Gene regulation across scales – how biophysical constraints shape evolution

by

Rok Grah

July, 2020

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 Rok Grah, titled *Gene regulation across scales – how biophysical constraints shape evolution*, is approved by:

Supervisor: Prof. Călin Guet, IST Austria, Klosterneuburg, Austria

Signature: _____

Supervisor: Prof. Gašper Tkačik, IST Austria, Klosterneuburg, Austria

Signature: _____

Committee Member: Prof. Nick Barton , IST Austria, Klosterneuburg, Austria

Signature: _____

Committee Member: Prof. John Marko, Northwestern, Chicago, USA

Signature: _____

Defense Chair: Prof. Christopher Wojtan, 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: _____

Rok Grah
July, 2020
signed page is on file

Abstract

In the thesis we focus on the interplay of the biophysics and evolution of gene regulation. We start by addressing how the type of prokaryotic gene regulation – activation and repression – affects spurious binding to DNA, also known as transcriptional crosstalk. We propose that regulatory interference caused by excess regulatory proteins in the dense cellular medium – global crosstalk – could be a factor in determining which type of gene regulatory network is evolutionarily preferred. Next, we use a normative approach in eukaryotic gene regulation to describe minimal non-equilibrium enhancer models that optimize so-called regulatory phenotypes. We find a class of models that differ from standard thermodynamic equilibrium models by a single parameter that notably increases the regulatory performance. Next chapter addresses the question of genotype-phenotype-fitness maps of higher dimensional phenotypes. We show that our biophysically realistic approach allows us to understand how the mechanisms of promoter function constrain genotype-phenotype maps, and how they affect the evolutionary trajectories of promoters. In the last chapter we ask whether the intrinsic instability of gene duplication and amplification provides a generic alternative to canonical gene regulation. Using mathematical modeling, we show that amplifications can tune gene expression in many environments, including those where transcription factor-based schemes are hard to evolve or maintain.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my supervisors, Călin and Gašper, for their patience and support. I am especially thankful for the encouraging environment and scientific freedom. I would also like to extend my thanks to the members of my thesis committee, John Marko and Nick Barton, for their support and advice.

Furthermore, I would like to thank my colleagues in Guet, Tkačik, and Barton groups for creating such a stimulating environment. In particular, I would like to thank Mato Lagator for his support and productive discussions. I would also like to thank Edelsbrunner group for all the discussions while playing table football. Moreover, I would like to thank 2015 cohort and "food people" for making these past years a wonderful experience.

I would also like to thank my friends, in particular Anže who was there already at the start of my life adventure, and my family for their constant support.
Last but not least, I would like to thank Nina – I am grateful for her constant support and encouragement which always drives me forward.

About the Author

After finishing his MSc in Physics in 2015 at the University of Ljubljana, Rok Grah joined IST Austria in September 2015. Being interested in interdisciplinary research, he joined the groups of Călin Guet and Gašper Tkačik to work on biophysical questions regarding modeling of gene regulation. For the duration of his PhD, Rok was a recipient of a DOC fellowship of the Austrian Academy of Sciences. He also received the Golden Sponge award for his teaching assistantship at IST Austria.

List of Publications

The main part of this thesis is or will be published in the following articles

1. **Grah R**, Friedlander T. The relation between crosstalk and gene regulation form revisited. *PLOS Computational Biology*. 16(2):1-24, 2020.

2. **Grah R**, Zoller B, Tkačik G. Normative models of enhancer function. *bioRxiv*. 2020. doi: 10.1101/2020.04.08.029405.

3. **Grah R**, Lagator M, Guet CC, Tkačik G. Evolving complex promoters for complex phenotypes. Manuscript. IST Austria, 2020.

4. Tomanek I*, **Grah R***, Lagator M, Andersson AMC, Bollback JP, Tkačik G, Guet CC. Gene amplification as a form of population-level gene expression regulation. *Nature Ecology & Evolution*. 4(4):612- 625, 2020.

* These authors contributed equally.

Table of Contents

List of Tables

List of Figures

1 Introduction

I would like to start the thesis with an anecdote. In my first year at IST, at a course on systems biology that was taught by my future supervisor Călin, we were discussing a paper on the repressilator [Elowitz and Leibler, 2000]. This is a genetic regulatory network, consisting of feedback loops between genes, each gene expressing a protein that represses the next gene. Learning about such an elegant system and solutions used in the paper, made me question why such systems were not explored earlier. This question goes very well in hand with other important results in science that seem completely obvious – why were not they thought of sooner? Călin's answer marked my future scientific path as it strongly influenced my decision to move towards biophysical sciences and to join his lab. Călin's answer to my question was that besides the obvious technical reasons why such systems weren't made earlier, it is that *many things that seem obvious to us now, weren't so obvious back then*. In other words, I believe that Călin was trying to tell me that in science best results are those that, in retrospect, seem obvious and trivial. Four years later, I now strongly believe that good science is about asking the right questions leading to a different perspective.

The work presented in this thesis focuses on modeling approaches for gene regulation and how these can give wider and more mechanistic understanding of the evolutionary consequences. Gene regulation, also known as regulation of gene expression, is a process of controlling production of gene products such as protein or RNA. As it allows a system to respond to an environmental change, it is a crucial process in all life. Therefore, understanding gene regulation, and with it gene expression, lies at the heart of understanding life.
The goal of modeling such processes is not to replace experiments but to offer

support where experimental approaches are difficult. Furthermore, one of the most important functions of modeling is to give rise to new questions, leading to different insights and perspective. It is then the experiments that tells us what is real. As prof. Kerševan from my undergraduate studies said, 'papir prenese vse'[1].

As a physicist coming to the world of quantitative biology and biophysics from a rather theoretical background, I believe that my time at IST was spent connecting different research areas together, all in the spirit of IST's interdisciplinarity.
As any transition to a new environment, it takes time to adapt and familiarize oneself with it[2]. Fully diving in the world of biophysics at the begining of my PhD, the work of Bintu et al [Bintu *et al.*, 2005a; Bintu *et al.*, 2005b] connected with the physicist in me. This paper shows quantitative approach to transcriptional regulation using thermodynamic models which are able to predict the expression of a gene from the DNA sequence. From my physics background I was already familiar with these type of models but had never applied them in biological context before. Why are the thermodynamic models of gene regulation important? First, they provide a highly quantitative mapping from promoter sequences to gene expression levels that is compatible with biophysical measurements. No other model is able to so accurately describe how individual nucleotides within binding sites affect gene expression. Second, thermodynamic models are based on biophysically realistic assumptions. They assume that we can use statistical mechanics to describe equilibrium probabilities of different molecules binding to the sequence of interest, and using these to describe the expression of the gene of interest. This main assumption, very basic in its core, directly leads to many biophysically realistic consequences. It means that, without further assumptions, many qualitative properties directly follow from the model itself. For example, it follows that the probability of binding is a sigmoid function of the binding energy, a realistic but often ignored fact in other models. Third, due to its realistic nature, the thermodynamic models can be used to mechanistically explain concepts widely used in evolutionary biology, such as

[1]Rough translation would be 'Making a model doesn't make it true'.
[2]see Chapter 5

epistasis. It is this type of approach that I have followed in much of my research. Due to these reasons, thermodynamic models are generally able to outperform other models of gene regulation. For example, pictorial models of gene regulation are able to give a simple understanding of the system, yet they lack quantitative power. Bioinformatics can give quantitative predictions but these are often lacking understanding of the model, or are limited by the data. Furthermore, machine learning models of gene regulations can give accurate predictions but do not give any insight and understanding of the process it is studying, thus failing one of the main points of any model – to examine and understand the system.
However, which model is best depends on the question one is trying to answer. As addressing biophysical constraints and mechanistic understanding constitutes a large portion of the thesis, the thermodynamic models lie at the center of the work presented here. Nonetheless, we also show how other models of gene regulation can be used.

Broadly speaking, the thesis explores the extent to which models of gene regulation explain cellular issues across scales, resulting from four projects, represented by four chapters. Each chapter shows how biophysical constraints limit the evolution of gene regulation and how understanding those limits can give insight into the realm of "possible" [Jacob, 1994]. Below we outline these four projects. We start with a broader system-level problem, discussing transcriptional crosstalk and its role in determining the regulatory network. We continue with the importance of understanding mechanistic details and biophysical constraints for both knowing what kind of systems can occur, and what can such systems do. We end with an evolutionary aspect of gene regulation, showing a new population level alternative to the canonical gene regulation which maintains most properties of gene regulation.

Chapter 2: The relation between crosstalk and gene regulation form revisited. Due to the large assembly of genes and regulators in a cell, erroneous binding and unbinding events called "crosstalk" could occur. For example, a non-cognate

transcription factor could bind to a promoter, wrongly activating gene expression. Such crosstalk can interfere with the gene's proper regulatory state (i.e., correct amount of expression at appropriate times) and is generally considered to be selected against. However, from an evolutionarily point of view, crosstalk can also help promote gene regulatory network evolvability. This shows that knowing how large number of regulatory proteins interact with various DNA targets is crucial in fully understanding gene regulatory networks and their dynamics.

Experimental measurements of crosstalk are possible but often limited to certain types of molecules that were priorly tagged by a fluorescent tag. Correctly estimating crosstalk would entail measuring not only all proteins binding to the DNA but doing so in a dynamical way. This means following the binding and unbinding dynamics to understand how often and which part of DNA is bound which is currently technologically very difficult. On the other hand, biophysically realistic models like thermodynamic models discussed above allow us to theoretically explore these systems. Such models were used to find that it is crucial to think of crosstalk as a global – not local – quantity which leads to qualitatively different constraints than considering crosstalk only at the level of individual gene regulatory elements.

