priori* assuming what these features are? How can cells read out the environmental state from these response trajectories and how close to the information-theoretical bounds is this readout process?

In chapters 3 and 4, we made steps towards answering these questions by focusing on the following questions: first, if we are given a full stochastic description of a biochemical reaction network, under what conditions can we theoretically compute information transmission through this network and various related bounds; Second, if we are given real data with no description of the network, what are tractable schemes to estimate the information transmission and how is this formalism useful to understand dynamical signaling networks?

We show a tractable Monte Carlo approximation scheme to estimate information for simple biological reactions, where the complete set of reactions are observed. Additionally, we introduce decoding-based model-free estimation methods and compare their performance to the commonly-used knn estimator. Our results show that

decoding-based estimators closely approach the optimal decoder performance and in many cases perform better than knn.

It is necessary to emphasize the flexibility of the decoding approach: decoding-based information estimation is based directly on the statistical problems of classification (for discrete input variable, U) or regression (for continuous input variable, U), so any classification / regression algorithm with good performance can provide the basis for information estimation. Statistical algorithms underlying decoding-based estimations have the extra advantage that: (i), we may be able to gain biological insight by inspecting which features of the response carry the relevant stimulus information (e.g., by looking at the linear kernels or features that neural networks extract in their various layers); (ii), pick a decoding algorithm based on features previously reported as relevant (e.g., The Gaussian decoder for second-order statistics as in Example 3), and (iii), estimate the information as a function of trajectory duration; and (iv), gain confidence in our estimates by testing their performance on withheld data.

By construction, decoding-based estimators only provide a lower bound to the true information. This, however, could turn out to be a smaller problem in practice than it appears in theory, especially for biochemical reaction networks. First, the extension to the Feder-Merhav bound [Hledík *et al.*, 2018] provides an estimate of how large the gap between the true information and the decoded estimation can be. The bound is not tight on our examples, and can only be applied when the optimal MAP decoder can be constructed [Tkačik *et al.*, 2015]. Second, and perhaps more importantly, information that can be decoded after single input presentations is the quantity that is likely more biologically relevant than the true channel capacity, if the organisms are under constraint to respond to the environmental changes quickly. Typically, organisms across the complexity scale operate under speed-accuracy tradeoffs [Heitz, 2014]: faster decisions based on noisy information lead to more errors and, conversely, with enough time to integrate sensory information errors can be reduced. When speed is at a premium or relevant inputs are sparse, decisions need to be taken after single input presentations. In this case, decoding-based estimation should not be viewed as an approximate but rather as the correct methodology for the biological problem at hand. Of course, there is still the question of whether the model-free decoders that we use on real data can achieve a performance that is close to the optimal MAP decoder that represents the

absolute performance limit. While there is no general way to answer this question, it appears that simple SVM decoding schemes work well when the response trajectories differ in their conditional mean, and neural networks as general approximators can be used to check for more complicated encoding features when data is plentiful. Unlike in neuroscience, there is much less clarity about what kind of read-out or decoding operations biochemical networks can mechanistically realize to mimic the functioning of our *in silico* decoders, and it may be challenging to biochemically implement even arbitrary linear classification of response trajectories. Until experimentally shown otherwise, it thus appears reasonable to proceed with the assumption that environmental signals can be read out from the time-dependent internal chemical state with a simple repertoire of computations.

We emphasize a simple yet important point. The decoding-based approach that we introduced here should also motivate us to look beyond methodological problems of significance and estimation, to truly biological problems of cellular decision making. Currently, data on biological regulatory processes is often analyzed by looking for "statistically significant differences" in the network response for, say, two possible network inputs. For example, one may report that the steady-state mean expression level of a certain gene is significantly larger in the stimulated vs unstimulated condition, with the statistical significance of the mean difference established through an appropriate statistical test that takes into account the number of collected population samples. While statistical significance is a necessary condition to validly report *any* difference in the response, it is very different from the question of whether *a single cell* could discriminate the two conditions given access only to its own expression levels. In theory, population-level statistics tell us with what confidence we, as scientists having access to N samples, can discriminate between conditions given some biological readout; decoding based information estimates, on the other hand, are relevant to the $N = 1$ case of individual cells. We hope that further work along the latter path can clarify and quantify better the difficult constraints and conditions under which real cells need to act based on individual noisy readouts of their stochastic biochemistry.

In chapter 5, our study of information transmission in single cells takes a different perspective and focuses on transcriptional regulation. The question of how accurately signals are transmitted when the source is noisy–in the sense introduced by Weaver

as the “semantic noise” [Shannon and Weaver, 1949], namely crosstalk, where non-cognate transcription factors can initiate transcription with potentially fatal effects for the cell, leads us to examine a non-equilibrium scheme that implements proofreading. This extension would mitigate the problem if it wouldn’t increase noise levels in gene expression. That trade-off helps us to determine a regime where proofreading is advantageous compared to the equilibrium scheme. In the case of metazoans, our results suggest that proofreading in transcriptional regulation could significantly improve the accuracy of information transmission. However, the study of this hypothetical mechanism in the presence of crosstalk suggests to reexamine the assumption that transcriptional initiation is an equilibrium process.

In conclusion, the results provided in this work - even though in their infancy - represent a good foundation for further studies investigating the dynamic intracellular representation of cell’s environment and out of equilibrium transcriptional regulation.

Bibliography

[Abadi *et al.*, 2015] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," 2015, Software available from tensorflow.org.

[Albeck *et al.*, 2013] John G Albeck, Gordon B Mills, and Joan S Brugge, "Frequency-modulated pulses of ERK activity transmit quantitative proliferation signals.," *Molecular cell*, 49(2):249–61, 2013.

[Alberts *et al.*, 2015] Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Roberts, and Peter Walter, *Molecular Biology of the Cell*, Garland Science, 6th edition, 2015.

[Alon, 2006] Uri Alon, *An Introduction to Systems Biology: Design Principles of Biological Circuits*, Chapman & Hall/CRC Mathematical and Computational Biology, 2006.

[Anderson *et al.*, 1958] Theodore Wilbur Anderson, Theodore Wilbur Anderson, Theodore Wilbur Anderson, Theodore Wilbur Anderson, and Etats-Unis Mathématicien, *An introduction to multivariate statistical analysis*, volume 2, Wiley New York, 1958.

[Barato *et al.*, 2013] Andre C. Barato, David Hartich, and Udo Seifert, "Rate of Mutual Information Between Coarse-Grained Non-Markovian Variables," *Journal of Statistical Physics*, 153:460–478, 2013.

[Barato *et al.*, 2014] Andre C Barato, David Hartich, and Udo Seifert, "Efficiency of cellular information processing," *New Journal of Physics*, 16:103024, 2014.

[Barlow, 1959] HB Barlow, "Sensory mechanisms, the reduction of redundancy, and intelligence," *NPL Symposium on the Mechanization of Thought Process*, (10):535–539, 1959.

[Batchelor *et al.*, 2008] Eric Batchelor, Caroline S. Mock, Irun Bhan, Alexander Loewer, and Galit Lahav, "Recurrent Initiation: A Mechanism for Triggering p53 Pulses in Response to DNA Damage," *Molecular Cell*, 30(3):277–289, 2008.

[Becker *et al.*, 2015] Nils B Becker, Andrew Mugler, and Pieter Rein Ten Wolde, "Optimal Prediction by Cellular Signaling Networks," *Physical Review Letters*, 115(25):258103, 2015.

[Behar and Hoffmann, 2010] Marcelo Behar and Alexander Hoffmann, "Understanding the temporal codes of intra-cellular signals," *Current Opinion in Genetics & Development*, 20(6):684–693, 2010.

[Bel1 *et al.*, 2010] Golan Bel1, Brian Munsky, and Ilya Nemenman, "The simplicity of completion time distributions for common complex biochemical processes," *Physical biology*, 7(1), 2010.

[Benabdeslem, Khalid and Bennani, 2006] Younès Benabdeslem, Khalid and Bennani, "Dendogram-based SVM for Multi-Class Classification," *Journal of Computing and Information Technology*, 14(4):283–289, 2006.

[Berg and Purcell, 1977] H C Berg and E M Purcell, "Physics of chemoreception.," *Biophysical journal*, 20:193–219, 1977.

[Bialek *et al.*, 1991] W Bialek, F Rieke, R R de Ruyter van Steveninck, and D Warland, "Reading a neural code.," *Science (New York, N.Y.)*, 252(5014):1854–7, 1991.

[Bialek, 2012] William Bialek, *Biophysics: Searching for Principles*, Princeton University Press, 2012.

[Bialek, 2018] William Bialek, "Perspectives on theory at the interface of physics and biology," *Reports on Progress in Physics*, 81(1):012601, 2018.

[Bialek and Setayeshgar, 2005] William Bialek and Sima Setayeshgar, "Physical limits to biochemical signaling," *Proceedings of the National Academy of Sciences*, 102(29):10040–10045, 2005.

[Bintu *et al.*, 2005] Lacramioara Bintu, Nicolas E Buchler, Hernan G Garcia, Ulrich Gerland, Terence Hwa, Jan Kondev, and Rob Phillips, "Transcriptional regulation by the numbers: models," *Current Opinion in Genetics & Development*, 15(2):116 – 124, 2005.

[Blahut, 1972] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, 18(4):460–473, 1972.

[Blake *et al.*, 2003] William J Blake, Mads KAErn, Charles R Cantor, and J J Collins, "Noise in eukaryotic gene expression.," *Nature*, 422(6932):633–7, 2003.

[Blakemore, 1975] R. P. Blakemore, "Magnetotactic bacteria," *Science*, 190:377–379, 1975.

[Blossey and Schiessel, 2011] Ralf Blossey and Helmut Schiessel, "Kinetic Proofreading in Chromatin Remodeling: The Case of ISWI/ACF," *Biophysical Journal*, 101(4):L30–L32, 2011.

[Borst and Theunissen, 1999] A Borst and F E Theunissen, "Information theory and neural coding," *Nature neuroscience*, 2:947–957, 1999.

[Bowsher and Swain, 2012] Clive G Bowsher and Peter S Swain, "Identifying sources of variation and the flow of information in biochemical networks.," *Proceedings of the National Academy of Sciences of the United States of America*, 109(20):E1320–8, 2012.

[Bowsher and Swain, 2014] Clive G Bowsher and Peter S Swain, "Environmental sensing, information transfer, and cellular decision-making," *Current Opinion in Biotechnology*, 28:149 – 155, 2014.

[Brennan *et al.*, 2012] Matthew D. Brennan, Raymond Cheong, and Andre Levchenko, "How Information Theory Handles Cell Signaling and Uncertainty," *Science*, 338(6105):334–335, 2012.

[Brunel and Nadal, 1998] Nicolas Brunel and J. P. Nadal, "Mutual information, Fisher information, and population coding.," *Neural computation*, 10(7), 1998.

[Cai *et al.*, 2008] Long Cai, Chiraj K Dalal, and Michael B Elowitz, "Frequency-modulated nuclear localization bursts coordinate gene regulation," *Nature*, 455:485–490, 2008.

[Carballo-Pacheco *et al.*, 2018] Martin Carballo-Pacheco, Jonathan Desponds, Tatyana Gravilchenko, Andreas Mayer, Roshan Prizak, Gautam Reddy, Ilya Nemenman, and Thierry Mora, "Receptor crosstalk improves concentration sensing of multiple ligands," *bioRxiv*, page 448118, 2018.

[Cepeda-Humerez *et al.*, 2015] Sarah A. Cepeda-Humerez, Georg Rieckh, and Gašper Tkačik, "Stochastic Proofreading Mechanism Alleviates Crosstalk in Transcriptional Regulation," *Physical Review Letters*, 115(24):248101, 2015.