In Chapter 2 we ask how the form of regulation, positive or negative, affects the extent of regulatory crosstalk. In particular, both positive (a gene is activated by the binding of its regulatory protein) and negative (a gene is active unless bound by its regulatory protein) regulation can lead to the same activation of a gene in response to an external signal. Due to this, researchers have pondered whether additional considerations could favour the choice of one mechanism over the other, or whether this choice is merely a coincidence ("evolutionary accident"). Different studies proposed various arguments that all concentrated on a single gene with a single regulator, regardless of the full regulatory network. Our study proposes that additional costs, such as global crosstalk which takes into account the whole network, could play a role in determining one form of regulation above the other. In other words, our work addresses a typically overlooked cost of protein production: that of regulatory interference caused by excess regulatory proteins in a dense cellular medium. The core of our results is based on using the thermodynamic

model which follows both correct and erroneous binding of regulatory proteins to the whole network of genes, with some genes having positive and some negative form of regulation.

Chapter 3: Normative models of enhancer function. In prokaryotes, thermodynamic models of gene regulation provide a highly quantitative mapping from promoter sequences to gene expression levels that is compatible with biophysical measurements. In eukaryotes, however, such accurate predictions are still missing. For example, in a set of eukaryotic promoter elements that increase transcriptional efficiency called enhancers, equilibrium models (like thermodynamic models) appear not to be adequate in describing its regulation. On the other hand, non-equilibrium models suffer from an exponential increase of complexity with increasing number of parameters, making their use quite limited.

In Chapter 3, we aim to describe minimal non-equilibrium enhancer models using the normative approach: finding such class of minimal models that optimizes so-called regulatory phenotypes. Examples of these are low transcription factor residence time, tunable cooperativity, and high specificity. The latter means decreasing transcriptional crosstalk by making transcription factor binding sites more distinct, again showing that crosstalk plays an important role in understanding gene regulatory networks. We find a class of models that differ from equilibrium models by a single parameter that introduces kinetic-proofreading scheme, thereby notably increasing the regulatory performance. Our solutions are the simplest generalization of the classic equilibrium regulatory schemes to non-equilibrium processes, thus still remaining simple enough to analyze and understand.

We further find that optimization of aforementioned regulatory phenotypes in our non-equilibrium models is in a trade off with gene expression noise, predicting bursty dynamics – an experimentally-observed hallmark of eukaryotic transcription. The modeling approach used here differs from that used in other chapters: the top-down normative approach based on optimization utilizes simple models without focusing on data fitting to lead to new insight and understanding.

Chapter 4: Evolving complex promoters for complex phenotypes. How genetic mutations (genotype) alter one or more organismal traits (phenotype) is the central problem of evolutionary biology. This genotype-phenotype (GP) mapping has been extensively studied in a range of experimental and theoretical systems, most of which indicate that the mapping is complex and non-linear. And yet, the wealth of experimentally determined maps has not resulted in comprehensive or generalizable understanding of the relationship between genotype and phenotype. In other words, we lack the ability to predict how genotype maps onto phenotype for most biological systems. One major area of focus for describing GP mapping has been the regulation of gene expression, due to its central role in enabling organisms to respond to environmental change and to coordinate inter-cellular processes. While offering unprecedented insights into how gene regulatory networks evolve, a majority of experimental work focused on a neighbourhood of only a handful of mutations away, making these descriptions local. Furthermore, main focus of most studies is the steady-state expression levels in cells. And yet, temporal dynamics of gene expression play an important role in determining how a biological system functions. For example, bistable behavior observed in various bacterial species is often enabled by having different rates at which relevant genes are turned on or off. Therefore, it is necessary to understand not only the steady state expression levels, but also how the expression dynamics (how rapidly the steady state is reached) affect organismal fitness.

In Chapter 4, we investigate complex promoters and complex phenotypes in realistic setting. We go beyond the typically studied single phenotype of a constitutive promoter and study how mutations in bacterial promoters alter gene expression dynamics between different environments. To achieve this goal, we extended the classical thermodynamic model that can accurately predict GP mapping for gene expression dynamics in a regulated bacterial promoter. This biophysical model allowed us to understand how the mechanisms of promoter function constrain GP mapping, how those constraints changed depending on whether we considered

only steady-state expression or the dynamics of expression, and how they affect the evolutionary trajectories of promoters.

Using a biophysically realistic modeling approach, we were able to gain new mechanistic insights into the function of a complex promoter, understanding not only what mutations do but also why. This is critical for developing a more predictive understanding of evolution, as it enables generalizing GP maps beyond a specific system being studied to a range of other systems that share similar features (regulated bacterial promoters).

Chapter 5: Gene amplification as a form of population-level gene expression regulation. Natural environments change periodically or stochastically with frequent or very rare fluctuations and life crucially depends on the ability to respond to such changes. Gene regulatory networks have evolved into an elaborate mechanism for such adjustments as populations were repeatedly required to cope with specific environmental changes. However, due to low single base-pair mutation rates, complex promoters cannot easily evolve on ecological time scales.

In Chapter 5, we ask whether the intrinsic instability of gene duplication and amplification provides a generic alternative to canonical gene regulation. By real-time monitoring of gene copy number mutations in *E. coli*, we show that gene duplications and amplifications enable adaptation to fluctuating environments by rapidly generating copy number, and hence expression level, polymorphism. This 'amplification-mediated gene expression tuning' occurs on timescales similar to canonical gene regulation and can deal with rapid environmental changes. With mathematical modeling, using population genetics, we show that amplifications also tune gene expression in stochastic environments where transcription factor-based schemes are hard to evolve or maintain. The fleeting nature of gene amplifications gives rise to a generic population-level mechanism that relies on genetic heterogeneity to rapidly tune expression of any gene, without leaving any genomic signature.

2 The relation between crosstalk and gene regulation form revisited

Genes differ in the frequency at which they are expressed and in the form of regulation used to control their activity. In particular, positive or negative regulation can lead to activation of a gene in response to an external signal. Previous works proposed that the form of regulation of a gene correlates with its frequency of usage: positive regulation when the gene is frequently expressed and negative regulation when infrequently expressed. Such network design means that, in the absence of their regulators, the genes are found in their least required activity state, hence regulatory intervention is often necessary. Due to the multitude of genes and regulators, spurious binding and unbinding events, called "crosstalk", could occur. To determine how the form of regulation affects the global crosstalk in the network, we used a mathematical model that includes multiple regulators and multiple target genes. We found that crosstalk depends non-monotonically on the availability of regulators. Our analysis showed that excess use of regulation entailed by the formerly suggested network design caused high crosstalk levels in a large part of the parameter space. We therefore considered the opposite 'idle' design, where the default unregulated state of genes is their frequently required activity state. We found, that 'idle' design minimized the use of regulation and thus minimized crosstalk. In addition, we estimated global crosstalk of *S. cerevisiae* using transcription factors binding data. We demonstrated that even partial network data could suffice to estimate its global crosstalk, suggesting its applicability to additional organisms. We found that *S. cerevisiae* estimated crosstalk is lower than that of a random network, suggesting that natural selection reduces crosstalk. In summary, our study highlights a new type of protein production cost which is typically overlooked: that of regulatory interference caused by the presence of excess regulators in the cell. It demonstrates the importance of whole-network descriptions, which could show effects

missed by single-gene models.

Published as **Grah R**, Friedlander T. The relation between crosstalk and gene regulation form revisited. *PLOS Computational Biology*. 16(2):1-24, 2020.

Some changes have been made to the text in order to integrate it into this thesis.

2.1 Introduction

Gene regulatory networks can employ different architectures that seemingly realize the same input-output relation. There is a basic dichotomy of gene regulation into positive and negative control. A gene controlled by positive regulation is, by default, not expressed and requires binding of an activator to its operator to induce it. In contrast, a gene controlled by negative regulation, is expressed by default, unless a repressor binds its operator and attenuates its activity. While a gene can be regulated using either mode, researchers have pondered whether additional considerations could favor the choice of one mechanism over the other, or whether this choice is merely a coincidence ("evolutionary accident"). Throughout the years, this question was addressed using different approaches. The seminal work of Michael Savageau [Savageau, 1974; Savageau, 1977; Savageau, 1983] proposed the so-called "Savageau demand rule", namely, that genes encoding frequently needed products ("high-demand") are often regulated by activators. Conversely, genes whose products are only needed sporadically ("low-demand"), tend to be regulated by repressors. Savageau argued that the intensity of selection depends on the extent to which the regulatory construct is used (later called the "use it or lose it" principle [Gerland and Hwa, 2009]). When infrequently used (as in activator regulating a low-demand or a repressor regulating a high-demand gene), selection to preserve is weak, rendering it unlikely to survive [Savageau, 1998]. A later evolutionary analysis mathematically formulated the problem as selection in an alternating environment and found the exact conditions under which the Savageau demand rule is expected to hold [Gerland and Hwa, 2009].

Recently, a comprehensive survey of regulatory topologies in *E. coli* and *B. subtilis*, found agreement between the experimentally observed topologies and their satisfaction of dynamic constraints, as verified in simulations. The authors found exceptions to the Savageau demand rule and proposed that evolutionary processes randomly pick a regulatory topology out of the many possible ones meeting the organism physiological constraints [Kumar Prajapat *et al.*, 2016].

An alternative reasoning for the observed correlation between a gene's demand

and its form of regulation was proposed using a biophysical, rather than evolutionary argument [Shinar *et al.*, 2006; Sasson *et al.*, 2012]. If a high-demand gene is regulated by an activator and a low-demand gene is regulated by a repressor, their regulatory binding sites are mostly occupied and protected from spurious binding of foreign regulators that could interfere with the gene's regulatory state. However, if this reasoning applies not just to one gene, but to many of them, it would also entail extravagant use of regulators [Kumar Prajapat *et al.*, 2016]. This would place heavy demands on protein expression systems, associated with reduced growth rate [Novick and Weiner, 1957; Koch, 1983; Kurland and Dong, 1996; Dekel and Alon, 2005; Kafri *et al.*, 2016].