[Cepeda-Humerez *et al.*, 2019] Sarah A Cepeda-Humerez, Jakob Ruess, and Gašper Tkačik, "Estimating information in time-varying signals," 2019.

[Chen *et al.*, 2014] Jiji Chen, Zhengjian Zhang, Li Li, Bi-Chang Chen, Andrey Revyakin, Bassam Hajj, Wesley Legant, Maxime Dahan, Timothée Lionnet, Eric Betzig, Robert Tjian, and Zhe Liu, "Single-Molecule Dynamics of Enhanceosome Assembly in Embryonic Stem Cells," *Cell*, 156(6):1274 – 1285, 2014.

[Cheong *et al.*, 2011] Raymond Cheong, Alex Rhee, Chiaochun Joanne Wang, Ilya Nemenman, and Andre Levchenko, "Information Transduction Capacity of Noisy Biochemical Signaling Networks," *Science*, 334(6054):354–358, 2011.

[Chevalier *et al.*, 2015] Michael Chevalier, Ophelia Venturelli, and Hana El-Samad, "The Impact of Different Sources of Fluctuations on Mutual Information in Biochemical Networks," *PLOS Computational Biology*, 11(10):e1004462, 2015.

[Cover and Thomas, 2005] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, John Wiley and Sons, 2005.

[Crane *et al.*, 2014] Matthew M. Crane, Ivan B. N. Clark, Elco Bakker, Stewart Smith, and Peter S. Swain, "A Microfluidic System for Studying Ageing and Dynamic Single-Cell Responses in Budding Yeast," *PLoS ONE*, 9(6):e100042, 2014.

[Crick, 1970] Francis Crick, "Central dogma of molecular biology," *Nature*, 227(5258):561–563, 1970.

[Crisanti *et al.*, 2018] Andrea Crisanti, Andrea De Martino, and Jonathan Fiorentino, "Statistics of optimal information flow in ensembles of regulatory motifs," *Physical Review E*, 97(2):022407, 2018.

[Cuadrado *et al.*, 2006] Myriam Cuadrado, Barbara Martinez-Pastor, Matilde Murga, Luis I Toledo, Paula Gutierrez-Martinez, Eva Lopez, and Oscar Fernandez-Capetillo, "ATM regulates ATR chromatin loading in response to DNA double-strand breaks," *The Journal of Experimental Medicine*, 203(2):297–303, 2006.

[Cyert, 2001] Martha S Cyert, "Regulation of Nuclear Localization during Signaling," 2001.

[Dalal *et al.*, 2014] Chiraj K Dalal, Long Cai, Yihan Lin, Kasra Rahbar, and Michael B Elowitz, "Pulsatile Dynamics in the Yeast Proteome," *Current Biology*, 24(18):2189–2194, 2014.

[de Ronde *et al.*, 2011] Wiet de Ronde, Filipe Tostevin, and Pieter Rein ten Wolde, "Multiplexing Biochemical Signals," *Phys. Rev. Lett.*, 107:048101, 2011.

[Dobrushin, 1958] R. L. Dobrushin, "A simplified method of experimental estimation of the entropy of a stationary distribution," *Tear. Veroyatnost. i Primenen; English transl. Theory Probab. Appl.*, 3:462464, 1958.

[Dubuis *et al.*, 2013a] Julien O. Dubuis, Reba Samanta, and Thomas Gregor, "Accurate measurements of dynamics and reproducibility in small genetic networks," *Molecular Systems Biology*, 9(1):1–22, 2013.

[Dubuis *et al.*, 2013b] Julien O. Dubuis, Gašper Tkačik, Eric F. Wieschaus, Thomas Gregor, and William Bialek, "Positional information, in bits," *Proceedings of the National Academy of Sciences*, 110(41):16301–16308, 2013.

[Egloff and Murphy, 2008] Sylvain Egloff and Shona Murphy, "Cracking the RNA polymerase II CTD code," *Trends in Genetics*, 24(6):280 – 288, 2008.

[Eldar and Elowitz, 2010] Avigdor Eldar and Michael B Elowitz, "Functional roles for noise in genetic circuits.," *Nature*, 467(7312):167–173, 2010.

[Elowitz *et al.*, 2002] Michael B. Elowitz, Arnold J. Levine, Eric D. Siggia, and Peter S. Swain, "Stochastic Gene Expression in a Single Cell," *Science*, 297(5584):1183–1186, 2002.

[Fakhouri *et al.*, 2010] Walid D Fakhouri, Ahmet Ay, Rupinder Sayal, Jacqueline Dresch, Evan Dayringer, and David N Arnosti, "Deciphering a transcriptional regulatory code: modeling short-range repression in the Drosophila embryo," *Molecular Systems Biology*, 6, 2010.

[Feder and Merhav, 1994] M. Feder and N. Merhav, "Relations between entropy and error probability," *IEEE Transactions on Information Theory*, 40(1):259–266, 1994.

[Feinerman *et al.*, 2008] Ofer Feinerman, Joël Veiga, Jeffrey R Dorfman, Ronald N Germain, and Grégoire Altan-Bonnet, "Variability and robustness in T cell activation from regulated heterogeneity in protein levels.," *Science (New York, N.Y.)*, 321(5892):1081–4, 2008.

[Friedlander *et al.*, 2016] Tamar Friedlander, Roshan Prizak, Călin C. Guet, Nicholas H. Barton, and Gašper Tkačik, "Intrinsic limits to gene regulation by global crosstalk," *Nature Communications*, 7, 2016.

[Garner *et al.*, 2016] Kathryn L Garner, Rebecca M Perrett, Margaritis Voliotis, Clive Bowsher, George R Pope, Thanh Pham, Christopher J Caunt, Krasimira Tsaneva-Atanasova, and Craig A McArdle, "Information transfer in gonadotropin-releasing hormone (GnRH) signaling: Extracellular signal-regulated kinase (ERK)-mediated feedback loops control hormone sensing," *Journal of Biological Chemistry*, 291(5):2246–2259, 2016.

[Garner *et al.*, 2017] Kathryn L. Garner, Margaritis Voliotis, Hussah Alobaid, Rebecca M. Perrett, Thanh Pham, Krasimira Tsaneva-Atanasova, and Craig A. McArdle, "Information Transfer via Gonadotropin-Releasing Hormone Receptors to ERK

and NFAT: Sensing GnRH and Sensing Dynamics," *Journal of the Endocrine Society*, 1(4):260–277, 2017.

[Gerland *et al.*, 2002] Ulrich Gerland, J. David Moroz, and Terence Hwa, "Physical constraints and functional characteristics of transcription factor–DNA interaction," *Proceedings of the National Academy of Sciences*, 99(19):12015–12020, 2002.

[Géron, 2017] Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, O'Reilly, 2017.

[Gillespie, 1977] Daniel T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[Gillespie, 1992] Daniel T. Gillespie, "A rigorous derivation of the chemical master equation," *Physica A: Statistical Mechanics and its Applications*, 188(1-3):404–425, 1992.

[Giorgetti *et al.*, 2010] Luca Giorgetti, Trevor Siggers, Guido Tiana, Greta Caprara, Samuele Notarbartolo, Teresa Corona, Manolis Pasparakis, Paolo Milani, Martha L. Bulyk, and Gioacchino Natoli, "Noncooperative Interactions between Transcription Factors and Clustered DNA Binding Sites Enable Graded Transcriptional Responses to Environmental Inputs," *Molecular Cell*, 37(3):418 – 428, 2010.

[Gotoh *et al.*, 1990] Yukiko Gotoh, Eisuke Nishida, Takashi Yamashita, Minako Hoshi, Minoru Kawakami, and Hikoichi Sakai, "Microtubule-associated-protein (MAP) kinase activated by nerve growth factor and epidermal growth factor in PC12 cells. Identity with the mitogen-activated MAP kinase of fibroblastic cells," *European Journal of Biochemistry*, 193(3):661–669, 1990.

[Granados *et al.*, 2018] Alejandro A Granados, Julian M J Pietsch, Sarah A Cepeda-Humerez, Iseabail L Farquhar, Gašper Tkačik, and Peter S Swain, "Distributed and dynamic intracellular organization of extracellular information.," *Proceedings of the National Academy of Sciences of the United States of America*, 115(23):6088–6093, 2018.

[Gregor *et al.*, 2007a] Thomas Gregor, David W. Tank, Eric F. Wieschaus, and William Bialek, "Probing the Limits to Positional Information," *Cell*, 130(1):153 – 164, 2007.

[Gregor *et al.*, 2007b] Thomas Gregor, Eric F Wieschaus, Alistair P. McGregor, William Bialek, and David W Tank, "Stability and Nuclear Dynamics of the Bicoid Morphogen Gradient," *Cell*, 130(1):141–152, 2007.

[Hafner *et al.*, 2017] Antonina Hafner, Jacob Stewart-Ornstein, Jeremy E Purvis, William C Forrester, Martha L Bulyk, and Galit Lahav, "p53 pulses lead to distinct patterns of gene expression albeit similar DNA-binding dynamics," *Nature Structural & Molecular Biology*, 24(10):840–847, 2017.

[Hansen and O'Shea, 2015] Anders S Hansen and Erin K O'Shea, "Limits on information transduction through amplitude and frequency regulation of transcription factor activity," *eLife*, 4:e06559, 2015.

[Hao and Shea, 2012] Nan Hao and Erin K O Shea, "Signal-dependent dynamics of transcription factor translocation controls gene expression," *Nature Publishing Group*, 19(1):31–39, 2012.

[Hasegawa, 2018] Yoshihiko Hasegawa, "Multidimensional biochemical information processing of dynamical patterns," *Physical Review E*, 97(2):22401, 2018.

[He *et al.*, 2010] Xin He, Md. Abul Hassan Samee, Charles Blatti, and Saurabh Sinha, "Thermodynamics-Based Models of Transcriptional Regulation by Enhancers: The Roles of Synergistic Activation, Cooperative Binding and Short-Range Repression," *PLOS Computational Biology*, 6:1–15, 09 2010.

[Heitz, 2014] Richard P Heitz, "The speed-accuracy tradeoff: history, physiology, methodology, and behavior," *Frontiers in neuroscience*, 8:150, 2014.

[Hledík *et al.*, 2018] Michal Hledík, Thomas Sokolowski, and Gašper Tkačik, "An Upper Bound on Mutual Information," 2018.

[Hopfield *et al.*, 1976] J J Hopfield, T Yamane, V Yue, and S M Coutts, "Direct experimental evidence for kinetic proofreading in amino acylation of tRNAIle," *Proceedings of the National Academy of Sciences*, 73(4):1164–1168, 1976.

[Hopfield, 1974] JJ Hopfield, "Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity.," *Proceedings of the National Academy of Sciences of the United States of America*, 71:4135–4139, 1974.

[Jacob and Monod, 1961] François Jacob and Jacques Monod, "Genetic regulatory mechanisms in the synthesis of proteins," *Journal of Molecular Biology*, 3(3):318–356, 1961.

[Janssens *et al.*, 2006] Hilde Janssens, Shuling Hou, Johannes Jaeger, Ah-Ram Kim, Ekaterina Myasnikova, David Sharp, and John Reinitz, "Quantitative and predictive model of transcriptional control of the Drosophila melanogaster even skipped gene," *Nature Genetics*, 38, 2006.

[Jothi *et al.*, 2009] Raja Jothi, S Balaji, Arthur Wuster, Joshua A Grochow, Jörg Gsponer, Teresa M Przytycka, L Aravind, and M Madan Babu, "Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture," *Molecular Systems Biology*, 5(1), 2009.

[Kaiser and Schreiber, 2002] A. Kaiser and T. Schreiber, "Information transfer in continuous processes," *Physica D: Nonlinear Phenomena*, 166(1-2):43–62, 2002.