While the above-mentioned studies examined the significance of regulatory architectures from different perspectives, they all concentrated on a single gene with a single regulator, regardless of the full regulatory network. It remains unanswered whether the choice of positive or negative regulation for a gene with low- or high-demand could have additional costs for the entire network. Specifically, transcription factors are known to have limited specificity and bind a variety of DNA targets, besides their cognate binding sites [Von Hippel *et al.*, 1974; Johnson *et al.*, 2005; Maerkl and Quake, 2007; Wunderlich and Mirny, 2009; Rockel *et al.*, 2012; Yona *et al.*, 2018]. The probability of such binding events naturally depends on their concentrations [Gerland *et al.*, 2002; Bintu *et al.*, 2005a]. Here, we revisit the argument that the Savageau demand rule minimizes transcriptional crosstalk, by accounting for crosstalk of multiple genes simultaneously, rather than the single-gene crosstalk considered earlier.

We use a mathematical global crosstalk model [Friedlander *et al.*, 2016], which was built upon the well-established thermodynamic model of gene regulation to calculate transcription factor (TF)-DNA interactions [Shea and Ackers, 1985; Von Hippel and Berg, 1986; Gerland *et al.*, 2002; Bintu *et al.*, 2005a; Kinney *et al.*, 2010; Lässig, 2007; Haldane *et al.*, 2014]. We have previously shown that while crosstalk affecting a particular gene can be reduced by different means, it always comes at the cost of elevating crosstalk in other genes [Friedlander *et al.*, 2016]. In contrast, the *global* crosstalk cannot be reduced below a certain threshold. Here, we analyze global

crosstalk levels under different regulatory strategies: either positive or negative regulation. We compare two extreme designs: a 'busy' one that implements the Savageau demand rule, in which a high (low)-demand gene is always regulated by an activator (repressor) and an opposite 'idle' design, in which a high (low)-demand gene is always regulated by a repressor (activator). We find that the 'busy' design maximizes regulator usage, whereas the 'idle' one minimizes it. We analyze the dependence of global crosstalk on the abundance of regulatory proteins in the cellular environment and find the exact conditions under which either 'idle' or 'busy' design minimizes crosstalk. We conclude that under most biologically plausible parameter values, the 'idle' design should yield lower *global* transcriptional crosstalk.

This chapter begins with the introduction of a general symmetric model for the analysis of transcriptional crosstalk in a many-TFs-many-genes setting, with combination of positive and negative regulation. We show that global crosstalk levels directly depend on the fraction of TFs in use and only indirectly on the choice of activation or repression as the form of regulation. We then analyze TF usage and crosstalk levels of the two extreme designs, i.e., 'busy' and 'idle' and then construct numerical simulations of a more general asymmetric gene usage model, that are in agreement with the analytical result. Lastly, we discuss the challenges in crosstalk calculation for real gene regulatory networks, in particular, the possible effect of data incompleteness, and show an example using *S. cerevisiae* TF data.

2.2 Results

2.2.1 A model of gene regulation using a combination of activators and repressors

We begin by introducing and analyzing a basic model with a simple form of gene regulation, assuming that each gene is regulated by a single transcription factor. We also assume identical properties for all genes and all transcription factors. Later we relax some of these simplifying assumptions and consider additional more complex gene regulatory architectures. We summarize these model variants in the main

text, and their full descriptions can be found in Section 2.5.1. We consider a cell that has a total of M genes, each of which is transcriptionally regulated to be either active or inactive. We assume that each gene is regulated by a single unique TF species - its cognate one. Each gene has a short DNA binding site to which its cognate TF binds. A fraction $0 \leq p \leq 1$ of the genes is regulated by activators and the remaining $1 - p$ fraction of genes is regulated by repressors. When no activator is bound, activator-regulated genes are inactive (or active at a low basal level) and only become active once an activator TF binds their binding site. In contrast, repressor-regulated genes are active, unless a repressor TF binds their binding site and inhibits their activity (Fig 2.1A). We assume that different environmental conditions require the activity of different subsets of the M genes. We assume however that all these subsets include an equal q proportion of genes $0 \leq q \leq 1$ that is needed to be active. The remaining $1 - q$ proportion should be inactive. These activity states are regulated by the binding and unbinding of the TFs specialized for these genes. We assume that only a subset of TFs necessary to maintain the desired regulatory pattern, is available to bind and regulate these genes. However, TFs often have limited specificity to their DNA targets and can occasionally bind slightly different sequences, albeit with lower probability [Maerkl and Quake, 2007; Wunderlich and Mirny, 2009; Sarai and Takeda, 1989; Fordyce *et al.*, 2010; Afek *et al.*, 2014; Yona *et al.*, 2018].

We define 'crosstalk' as the average fraction of genes found in any erroneous regulatory state: a gene that should be activated (repressed) but is not, because its cognate TF fails to bind or because its binding site which should remain unoccupied is bound by a non-cognate TF and also events of activation (repression) in response to a non-cognate signal (or in a wrong dynamic range) because a non-cognate activator (repressor) binds instead of the cognate one - see summary in Fig 2.1C. To quantitate the probability of these events, we use the thermodynamic model of gene regulation [Shea and Ackers, 1985; Von Hippel and Berg, 1986; Gerland *et al.*, 2002; Bintu *et al.*, 2005a; Landman *et al.*, 2017]. Importantly, this model assumes that gene activity is proportional to the equilibrium binding probability of its transcription factor to its regulatory binding site. Hence, we use a quasi-static, rather than kinetic,

Activity	Regulated by	Proportion of genes using this regulatory strategy
active	activator	a, where $a \leq q, p$
active	repressor	$q - a$
inactive	activator	$p - a$
inactive	repressor	$(1 - p) - q + a$

Table 2.1: **We distinguish 4 sets of genes according to their state of activity (active/ inactive) and form of regulation (activation/ repression).**

description where we assume that the system switches between different states of equilibrium. A mathematical model for crosstalk for the special case in which all TFs are activators ($p = 0$) was derived and analyzed in our previous work [Friedlander *et al.*, 2016]. Here, we analyze a more general model with a combination of activators and repressors. The reader can find the details of both models in Section 2.5.1.

Both activity and inactivity of genes can be attained by means of either activator or repressor regulation. Accordingly, our model distinguishes between four sets of genes (see Table 2.1 and Fig 2.1B):

The probability that a particular gene i is in the x_{bound} or x_{unbound} crosstalk states, depends on the concentration of competing non-cognate TFs, C_j, $j \neq i$ and on the number of mismatches, d_{ij}, between each competing TF j and the regulatory binding site of gene i, where we assume equal energetic contributions of all positions in the binding site. Consequently, the similarity between binding sites regulated by distinct TFs is a major determinant of crosstalk. We introduce an average measure of similarity between binding site i and all other binding sites $j \neq i$ [Friedlander *et al.*, 2016]:

$$S_i \equiv \langle e^{-\epsilon d_{ij}} \rangle_{P(d)} = \frac{1}{C} \sum_{j \neq i} C_j e^{-\epsilon d_{ij}} = \frac{1}{T} \sum_{j \neq i} e^{-\epsilon d_{ij}}. \tag{2.1}$$

As only a subset of the genes is regulated, the summation of only the corresponding subset of TFs available to bind is taken. S_i is defined as the average of the Boltzmann factors, $e^{-\epsilon d_{ij}}$, taken over the distribution of mismatch values $P(d)$

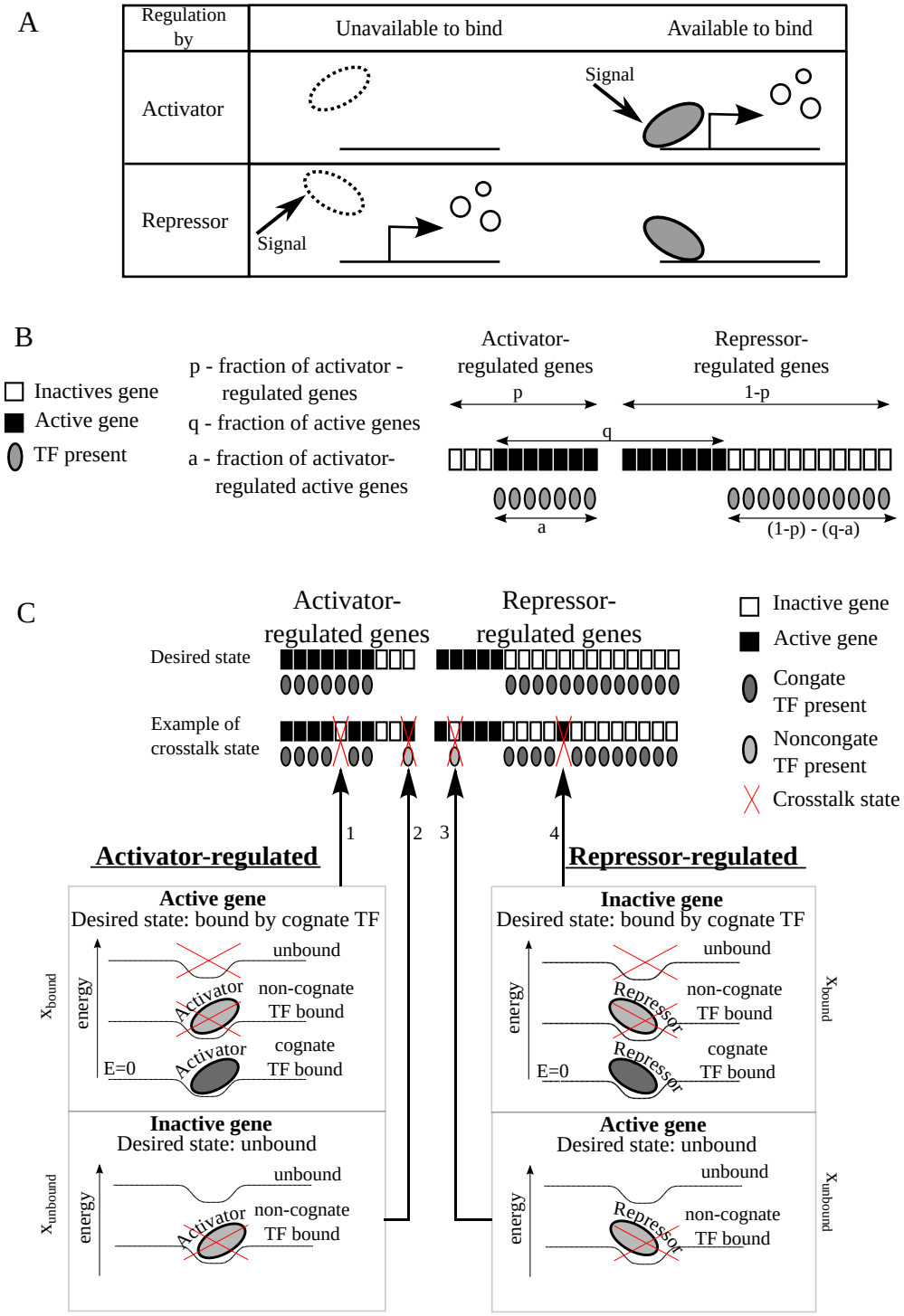

Figure 2.1: (Continued on the following page.)