[Kaizu *et al.*, 2014] Kazunari Kaizu, Wiet De Ronde, Joris Paijmans, Koichi Takahashi, Filipe Tostevin, and Pieter Rein Ten Wolde, "The berg-purcell limit revisited," *Biophysical Journal*, 106:976–985, 2014.

[Khan *et al.*, 2007] Shiraj Khan, Sharba Bandyopadhyay, Auroop R. Ganguly, Sunil Saigal, David J. Erickson, Vladimir Protopopescu, and George Ostrouchov, "Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data," *Physical Review E*, 76(2):026209, 2007.

[Kinney *et al.*, 2010] Justin B. Kinney, Anand Murugan, Curtis G. Callan, and Edward C. Cox, "Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence," *Proceedings of the National Academy of Sciences*, 107(20):9158–9163, 2010.

[Kraskov *et al.*, 2004] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger, "Estimating mutual information," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 69, 2004.

[Kuge *et al.*, 1997] Shusuke Kuge, Nic Jones, and Akio Nomoto, "Regulation of yAP-1 nuclear localization in response to oxidative stress," *EMBO Journal*, 16(7):1710–1720, 1997.

[Kuhlman *et al.*, 2007] Thomas Kuhlman, Zhongge Zhang, Milton H. Saier, and Terence Hwa, "Combinatorial transcriptional control of the lactose operon of Escherichia coli," *Proceedings of the National Academy of Sciences*, 104(14):6043–6048, 2007.

[Lahav *et al.*, 2004] Galit Lahav, Nitzan Rosenfeld, Alex Sigal, Naama Geva-Zatorsky, Arnold J Levine, Michael B Elowitz, and Uri Alon, "Dynamics of the p53-Mdm2 feedback loop in individual cells," *NATURE GENETICS*, 36(2), 2004.

[Lajnef *et al.*, 2015] Tarek Lajnef, Sahbi Chaibi, Perrine Ruby, Pierre Emmanuel Aguera, Jean Baptiste Eichenlaub, Mounir Samet, Abdennaceur Kachouri, and Karim Jerbi, "Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines," *Journal of Neuroscience Methods*, 250:94–105, 2015.

[Lang *et al.*, 2014] Alex H. Lang, Charles K. Fisher, Thierry Mora, and Pankaj Mehta, "Thermodynamics of Statistical Inference by Cells," *Physical Review Letters*, 113:148103, 2014.

[Levine *et al.*, 2013] Joe H. Levine, Yihan Lin, and Michael B Elowitz, "Functional Roles of Pulsing in Genetic Circuits," *Science (New York, N.Y.)*, 342(December):1193–1200, 2013.

[Li *et al.*, 2014] Gene-Wei Li, David Burkhardt, Carol Gross, and Jonathan S Weissman, "Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources.," *Cell*, 157(3):624–35, 2014.

[Li *et al.*, 2008] Xiao-yong Li, Stewart MacArthur, Richard Bourgon, David Nix, Daniel A Pollard, Venky N Iyer, Aaron Hechmer, Lisa Simirenko, Mark Stapleton,

Cris L. Luengo Hendriks, Hou Cheng Chu, Nobuo Ogawa, William Inwood, Victor Sementchenko, Amy Beaton, Richard Weiszmann, Susan E Celniker, David W Knowles, Tom Gingeras, Terence P Speed, Michael B Eisen, and Mark D Biggin, "Transcription Factors Bind Thousands of Active and Inactive Regions in the Drosophila Blastoderm," *PLOS Biology*, 6(2):1–24, 02 2008.

[Little *et al.*, 2013] Shawn C. Little, Mikhail Tikhonov, and Thomas Gregor, "Precise Developmental Gene Expression Arises from Globally Stochastic Transcriptional Activity," *Cell*, 154(4):789 – 800, 2013.

[Liu *et al.*, 2011] Chenli Liu, Xiongfei Fu, Lizhong Liu, Xiaojing Ren, Carlos K.L. Chau, Sihong Li, Lu Xiang, Hualing Zeng, Guanhua Chen, Lei-Han Tang, Peter Lenz, Xiaodong Cui, Wei Huang, Terence Hwa, and Jian-Dong Huang, "Sequential Establishment of Stripe Patterns in an Expanding Cell Population," *Science*, 334(6053):238–241, 2011.

[MacKay and McCulloch, 1952] Donald M. MacKay and Warren S. McCulloch, "The limiting information capacity of a neuronal link," *The Bulletin of Mathematical Biophysics*, 14(2):127–135, 1952.

[Maerkl and Quake, 2007] Sebastian J. Maerkl and Stephen R. Quake, "A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors," *Science*, 315(5809):233–237, 2007.

[Mancini *et al.*, 2013] F Mancini, C H Wiggins, M Marsili, and A M Walczak, "Time-dependent information transmission in a model regulatory circuit," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 88(2):22708, 2013.

[Marre *et al.*, 2015] Olivier Marre, Vicente Botella-Soler, Kristina D Simmons, Thierry Mora, Gašper Tkačik, and Michael J Berry II, "High accuracy decoding of dynamical motion from a large retinal population," *PLoS computational biology*, 11(7):e1004304, 2015.

[Maturana R and J. Varela, 1973] Humberto Maturana R and Francisco J. Varela, *De máquinas y seres vivos : autopoiesis : la organización de lo vivo*, Editorial Universitaria, 1973.

[Mc Mahon *et al.*, 2015] Siobhan S Mc Mahon, Oleg Lenive, Sarah Filippi, and Michael P H Stumpf, "Information processing by simple molecular motifs and susceptibility to noise.," *Journal of the Royal Society, Interface*, 12(110):0597, 2015.

[McCulloch and Pitts, 1943] Warren S. McCulloch and Walter Pitts, "A logical calculus of the ideas immanent in nervous activity," *The Bulletin of Mathematical Biophysics*, 5(4):115–133, 1943.

[McGinnis and Krumlauf, 1992] William McGinnis and Robb Krumlauf, "Homeobox genes and axial patterning," *Cell*, 68(2):283 – 302, 1992.