between binding sites i and j, $\forall j$. In the last equality in Eq 2.1, we assume that all available TFs are found in equal concentrations $C_j = C/T, \forall j$, where C is the total TF concentration and T is the number of distinct TF species available. Eq 2.1

Figure 2.1: **Gene regulation can employ different combinations of activators and repressors to implement the same gene expression pattern. (A)** A signal can cause gene activation by either positive (first row) or negative (second row) control. **(B)** We consider a total of M genes in a cell, of which a fraction $0 \leq p \leq 1$ is regulated by activators, and the remaining $1 - p$ is regulated by repressors. Assume that only a fraction $q < 1$ of these genes should be active under certain conditions (black squares), while the remaining genes should be inactive (white squares). In general, $a \leq q, p$ of this q proportion is activator-regulated and $q - a$ is repressor-regulated. Here, we illustrate all four cases of active/inactive genes regulated by activator/repressor and define all the variables. Gray ellipses represent TFs (of either type) required to maintain the regulatory state of the genes. **(C)** Different genes are regulated by different TF species, where TF specificity is determined by short regulatory DNA sequences (binding sites) adjacent to the gene. Each such binding site can be at different levels of energy depending on its occupancy. It is in the lowest $E = 0$ (most favorable) level when bound by its cognate TF; it can be in a variety of higher energy levels if a non-cognate TF binds or if the site remains unoccupied (lower panel). The upper panel shows the crosstalk-free 'desired state' (first row), where each TF binds its cognate target. Below (second row), four different possibilities in which binding of a TF to non-cognate binding sites or failure to bind lead to crosstalk. An activator-regulated gene should ideally be regulated by its cognate activator (right-inclined ellipse), in order to become active. If this cognate TF fails to bind when the gene should be active (1), or if another TF binds when the gene should remain inactive (2), we consider this as crosstalk. For a repressor-regulated gene, crosstalk states occur when a non-cognate repressor binds when the gene should be active (3), or if the cognate repressor fails to bind when the gene should be inactive (4). We present cognate TFs by dark gray and non-cognate ones by light gray. Activators are represented by right-inclining and repressors by left-inclining ellipses. Crosstalk states are marked by red crosses.

can also be used for general TF concentrations, as observed in experiments. We demonstrate this calculation in Section 2.5.8. We found that allowing different concentrations for activators and repressors does not reduce crosstalk below this equal concentration scheme (Section 2.5.1). We also assume full symmetry between binding sites i, such that $S_i = S \; \forall i$. A numerical analysis of a more general case with non-uniform S_i values can be found in Fig 2.6 in Section 2.5.1. The value of S can be either estimated using binding site data (see below) or analytically calculated under different assumptions on the pairwise mismatch distribution $P(d)$. In the following, we use rescaled variables: $s = S \cdot M$ for rescaled similarity between binding sites, the fraction of available TFs ($t = T/M$) and the rescaled total TF concentration

($c = C/M$).

We distinguish crosstalk states of genes whose desired state of activity requires unoccupied binding sites (x_{unbound}), and those requiring occupation by a cognate regulator (x_{bound}). x_{unbound} crosstalk includes the cases of an activator-regulated gene that should remain inactive as well as that of a repressor-regulated gene that should be active, both requiring an unoccupied binding site. For these genes, the cognate TF is not available to bind and any binding event by another (non-cognate) regulator is considered crosstalk. x_{bound} crosstalk includes both an activator-regulated gene that should be active and a repressor-regulated one that should be inactive. For these, crosstalk states occur either if the binding site remains unbound or if it is occupied by a non-cognate regulator, in which case, the regulatory state is not guaranteed. For illustration of all possible crosstalk states, see Fig 2.1C. Using equilibrium statistical mechanics, these crosstalk probabilities for a single gene i are [Von Hippel and Berg, 1986; Gerland *et al.*, 2002; Friedlander *et al.*, 2016]:

$$x_{\text{bound}} = \frac{e^{-E_a} + \sum_{j\neq i} C_j e^{-\epsilon d_{ij}}}{C_i + e^{-E_a} + \sum_{j\neq i} C_j e^{-\epsilon d_{ij}}} = \frac{e^{-E_a} + cs}{c/t + e^{-E_a} + cs} \tag{2.2}$$

$$x_{\text{unbound}} = \frac{\sum_{j\neq i} C_j e^{-\epsilon d_{ij}}}{e^{-E_a} + \sum_{j\neq i} C_j e^{-\epsilon d_{ij}}} = \frac{cs}{e^{-E_a} + cs}. \tag{2.3}$$

E_a is the energy difference between cognate bound and unbound states. The expression $\sum_{j\neq i} C_j e^{-\epsilon d_{ij}}$ captures the sum of all interactions of binding site i with foreign regulators.

2.2.2 Global crosstalk depends on the use of regulators

We define the global crosstalk, X, of a cell as the average fraction of genes found in any of the crosstalk states. For a given value of a, we average over different choices of a active genes out of the p activator-regulated and over different choices of $q - a$ out of the $(1 - p)$ repressor-regulated proportions. The weighted sum over these four types of contributions provides the average total crosstalk, X, of the whole system:

$$X = \overbrace{a \cdot x_{\text{bound}} + (p-a) \cdot x_{\text{unbound}}}^{\text{Contribution of activator-regulated genes}} + \overbrace{(q-a) \cdot x_{\text{unbound}} + (1-p-q+a) \cdot x_{\text{bound}}}^{\text{Contribution of repressor-regulated genes}} \qquad (2.4)$$
$$= t \cdot x_{\text{bound}} + (1-t) \cdot x_{\text{unbound}}.$$

As Eq 2.4 shows, X simply depends on the fraction of available TF species $t = 1 - p - q + 2a$, where $t = T/M$, regardless of their role as activators or repressors. Importantly, global crosstalk does not directly depend on the fraction of active genes q. This is a generalization of the result obtained in [Friedlander *et al.*, 2016], where the special cases of $t = q$ (all TFs are activators) and $t = 1 - q$ (all TFs are repressors) were studied. To obtain a lower bound on crosstalk values for given similarity, s, and fraction of available TFs, t, we substitute the expressions for x_{bound} and x_{unbound} (Eq 2.2 and Eq 2.3 into Eq 2.4). We then minimize X with respect to the total TF concentration, c. Such minimization is possible because global crosstalk balances between some binding sites that should be bound and others that should be unbound. For the former, higher c increases their chance to be bound by their cognate TFs and thus reduces crosstalk. For the latter, their cognate TF is absent and thus higher c increases their chance to be bound by foreign TFs, namely increases crosstalk. We then obtain the expression for minimal crosstalk:

$$X^*(t,s) = t\left(-s(1-t) + 2\sqrt{s(1-t)}\right). \qquad (2.5)$$

Hence, the lower bound on crosstalk X^* only depends on two macroscopic variables: s (similarity between binding sites) and t (fraction of available TFs). The higher the similarity s, the larger the resulting crosstalk X^*, where to first order, $X^* \sim \sqrt{s}$ (Fig 2.2A). The dependence on t is more complicated and non-monotonic: for low t values, $t < t^*(s)$ (we show in Section 2.5.1 that $t^*(s) \geq 2/3$), X^* increases with t. Intuitively, the number of available TF species positively correlates with the number of crosstalk opportunities. Contrary to this intuition, for high TF usage beyond the threshold value t^*, we find the opposite trend, where X^* *decreases* with increasing TF usage, t. This non-monotonic dependence of X^* on t comes about since the optimal concentration $c^*(s,t)$ is tailored specifically for each t value. That is because

the relative weight of binding sites that should be bound vs. those that should be unbound, shifts with t. High TF usage though always comes at the cost of an exponential increase in the optimal TF concentration, c^*, (Eq. S4), where for high s values, c^* diverges to infinity $c^* \to \infty$ (see Fig 2.2B). We discuss below the biological relevance of the high t regime. We derived this model for the simple regulatory network shown in Fig 2.1C. Eqs 2.2-2.5 can be analogously derived for more complex network architectures, as we demonstrate in Section 2.5.10.

2.2.3 Mode of regulation affects global crosstalk because it affects TF usage

A particular gene activity pattern can be obtained by different combinations of positive and negative regulation, yielding seemingly identical gene functionality. One may then ask whether these various TF-gene associations differ in the resulting global crosstalk. Following Eq 2.5, crosstalk only depends on the fraction of available TF species, t, regardless of the underlying association of a gene with either activator or repressor. It is thus sufficient to consider how different regulatory strategies affect TF usage, rather than analyzing the whole network architecture, thereby significantly simplifying the analysis. Using our model, we calculate the global crosstalk for any combination of the fraction of active genes, q, with any mixture of activators and repressors defined by p, thereby covering all possible gene-regulator associations with either activators or repressors. While each point represents a fixed fraction of active genes, this model can also be used to study a varying number of active genes, by taking a distribution of points over the q-axis (see Section 2.5.7 for an example). Specifically, we focus on the two extreme gene-regulator associations, which we call the 'busy' and 'idle' network designs. The 'busy' design means that gene regulation is operative most of the time. It is implied by the "Savageau demand rule" [Savageau, 1977], because the gene's default state of activity is not its commonly needed state. Under the opposite 'idle' design, the default state of each gene is its more commonly needed regulatory state. Hence, regulation is inoperative most of the time (see Fig 2.2C). Hybrids of these two extreme designs are also possible.