[Milo *et al.*, 2010] Ron Milo, Paul Jorgensen, Uri Moran, Griffin Weber, and Michael Springer, "BioNumbers–the database of key numbers in molecular and cell biology," *Nucleic Acids Research*, 38:D750–D753, 2010.

[Milo and Philips, 2016] Ron Milo and Rob Philips, *Cell Biology by the Numbers*, Garland Science, 2016.

[Mirny, 2010] Leonid A. Mirny, "Nucleosome-mediated cooperativity between transcription factors," *Proceedings of the National Academy of Sciences*, 107(52):22534–22539, 2010.

[Mitchell *et al.*, 2009] Amir Mitchell, Gal H. Romano, Bella Groisman, Avihu Yona, Erez Dekel, Martin Kupiec, Orna Dahan, and Yitzhak Pilpel, "Adaptive prediction of environmental changes by microorganisms," *Nature*, 460(7252):220–224, 2009.

[Murphy, 2012] Kevin P. Murphy, *Machine Learning: A Probabilistic Perspective*, The MIT Press, 2012.

[Nandagopal *et al.*, 2018] Nagarajan Nandagopal, Leah A. Santat, Lauren LeBon, David Sprinzak, Marianne E. Bronner, and Michael B. Elowitz, "Dynamic Ligand Discrimination in the Notch Signaling Pathway," *Cell*, 172(4):869–880.e19, 2018.

[Nguyens *et al.*, 1993] Tien T Nguyens, Jean-Claude Scimeca, Chantal Filloux, Pascal Peraldi, Jean-Louis Carpentier, and Emmanuel Van Obberghen, "Co-regulation of the Mitogen-activated Protein Kinase, Extracellular Signal-regulated Kinase 1, and the 90-kDa Ribosomal S6 Kinase in PC12 Cells," *THE JOURNAL OF BIOLOGICAL CHEMISTRY*, 268(13):9803–9810, 1993.

[Ninio, 1975] J Ninio, "Kinetic amplification of enzyme discrimination.," *Biochimie*, 57:587–595, 1975.

[Olshausen and Field, 2004] Bruno A Olshausen and David J Field, "Sparse coding of sensory inputs," *Current Opinion in Neurobiology*, 14(4):481–487, 2004.

[Ouldridge *et al.*, 2017] Thomas E. Ouldridge, Christopher C. Govern, and Pieter Rein ten Wolde, "Thermodynamics of computational copying in biochemical systems," *Physical Review X*, 7(2):021004, 2017.

[Ozbudak *et al.*, 2002] Ertugrul M Ozbudak, Mukund Thattai, Iren Kurtser, Alan D Grossman, and Alexander van Oudenaarden, "Regulation of noise in the expression of a single gene.," *Nature genetics*, 31(1):69–73, 2002.

[Paninski, 2003] Liam Paninski, "Estimation of Entropy and Mutual Information," *Neural Computation*, 15:1191–1253, 2003.

[Paulsson, 2004] Johan Paulsson, "Summing up the noise in gene networks.," *Nature*, 427:415–418, 2004.

[Peccoud and Ycart, 1995] J. Peccoud and B. Ycart, "Markovian Modeling of Gene-Product Synthesis," *Theoretical Population Biology*, 48(2):222 – 234, 1995.

[Phatnani and Greenleaf, 2006] Hemali P. Phatnani and Arno L. Greenleaf, "Phosphorylation and functions of the RNA polymerase II CTD," *Genes and Development*, 20:2922–2936, 2006.

[Potter *et al.*, 2017] Garrett D. Potter, Tommy A. Byrd, Andrew Mugler, and Bo Sun, "Dynamic Sampling and Information Encoding in Biochemical Networks," *Biophysical Journal*, 112(4):795–804, 2017.

[Ptashne and Gann, 2002] M Ptashne and A Gann, *Genes and Signals*, Cold Spring Harbor Press, 2002.

[Purvis and Lahav, 2013] Jeremy E Purvis and Galit Lahav, "Encoding and Decoding Cellular Information through Signaling Dynamics," *Cell*, 152(5):945–956, 2013.

[Quiroga and Panzeri, 2009] Rodrigo Quian Quiroga and Stefano Panzeri, "Extracting information from neuronal populations: information theory and decoding approaches," *Nature Reviews Neuroscience*, 10(3):173, 2009.

[Raj *et al.*, 2006] Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi, "Stochastic mRNA Synthesis in Mammalian Cells," *PLOS Biology*, 4(10):1–13, 09 2006.

[Rhee *et al.*, 2012] Alex Rhee, Raymond Cheong, and Andre Levchenko, "The application of information theory to biochemical signaling systems," *Physical Biology*, 9(4):045011, 2012.

[Rieckh and Tkačik, 2014] Georg Rieckh and Gašper Tkačik, "Noise and information transmission in promoters with multiple internal states," *Biophysical Journal*, 106:1194–1204, 2014.

[Rieke and Baylor, 1998] F. Rieke and D. A. Baylor, "Single-photon detection by rod cells of the retina," *Reviews of Modern Physics*, 70(3):1027–1036, 1998.

[Rieke *et al.*, 1993] F. Rieke, D. Warland, and W. Bialek, "Coding Efficiency and Information Rates in Sensory Neurons," *EPL (Europhysics Letters)*, 22(2):151, 1993.

[Rockel *et al.*, 2013] Sylvie Rockel, Marcel Geertz, Korneel Hens, Bart Deplancke, and Sebastian J. Maerkl, "iSLIM: a comprehensive approach to mapping and characterizing gene regulatory networks," *Nucleic Acids Research*, 41(4):e52, 2013.

[Rosenblatt, 1957] Frank Rosenblatt, "The Perceptron - A Perceiving and Recognizing Automaton," Technical report, Cornell Aeronautical Laboratory, 1957.