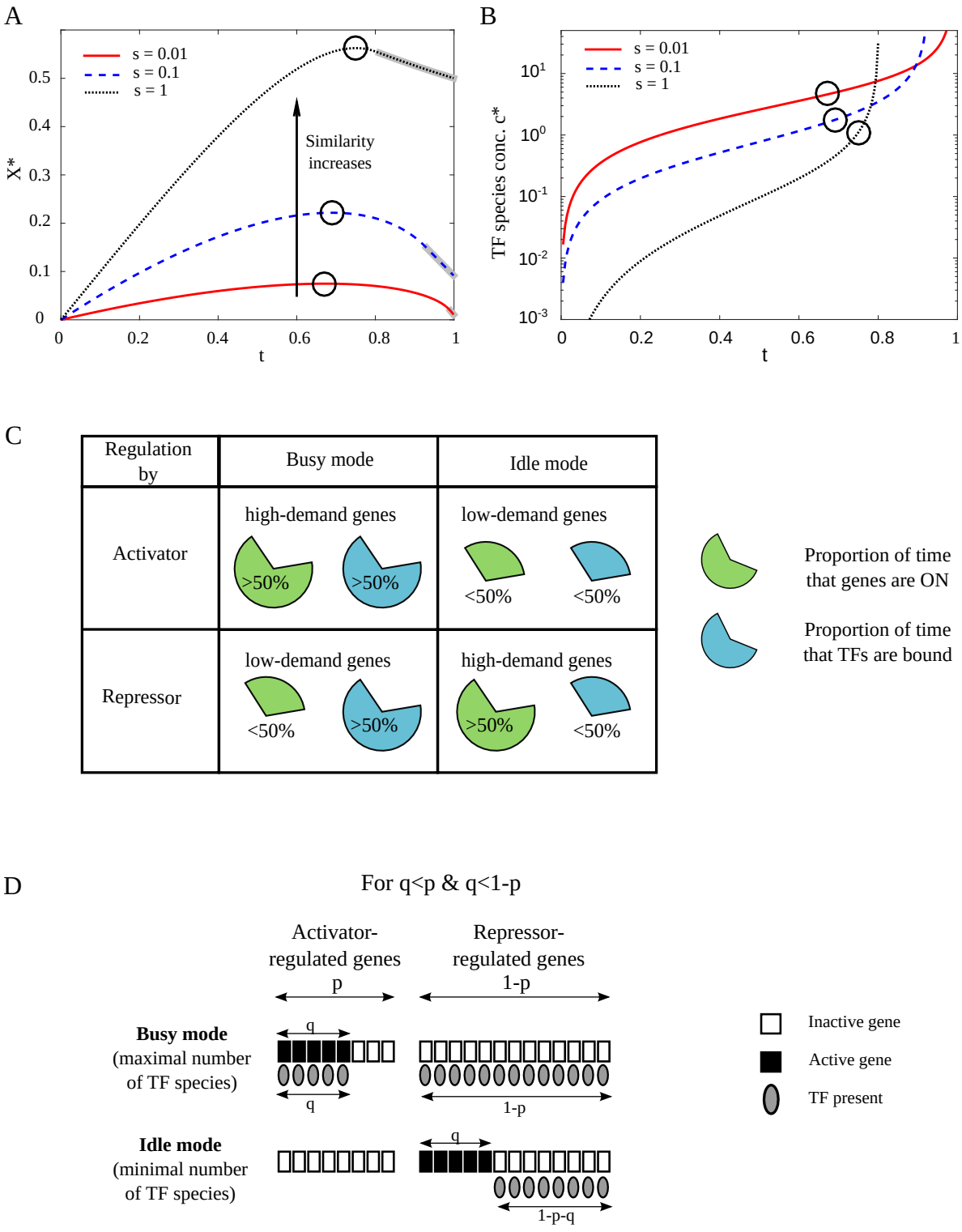

Figure 2.2: (Continued on the following page.)

To represent the 'busy' design, we associate as much of the q active proportion as possible with activators, and only if the total fraction of activators is smaller than

Figure 2.2: **Crosstalk depends on the fraction of available TFs, which varies between regulatory designs. (A)** We illustrate minimal crosstalk, X^*, vs. t, the fraction of available TFs, for different values of similarity, s. In most of the parameter regime (for $t < t^*$, $t^* \geq 2/3$), minimal crosstalk, X^*, increases with t. Black circles denote the maxima of the curves. Crosstalk monotonically increases with similarity between binding sites. The anomalous regime where TF concentration needed to minimize crosstalk mathematically diverges to infinity, is gray-shaded around the curves. **(B)** The optimal TF concentration, c^*, needed to minimize crosstalk increases sharply with t. c^* diverges to infinity at the boundary with the anomalous regime, which for high similarity s, occurs already at lower TF usage t. Circles represent the maximal X^* values for each curve (as in (A)). **(C)** Different genes are expressed to different extents, where here, we grossly classify them as either high- (more than half of the time) or low-demand (less than half). If a high-demand gene is regulated by an activator or if a low-demand gene is regulated by a repressor, demand for the regulator will be high ('busy design'). Conversely, if the same high-demand gene is regulated by a repressor and the low-demand gene is regulated by an activator, the regulator is only required for a small fraction of the time ('idle design'). **(D)** Each of the q active genes and $1-q$ inactive genes can be assigned either positive or negative regulation. We illustrate the two extremes maximizing (minimizing) TF usage: in the 'busy' ('idle') design, as many active genes as possible are assigned positive (negative) regulation and as many inactive genes as possible are assigned negative (positive) regulation. The scheme shows an example with the proportion of active genes q, the proportion of activator-regulated genes p and the proportion of repressor-regulated genes $(1-p)$ such that $q \leq p, 1-p$. Other combinations are shown in Fig 2.9 in Section 2.5.3.

the fraction of active genes ($p < q$), the remaining $q - p$ proportion is regulated by repressors. Thus the fraction of activator-regulated active genes is $a = \min(p, q)$. Conversely, under the 'idle' design, we associate as much of the q active proportion as possible with repressors. Only if the fraction of repressors is smaller than the proportion of active genes ($1 - p < q$), the remaining active genes pursue positive regulation, hence $a = q - \min((1-p), q)$. The corresponding fractions of TFs in use (including both activators and repressors) in these two extremes are then:

$$t_{\text{busy}} = 1 - |p - q|, \tag{2.6}$$

$$t_{\text{idle}} = |1 - p - q|. \tag{2.7}$$

In Fig 2.2D, we illustrate regulation following these two extreme designs. The TF assignments defined in Eq 2.6 and Eq 2.7 are the two extremes in TF usage. Namely, for any general regulatory scheme, the fraction of TFs needed to regulate a given fraction of genes q is $t_{\text{idle}} \leq t \leq t_{\text{busy}}$ (see Section 2.5.2 for formal proof). In Fig 2.3A, we illustrate the difference in the fraction of available TFs between the two extreme designs $\Delta t = t_{\text{busy}} - t_{\text{idle}} = 1 - |p - q| - |1 - p - q| > 0$, demonstrating that the 'busy' design always requires more regulators than the 'idle' design (see Section 2.5.5).

Using Eq 2.5, we obtain exact expressions for X^* under these extreme designs (see Section 2.5.4). In Fig 2.3B, we show $\Delta X^* = X^*_{\text{idle}} - X^*_{\text{busy}}$, the difference in minimal crosstalk X^* between the two extreme designs, for all (p, q) combinations. We find that the 'idle' design yields less crosstalk in a large part of this parameter space. The 'busy' design still involves less crosstalk for parameter combinations centered around the diagonal $p = q$, whereas the 'idle' design always performs best on the anti-diagonal $1 - p = q$. This is due to the fact that on the diagonal, the fraction of activators, p, equals exactly the fraction of genes that should be active q, resulting in full usage of all existing TFs, $t = 1$. On the anti-diagonal $1 - p = q$, the fraction of genes that should be active, q, equals exactly the fraction of repressors $1 - p$. Thus, the default state of all genes is the desired regulatory state requiring no TF usage at all, $t = 0$, which makes the 'idle' design most advantageous.

In the region in which the 'busy' design yields the lowest crosstalk, this comes at the cost of using a larger fraction of existing TF species, as depicted in Fig 2.3C. The 'idle' design, in contrast, requires a much smaller fraction of TF species. Furthermore, the two designs differ not only in the fraction of TFs needed but also in their concentrations. To achieve the lower bound, the 'busy' design always requires a higher total TF concentration, c^* (Fig 2.3D).

The explanation for the alternating crosstalk advantage between the two extreme designs lies in the non-monotonic dependence of crosstalk on TF usage, t (Fig 2.2A). For $t(p, q) < t^*(s)$, crosstalk *increases* and for $t(p, q) > t^*(s)$, it *decreases* with t. Thus, for (p, q) combinations for which $t_{\text{idle}} < t_{\text{busy}} < t^*$, 'idle' design will yield lower crosstalk, whereas if $t^* < t_{\text{idle}} < t_{\text{busy}}$, 'busy' will be more advantageous (see Section 2.5.2 for more details). While 'idle' and 'busy' represent the two extremes, a

continuum of regulatory designs interpolating between these two extremes can be defined. We show, however, that minimal crosstalk is always obtained by one of the two extremes, due to the concavity of $X^*(t)$ (see Section 2.5.2).