[S., 1963] Bellini S., "Su di un particolare comportamento di batteri d'acqua dolce (On a unique behavior of freshwater bacteria)," Institute of Microbiology,University of Pavia, 1963.

[Samengo, 2002] Inès Samengo, "Information loss in an optimal maximum likelihood decoding," *Neural Computation*, 14(4):771–779, 2002.

[Sandelin *et al.*, 2004] Albin Sandelin, Wynand Alkema, P ar Engström, Wyeth W. Wasserman, and Boris Lenhard, "JASPAR: an open–access database for eukaryotic transcription factor binding profiles," *Nucleic Acids Research*, 32, 2004.

[Saunders *et al.*, 2006] Abbie Saunders, Leighton J Core, and John T Lis, "Breaking barriers to transcription elongation.," *Molecular cell biology*, 7:557–567, 2006.

[Savir and Tlusty, 2013] Yonatan Savir and Tsvi Tlusty, "The ribosome as an optimal decoder: A lesson in molecular recognition," *Cell*, 153:471–479, 2013.

[Schrödinger and Penrose, 1992] Erwin Schrödinger and Roger Penrose, *What is Life?: With Mind and Matter and Autobiographical Sketches*, Canto. Cambridge University Press, 1992.

[Selimkhanov *et al.*, 2014] Jangir Selimkhanov, Brooks Taylor, Jason Yao, Anna Pilko, John Albeck, Alexander Hoffmann, Lev Tsimring, and Roy Wollman, "Accurate information transmission through dynamic biochemical signaling networks.," *Science (New York, N.Y.)*, 346(6215):1370–3, 2014.

[Shannon and Weaver, 1949] C E Shannon and W Weaver, *The Mathematical Theory of Communication*, volume 27, Urbana : University of Illinois Press, 1949.

[Sharifpoor *et al.*, 2011] Sara Sharifpoor, Alex N. Nguyen Ba, Ji-Young Young, Dewald van Dyk, Helena Friesen, Alison C. Douglas, Christoph F. Kurat, Yolanda T. Chong, Karen Founk, Alan M. Moses, and Brenda J. Andrews, "A quantitative literature-curated gold standard for kinase-substrate pairs," *Genome Biology*, 12(4):R39, 2011.

[Shea and Ackers, 1985] Madeline A. Shea and Gary K. Ackers, "The OR control system of bacteriophage lambda: A physical-chemical model for gene regulation," *Journal of Molecular Biology*, 181(2):211 – 230, 1985.

[Slonim *et al.*, 2005] Noam Slonim, Gurinder S. Atwal, Gašper Tkačik, and William Bialek, "Estimating mutual information and multi–information in large networks," 2005.

[Smith and Grima, 2018] Stephen Smith and Ramon Grima, "Spatial Stochastic Intracellular Kinetics: A Review of Modelling Approaches," *Bulletin of Mathematical Biology*, pages 1–50, 2018.

[Sokolowski and Tkačik, 2015a] Thomas R Sokolowski and Gašper Tkačik, "Optimizing information flow in small genetic networks. IV. Spatial coupling," *Physical Review E*, 91(6):62710, 2015.

[Sokolowski and Tkačik, 2015b] Thomas R Sokolowski and Gašper Tkačik, "Optimizing information flow in small genetic networks. IV. Spatial coupling," *Physical Review E*, 91(6):062710, 2015.

[Sokolowski *et al.*, 2016] Thomas R Sokolowski, Aleksandra M Walczak, William Bialek, and Gašper Tkačik, "Extending the dynamic range of transcription factor action by translational regulation," *Physical Review E*, 93(2):022404, 2016.

[Spencer *et al.*, 2009] Sabrina L. Spencer, Suzanne Gaudet, John G. Albeck, John M. Burke, and Peter K. Sorger, "Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis," *Nature*, 459(7245):428–432, 2009.

[Strong *et al.*, 1998] S. Strong, Roland Koberle, Rob de Ruyter van Steveninck, and William Bialek, "Entropy and Information in Neural Spike Trains," *Physical Review Letters*, 80:197–200, 1998.

[Swain *et al.*, 2002] Peter S Swain, Michael B Elowitz, and Eric D Siggia, "Intrinsic and extrinsic contributions to stochasticity in gene expression.," *Proceedings of the National Academy of Sciences of the United States of America*, 99:12795–12800, 2002.

[Tagkopoulos *et al.*, 2008] Ilias Tagkopoulos, Yir-Chung Liu, and Saeed Tavazoie, "Predictive Behavior Within Microbial Genetic Networks," *Science*, 320(5881):1313–1317, 2008.

[Thiermann *et al.*, 2018] Ryan Thiermann, Alison Sweeney, and Arvind Murugan, "Information content of downwelling skylight for non-imaging visual systems," preprint on bioRxiv, 2018.

[Thomas and Eckford, 2016] P. J. Thomas and A. W. Eckford, "Capacity of a Simple Intercellular Signal Transduction Channel," *IEEE Transactions on Information Theory*, 62(12):7358–7382, 2016.

[Tkačik and Bialek, 2016] Gašper Tkačik and William Bialek, "Information Processing in Living Systems," *Annual Review of Condensed Matter Physics*, 7(1):89–117, 2016.

[Tkačik *et al.*, 2008] Gašper Tkačik, Curtis G Callan, and William Bialek, "Information flow and optimization in transcriptional regulation.," *Proceedings of the National Academy of Sciences of the United States of America*, 105(34):12265–70, 2008.

[Tkačik *et al.*, 2015] Gašper Tkačik, Julien O. Dubuis, Mariela D. Petkova, and Thomas Gregor, "Positional Information, Positional Error, and Readout Precision in Morphogenesis: A Mathematical Framework," *Genetics*, 199(1):39–59, 2015.

[Tkačik and Walczak, 2011] Gašper Tkačik and Aleksandra Walczak, "Information transmission in genetic regulatory networks: a review," *Journal of Physics: Condensed Matter*, 23, 2011.

[Tkačik *et al.*, 2009a] Gašper Tkačik, Aleksandra M Walczak, and William Bialek, "Optimizing information flow in small genetic networks," *Physical Review E*, 80(3):031920, 2009.

[Tkačik *et al.*, 2009b] Gašper Tkačik, Aleksandra M Walczak, and William Bialek, "Optimizing information flow in small genetic networks," *Physical Review E*, 80(3):031920, 2009.

[Tkačik *et al.*, 2012] Gašper Tkačik, Aleksandra M Walczak, and William Bialek, "Optimizing information flow in small genetic networks. III. A self-interacting gene," *Physical Review E*, 85(4):41903, 2012.

[Tkacik *et al.*, 2012] Gašper Tkacik, Aleksandra M Walczak, and William Bialek, "Optimizing information flow in small genetic networks. III. A self-interacting gene," *Phys. Rev. E*, 85(4):041903, 2012.

[Tkačik *et al.*, 2008a] Gašper Tkačik, Curtis G Callan, and William Bialek, "Information capacity of genetic regulatory elements.," *Physical review. E, Statistical, nonlinear, and soft matter physics*, 78:011910, 2008.

[Tkačik *et al.*, 2008b] Gašper Tkačik, Thomas Gregor, and William Bialek, "The role of input noise in transcriptional regulation," *PLoS ONE*, 3, 2008.

[Todeschini *et al.*, 2014] Anne-Laure Todeschini, Adrien Georges, and Reiner a Veitia, "Transcription factors: specific DNA binding and specific gene regulation.," *Trends in genetics*, 30:211–9, 2014.

[Tostevin and ten Wolde, 2009] Filipe Tostevin and Pieter Rein ten Wolde, "Mutual Information between Input and Output Trajectories of Biochemical Networks," *Phys. Rev. Lett.*, 102:218101, 2009.

[Tostevin and Ten Wolde, 2010] Filipe Tostevin and Pieter Rein Ten Wolde, "Mutual information in time-varying biochemical systems," *Physical Review E*, 81(6):061917, 2010.

[Traverse *et al.*, 1992] S Traverse, N Gomez, H Paterson, C Marshall, and P Cohen, "Sustained activation of the mitogen-activated protein (MAP) kinase cascade may be required for differentiation of PC12 cells. Comparison of the effects of nerve growth factor and epidermal growth factor.," *The Biochemical journal*, 288 (Pt 2)(2):351–5, 1992.

[Tudelska *et al.*, 2017] Karolina Tudelska, Joanna Markiewicz, Marek Kochańczyk, Maciej Czerkies, Wiktor Prus, Zbigniew Korwek, Ali Abdi, Sławomir Błoński, Bogdan Kaźmierczak, and Tomasz Lipniacki, "Information processing in the NF-κB pathway," *Scientific Reports*, 7(1):15926, 2017.

[Tuğrul *et al.*, 2015] Murat Tuğrul, Tiago Paixão, Nicholas H. Barton, and Gašper Tkačik, "Dynamics of Transcription Factor Binding Site Evolution," *PLOS Genetics*, 11(11):1–28, 11 2015.

[Tyas *et al.*, 2000] L Tyas, V A Brophy, A Pope, A J Rivett, and J M Tavaré, "Rapid caspase-3 activation during apoptosis revealed using fluorescence-resonance energy transfer.," *EMBO reports*, 1(3):266–70, 2000.

[Uda *et al.*, 2013] Shinsuke Uda, Takeshi H Saito, Takamasa Kudo, Toshiya Kokaji, Takaho Tsuchiya, Hiroyuki Kubota, Yasunori Komori, Yu-ichi Ozaki, and Shinya Kuroda, "Robustness and Compensation of Information Transmission of Signaling Pathways," *Science*, 341(6145):558–561, 2013.

[Van Kampen, 2007] N.G. Van Kampen, "Stochastic Processes in Physics and Chemistry," 2007.

[Varela *et al.*, 1974] F.G. Varela, H.R. Maturana, and R. Uribe, "Autopoiesis: The organization of living systems, its characterization and a model," *Biosystems*, 5(4):187 – 196, 1974.

[Vasicek, 1973] Oldrich A. Vasicek, "A note on using cross-sectional information in bayesian estimation of security betas," *The Journal of Finance*, 28(5):1233–1239, dec 1973.

[Voliotis *et al.*, 2018] Margaritis Voliotis, Kathryn L. Garner, Hussah Alobaid, Krasimira Tsaneva-Atanasova, and Craig A. McArdle, "Gonadotropin-releasing hormone signaling: An information theoretic approach," *Molecular and Cellular Endocrinology*, 463:106 – 115, 2018.

[Voliotis *et al.*, 2014] Margaritis Voliotis, Rebecca M. Perrett, Chris McWilliams, Craig A. McArdle, and Clive G. Bowsher, "Information transfer by leaky, heterogeneous, protein kinase signaling systems," *Proceedings of the National Academy of Sciences*, 111(3):E326–E333, 2014.

[Walczak *et al.*, 2010a] Aleksandra M Walczak, Gašper Tkačik, and William Bialek, "Optimizing information flow in small genetic networks. II. Feed-forward interactions," *Physical Review E*, 81(4):041905, 2010.

[Walczak *et al.*, 2010b] Aleksandra M Walczak, Gašper Tkačik, William Bialek, and Joseph Henry, "Optimizing information flow in small genetic networks. II. Feed-forward interactions," *Physical Review E*, 81(4):041905, 2010.

[Wunderlich and Mirny, 2009] Zeba Wunderlich and Leonid A. Mirny, "Different gene regulation strategies revealed by analysis of binding motifs," *Trends in Genetics*, 25:434–440, 2009.

[Yatsenko *et al.*, 2015] Dimitri Yatsenko, Krešimir Josić, Alexander S. Ecker, Emmanouil Froudarakis, R. James Cotton, and Andreas S. Tolias, "Improved Estimation and Interpretation of Correlations in Neural Circuits," *PLOS Computational Biology*, 11(3):e1004083, 2015.