We previously found that for some parameter combinations of similarity, s, and fraction of active genes, q, the mathematical expression for X^* (Eq 2.5) has no biological relevance [Friedlander *et al.*, 2016]. Specifically, for similarity between binding sites which is too high $s > \frac{1}{1-t}$, regulation is ineffective and the lower bound on crosstalk X^* is obtained with no regulation at all. Another biologically irrelevant regime occurs for high TF usage $t > t_{\max}$ (see SI of [Friedlander *et al.*, 2016]). Then the concentration needed to obtain minimal crosstalk formally diverges to infinity $c^* \to \infty$. These biologically implausible regimes put an upper bound to the total number of genes that an organism can effectively regulate [Itzkovitz *et al.*, 2006; Friedlander *et al.*, 2016]. The results shown in Fig 2.3 only refer to crosstalk values obtained in the 'regulation regime' where c^* is finite and positive, $0 < c^* < \infty$. Specifically, we find that when similarity, s, increases, parts of the parameter space shown in Fig 2.3A indeed move into the anomalous regimes. In particular, the high TF usage region around the diagonal $p = q$, where the 'busy' design outperforms in crosstalk reduction, vanishes due to this anomaly (see Fig 2.3E where anomalous regions are blackened). For high similarity values $s > 5$, the 'idle' design yields lower crosstalk in the entire biologically relevant parameter space – see Section 2.5.6 and Fig 2.10.

2.2.4 The distribution of crosstalk in a stochastic gene activity model

So far, we considered a deterministic model in which the numbers of active genes and available TF species were fixed, resulting in a single crosstalk value per (p, q) configuration. In reality, these numbers can temporally fluctuate, for example, because of the bursty nature of gene expression [Golding *et al.*, 2005; Wang *et al.*, 2009]. In the deterministic model, we also assumed uniform gene usage, such that all genes are equally likely to be active. In reality, however, some genes are active

more frequently than others.

To account for this, we study crosstalk in a probabilistic gene activity model. We assume independence between activities of different genes, where each gene i, $i = 1...M$, has demand (probability to be active) D_i. We then numerically calculate crosstalk for a set of genes. This approach enables us to incorporate a varying number of active genes and a non-uniform gene demand and compare our results to the deterministic model studied above. To comply with its demand D_i, each gene i is regulated with probability γ_i, where $\gamma_i = D_i$ if regulation is positive and $\gamma_i = 1 - D_i$ if it is negative. We then obtain exact solutions for the distributions of t and X^* (Eq 2.26, Eq 2.27 and Section 2.5.7). In Fig 2.3F, we illustrate the X^* distributions for two values of t, representative of the two extreme designs. We find excellent agreement between this analytical solution and stochastic simulation results. The distribution of X^* is typically narrow, such that for practical purposes, the distribution mean, calculated using the deterministic activation model, serves as an excellent estimator of crosstalk values. For more details on this calculation and for approximation of the distribution width, see Section 2.5.7.

2.2.5 Data-based crosstalk calculation

Similarity and crosstalk, considered in our analytical model, can be estimated from bioinformatic data. As direct thorough measurements of TF binding preferences are available for only a few TFs [Maerkl and Quake, 2007; Fordyce *et al.*, 2010; Afek *et al.*, 2014], we use statistical estimates based on multiple binding sites to which a particular TF binds (PCM) to determine its binding energetics to various sequences. Specifically, we use data of 23 *S. cerevisiae* transcription factors collected from the scerTF database [Gasch *et al.*, 2000; Spivak and Stormo, 2012]. PCMs are $4 \times L$ matrices that provide the total number of counts for each nucleotide at each of the L binding site positions, taken over multiple binding sites of the particular transcription factor. They allow us to compute the mismatch energy penalties for every position and nucleotide in a given binding site sequence and then numerically calculate crosstalk.

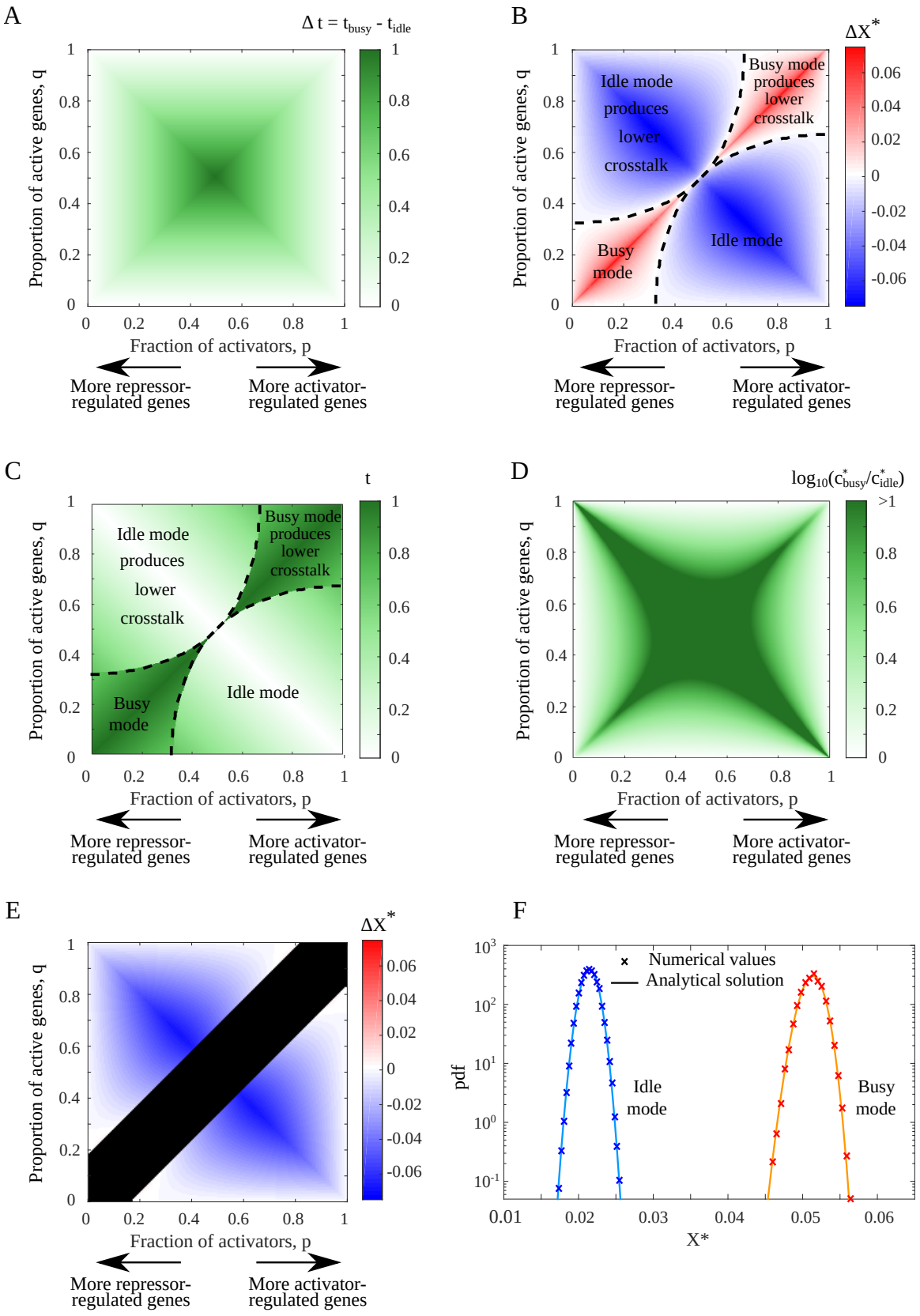

Figure 2.3: (Continued on the following page.)

Figure 2.3: **'Idle' design yields lower crosstalk than the 'busy' in a large part of the parameter regime. (A)** The 'busy' design always requires more TFs compared to the 'idle' design. Here we illustrate Δt, the difference in the fraction of TFs in use between the two designs for different values of p and q (shown in color scale). **(B)** The difference in minimal total crosstalk ($\Delta X = X^*_{\text{idle}} - X^*_{\text{busy}}$) between 'idle' and 'busy' designs, shown in color scale, as a function of p and q. In a large part of the parameter regime (colored blue), lower crosstalk is achieved by the 'idle' design. The 'busy' design is most beneficial on the diagonal $p = q$ (red region), but this requires use of all TFs and comes at the cost of an enormously high TF concentration. The 'idle' design is most beneficial around the anti-diagonal $q = 1 - p$, where regulation can proceed with no TFs at all and crosstalk is close to zero. **(C)** Fraction of TFs in use (shown in color scale) when the design providing minimal crosstalk ('idle' or 'busy' as in (B)) is used, as a function of p and q. Black dashed lines mark the borders between the regions where 'busy' or 'idle' designs provide lower crosstalk. While 'idle' design mostly requires a minority (< 50%) of the TFs, the 'busy' design always necessitates a majority (> 50%) of TFs to be in use. $s = 10^{-2}$ was used in (B)-(C). **(D)** Ratio between TF concentrations providing minimal crosstalk in either design $c^*_{\text{busy}}/c^*_{\text{idle}}$. 'Busy' design always requires higher TF concentrations. **(E)** For higher similarity s between binding sites, parts of the parameter space fall into the anomalous regime where the optimal TF concentration diverges to infinity. We plot here the difference in optimal crosstalk $\Delta X = X^*_{\text{idle}} - X^*_{\text{busy}}$ between designs for $s = 1$. Black areas denote the anomalous regime. Importantly, the region where the 'busy' design was beneficial for low s (see (B)) falls into this anomalous regime. **(F)** Analytical solution of the stochastic model for the distribution of crosstalk values, is in excellent agreement with stochastic simulation results. The distributions obtained are narrow, suggesting that their mean value is representative. Crosstalk values only depend on TF usage, regardless of the exact underlying model. Parameter values: total number of genes $M = 3000$, proportion of activator-regulated genes $p = \frac{1}{3}$, regulation probability $\gamma_i = \gamma = 0.12$ for 'idle' design and $\gamma_i = \gamma = 0.92$ for 'busy' design, with $2 \cdot 10^6$ realizations.

In our theoretical model, we made several simplifying assumptions to allow for an analytical solution. In particular, we assumed uniform properties for all binding sites, assigned equal energetic contributions to all nucleotides in the sequence and assumed that all TFs regulate an equal number of genes (a single gene per TF, in the basic model). The availability of TF binding data allows us to relax these assumptions, and consider variation in binding energies and promiscuity among TFs, as well as the actual unequal energetic contributions of the different positions in each binding site.

For simplicity, we still assume equal concentrations for all available TFs and

calculate a lower bound on crosstalk if concentrations are optimized. In Section 2.5.8 we demonstrate how crosstalk calculation can be implemented for general TF concentration values and show an example using experimentally measured concentrations [Ghaemmaghami *et al.*, 2003]. Due to paucity of data on epistatic effects between distinct binding site positions, we still assume additivity in the energetic contributions of different positions in the sequence. The latter assumption is considered reasonable for up to 3-4 bp substitutions [Maerkl and Quake, 2007].

Similarity values vary between genes even within the same organism. We begin by numerically calculating the similarity s_i between consensus binding sequences of different transcription factors (see Section 2.4). In Fig 2.4A, we show the distribution of similarity values of genes associated with 23 *S. cerevisiae* transcription factors (top). We find a broad distribution of s_i values spanning over 5 orders of magnitude, where its median is around $10^{-4} - 10^{-3}$. This finding is in marked contrast to the full symmetry and equal s_i values for all TFs assumed in our analytical solution.

While we find that s_i values are very variable, the largest contributions to global crosstalk are made by the few most promiscuous TFs (those with high s_i values). In the following, we fit an effective similarity value that would best capture the numerically calculated crosstalk values, had all TFs had uniform s_i values, as in the mathematical model (denoted by red arrow in Fig 2.4A). In this example, we find that $s_{\text{effective}}$ is almost equal to the median s_i value (black arrow there).

Numerical crosstalk calculation: incorporation of a complex TF-gene interaction network. In the analytical model, we assumed that each TF regulates only a single unique gene. Yet, in real gene regulatory networks, the same TF species often regulates multiple genes and some genes are regulated by a combination of different TFs. To account for this, we expand (using SGD) our dataset to include all the 2126 genes regulated by the 23 *S. cerevisiae* TFs for which we have PCMs and considered all possible TF-gene interactions in this set. Notably, there is high variability in the number of genes regulated by each TF. For different values of t (proportion of available TFs), we randomly choose a subset of TFs to be available

and accordingly compute the crosstalk probabilities for all genes, accounting for all possible TF-binding site (BS) combinations. We repeat this procedure for 20 different t values, with 100 independent draws of available TFs for each. In the crosstalk calculation, we assume that all available TFs have equal concentrations. In contrast to the analytical calculation, where we included crosstalk contributions from all TFs, here, only binding states associated with transcription factors that are chosen to be available, are considered. In the analytical model we assumed full symmetry between all TFs and all binding sites. Hence a single similarity value s was sufficient. In contrast, in a real network, we obtain a variety of similarity values (Fig 2.4A). As each TF regulates multiple binding sites, we now calculate similarity between the consensus sequences of the different TFs, and refer to similarity between TFs, rather than similarity between binding sites. In order to compare similarity values of different networks, we fit the numerically calculated crosstalk with the analytical model, where a single $s_{\text{effective}}$ value is used for all TFs. Fig 2.4B shows both the numerically calculated crosstalk and the analytically predicted one (using $s_{\text{effective}}$) for this more complex interaction network (solid and dashed lines, correspondingly). The gray shading represents ±1 standard deviation around the mean value of the numerically calculated crosstalk.

Data incompleteness could affect crosstalk estimates. Global crosstalk accounts for the combined effects of all of the organism's TFs and binding sites. Unfortunately, data of TF binding preferences is incomplete. Moreover, the accuracy of PCMs depends on the number of known binding sites associated with the TF of interest. Due to these technical limitations, we focused on only 23 *S. cerevisiae* TFs for which > 5 binding sites (per TF) are known. However, this small subset of TFs regulates one third (!) of the yeast genes. Motivated by that, we ask how representative is a crosstalk estimation of the entire network based on this small TF subset. In other words, what fraction of the TFs (or genes they regulate) would suffice to reliably estimate the global network crosstalk.

This crosstalk estimation problem is further complicated by the diversity of s_i values we find among TFs. To generally address these questions, we simulate

synthetic gene regulatory networks, each integrating 300 TFs. We simulate the binding preferences of these TFs using the PCM statistics of the 23 yeast TFs. We then sample subnetworks of different sizes from these full networks and numerically calculate crosstalk for each subnetwork (see Section 2.4).

We sample the full networks in two manners: we either randomly choose a subset of TFs ("random subnetworks") or deterministically select the TFs showing the highest similarity with respect to the full network ("ordered subnetworks"). The latter choice is motivated by the prior information that the few yeast TFs for which we have reliable data, are not a random subset, but rather the subset that has the largest number of binding sites. This choice is then a worst-case estimate of global crosstalk. To compare different networks on an equal basis, we estimate the effective similarity $s_{\text{effective}}$ fitted for each subnetwork. Fig 2.4C shows the distributions and medians of $s_{\text{effective}}$ values obtained, as a function of the subnetwork size. Each distribution is based on independent draws of 100 full networks. From each full network, we sample one random and one ordered subnetwork of each size.

We find, that small-size "ordered" subnetworks exhibit higher median $s_{\text{effective}}$ values but narrower distributions than the "random" subnetworks, as expected. Both "ordered" and "random" subnetworks converge to the same $s_{\text{effective}}$ value for the full network (of size 300). The $s_{\text{effective}}$ distribution for the full size represents variation between various full networks of same size, which is significantly smaller than the variation due to limited sampling, observed for the smaller networks. As the "ordered" subnetworks deliberately include the most promiscuous TFs, their $s_{\text{effective}}$ is an over-estimate of the full network measure. In contrast, we find, that $s_{\text{effective}}$ estimated for random subnetworks is an under-estimate of the full network $s_{\text{effective}}$. In our synthetic data, we allowed for binding site length variation among TFs (the PCM dimension). Interestingly, we find positive correlation between the TF's promiscuity s_i and its consensus binding site length. An opposite effect is found for the length of DNA binding sites (see Fig 2.15).

Considering the sufficiency of the sample size, for an "ordered" subnetwork, a sample of ~ 50 (out of 300) TFs provides variation close to the full network measure, whereas for "random" subnetworks, a larger sample size of around ~ 100 TFs (out of

300) is needed. Either way, we conclude that a global crosstalk estimate is possible with only a subset of the network TFs. We compare our calculated s_i values of yeast data (red cross) to the estimated $s_{\text{effective}}$ distributions of this subnetwork size. Interestingly, the yeast estimated crosstalk value falls below the median value for both "random" and "ordered" sampling approaches. This may imply that selection to reduce crosstalk is at work, yielding similarity values which are lower than what one would expect at random [Qian and Kussell, 2016].

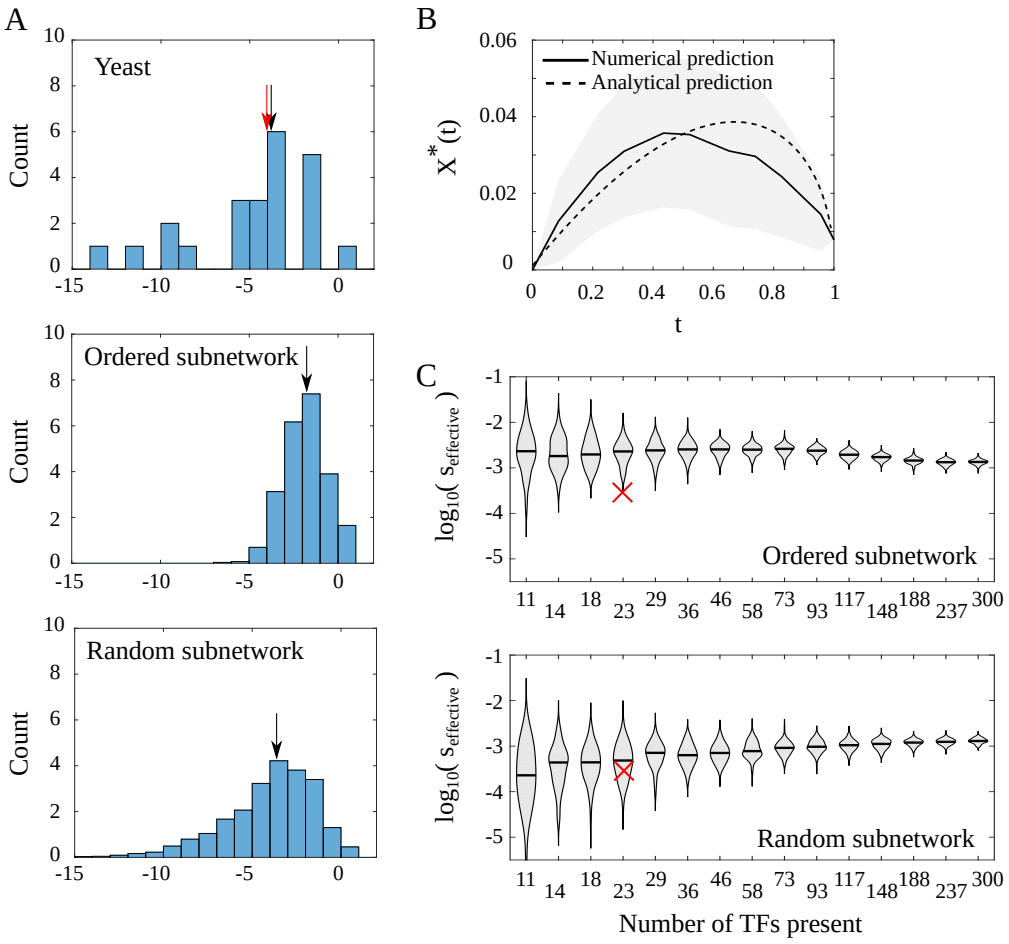

Figure 2.4: (Continued on the following page.)

Figure 2.4: **Data-based crosstalk estimates. (A)** Inter-TF similarity values of *S. cerevisiae* TFs (top), and of synthetic data (middle and bottom) exhibit broad distributions spanning a few orders of magnitude. The distribution median values are marked by black arrows. The red arrow in the yeast data represents $s_{\text{effective}}$ of the yeast data, and nearly overlaps with the distribution median. Synthetic data were created by randomly drawing PCMs representing all TFs of an artificial network. Then, sub-networks of 23 TFs were sampled by either taking the 23 most promiscuous TFs (middle) or randomly choosing them (bottom). The figures show similarity distributions amongst TFs in these artificial networks, averaged over 100 repeated draws. s_i values here are with respect to all TFs in the network, regardless of their (un)availability. **(B)** Numerical prediction of minimal global crosstalk depending on TF availability t for *S. cerevisiae* (solid line) compared to an analytical prediction based on a single $s_{\text{effective}}$ value common to all genes (dashed line). This effective similarity value was chosen to provide the best fit to the numerical curve. The curves represent estimation of crosstalk for the network of all 2126 *S. cerevisiae* downstream genes regulated by the 23 TFs, for which we have PCMs. The numerical curve represents the mean over 10^3 realizations for each t value, where the exact subset of available TFs was randomly drawn. The surrounding gray shadings show ±1 standard deviation around the mean. The discrepancy between numerical and analytical calculations is attributed to the broad distribution of s_i values for the numerical calculation, whereas the analytical calculation assumes a uniform s_i value for all TFs. **(C)** Violin plots of $s_{\text{effective}}$ for different subnetwork sizes for *ordered* and *random* subnetworks. Ordered subnetworks are the subsets of TFs having highest similarity s_i with respect to the whole network. Random subnetworks include a random subset of the full network TFs. For each subnetwork, we numerically calculated crosstalk and fitted the $s_{\text{effective}}$ which would best capture the crosstalk function if all TFs had a uniform s value. The violin plots represent distributions of effective similarity values from 100 different randomly drawn subnetworks, each coming from an independently drawn full network of 300 TFs. The red x represents the $s_{\text{effective}}$ value of the 23 yeast TFs (same value as the red arrow in A). For details on the numerical calculations of similarities and crosstalk, see Section 2.4. All violin plots exhibit broad $s_{\text{effective}}$ distributions which are broadest for the smallest subnetworks, as expected. For "ordered" subnetworks, the median $s_{\text{effective}}$ value is high for the small subnetworks (which were chosen to contain the most promiscuous TFs) and then slightly decreases for bigger subnetworks. For random subnetworks, the trend is opposite.

2.3 Discussion

We studied the susceptibility of different gene regulatory networks to transcriptional crosstalk. We found a lower bound on crosstalk $X^* = X^*(t, s)$, which is fully determined by two macroscopic "thermodynamic-like" variables, regardless of other

microscopic details of the network. These are the fraction of available TF species, t, and the average similarity between distinct binding site sequences, s. This emergent simplification enabled us to analyze crosstalk for classes of gene regulatory networks, regardless of other network details. We showed that different network designs may vary in t, the TF usage they require, and hence differ in the crosstalk levels they incur, even if they have the same gene activity pattern. We analyzed two extremes: a 'busy' design, which maximizes the use of regulators and is equivalent to the previously proposed Savageau demand rule [Savageau, 1974] and the opposite 'idle' design, that minimizes the use of regulators. Interestingly, crosstalk is minimized by either of these extremes, and not by any hybrid design. We found that, in a large part of the parameter regime, crosstalk increased with t, and consequently minimized by the 'idle' design. In the remaining part, crosstalk was minimized by the 'busy' design, but came at a cost of a much higher TF concentration requirement. Our basic analysis refers to a simple network architecture. We exemplify in Section 2.5.10 how the crosstalk expressions Eqs 2.2-2.5 can be generalized to describe more complex regulatory architectures. We also studied a stochastic gene activation variant of the model, where the number of active genes can fluctuate. We found that it is well-approximated by the deterministic activation model, because the distributions of TF availability and minimal crosstalk are typically very narrow and centered around their mean value.

Where are real organisms located in the (t, s) parameter space? Reports of the number of co-expressed genes greatly vary between organisms and depend on growth conditions. For example: $\sim$ 10,000 different genes were reported to be co-expressed in a mouse cell (< 50% of total) [Carter *et al.*, 2005; Islam *et al.*, 2014], 10,000-12,000 (< 50%) genes were estimated to be co-expressed in human HeLa cells [Nagaraj *et al.*, 2011], 3300-3500 out of 4290 genes (76%-82%) were co-expressed in *E. coli* during exponential growth [Tao *et al.*, 1999; Wei *et al.*, 2001] and 75%-80% of the genes were co-expressed in *S. cerevisiae* [Ghaemmaghami *et al.*, 2003; Lewin, 2007].

Values of similarity between distinct TF binding sites, vary not only between organisms, but also between modules and distinct genes within the same organism

(see Fig 2.4). We estimated s_i and the resultant minimal crosstalk values for 23 *S. cerevisiae* TFs using PCM data. We found an extremely broad distribution of single-TF s_i values spanning > 5 orders of magnitude, with a median between $10^{-4} - 10^{-3}$. Global crosstalk, however, is determined by the few high-similarity TFs. To bridge the gap between the high diversity of s_i in real networks and our uniform s analytical solution, we fitted a single $s_{\text{effective}}$ value which would best capture the numerically calculated network crosstalk. For the yeast data, we found that this $s_{\text{effective}}$ is very close to the distribution median. Using our estimates for s and t, we estimated minimal crosstalk X^* for this subnetwork of *S. cerevisiae* to be in the range 0.03-0.04 (see Fig 2.4B), if 30%-80% of the TFs are present. Our analysis showed that, for relatively low s values, as we found for yeast, there was a regime in the parameter space in which 'busy' yields the lowest crosstalk. The choice of network design that minimizes crosstalk ('busy'/'idle') depends on the proportion of co-activated genes and on the proportion of activators. For organisms with high s values, the regime in which 'busy' is beneficial is actually anomalous, and hence biologically irrelevant. Such higher s is expected for organisms with shorter binding sites.

Binding site data is often incomplete. To assess the validity of whole-network crosstalk estimation based on a small subset of TFs, we constructed synthetic gene regulatory networks, sampled some subnetworks and then compared the s estimation of full and partial networks. In the *S. cerevisiae* case, we found that a full network crosstalk estimate is possible with binding information of only 16%-33% of the TFs.

Here, we used a symmetric and admittedly simplified gene regulatory network model. Our analysis determined a lower bound for crosstalk, assuming that TF concentrations are accurately tuned. In reality, TFs are not necessarily expressed and degraded at a precise time [Price *et al.*, 2013] and crosstalk is thus expected to be higher. In Section 2.5.8 we demonstrate crosstalk calculation with general TF concentrations, obtained in experiments. Relaxation of other simplifying assumptions made in our analytical model opens new research avenues for future work. Most importantly, we assumed uniform similarity values of all TFs and all BS, whereas *S. cerevisiae* data analysis showed diversity in TF properties. In principle, a distribution of s values

can be incorporated into the model, but would significantly complicate averaging over different sets of active genes (but see a simple example in Section 2.5.1. Other simplifications include the averaging over gene sets of same-size as representatives of different environmental conditions, whereas, in reality, the number of expressed genes could vary between environments (e.g., growth media [Tao *et al.*, 1999]). We averaged over all possible choices of active genes, although only some of these activity combinations occur naturally. We also assumed that every gene has a regulator, and vice versa, although this is not always the case. Hershberg and co-workers found an imbalance between genes and regulators, where orphan repressors with no genes and orphan genes with no activators, transiently exist, and could also contribute to crosstalk [Hershberg and Margalit, 2006]. Relaxation of these assumptions would require a more comprehensive characterization of gene regulatory networks and co-expression patterns than is known to date.

Our study addressed a typically overlooked cost of protein production: that of regulatory interference caused by excess regulatory proteins in the dense cellular medium. This cost is distinct from the energetic burden of unnecessary protein production, which was found to delay growth [Koch, 1983; Kurland and Dong, 1996; Dekel and Alon, 2005; Shachrai *et al.*, 2010].

It was previously shown that transcriptional error for a single gene is minimized when its binding site is occupied [Shinar *et al.*, 2006] - a regulatory strategy equivalent to the Savageau demand rule. However, single-gene models neglected the increase in erroneous interactions that can occur following network augmentation beyond the single gene. The regulatory cost increases super-linearly with the number of molecular species and regulatory interactions and can therefore only be determined when the network is considered as whole. This would result in a different mathematical solution to minimize global crosstalk, compared to the single-gene case. For comparison between single-gene and global crosstalk models see Section 2.5.9.

Selection to reduce global regulatory crosstalk [Hahn *et al.*, 2003; Qian and Kussell, 2016], was reported in previous bioinformatic studies. Our finding that effective similarity obtained for the *S. cerevisiae* gene regulatory network is lower than the median effective similarity obtained in random networks with similar

parameters, corroborated these reports (Fig 2.4C). Yet, crosstalk is not fully eliminated by selection. Despite the functional interference it causes in the short run, crosstalk is thought to promote evolvability in both gene regulatory and signalling networks in the long run [Shultzaberger *et al.*, 2012; Payne and Wagner, 2014; Aakre *et al.*, 2015; Friedlander *et al.*, 2017; Rowland *et al.*, 2017]. However, the interplay between these two opposing effects of crosstalk, is still poorly understood.

Crosstalk reduction is one of several functional considerations shaping the evolution of gene regulatory networks. Other considerations include the network dynamical properties [Alon, 2007] and protein production requirements [Kumar Prajapat *et al.*, 2016]. Above all, evolution is a random process and certain network designs become fixed and continue propagating [Wagner, 2008; Fontana and Buss, 1994; Friedlander *et al.*, 2013; Martin *et al.*, 2016]. For example, new transcription factors often evolve by duplication of an existing TF followed by sub- or neo-functionalization, thereby preserving the form of regulation of the ancestral TF [Nguyen and Saier, 1995]. Taken together, a generalized model for network evolution, which would incorporate the effects of crosstalk on different time scales, alongside traditional selection on the network to achieve a certain input-output goal, remains to be formulated.

2.4 Methods

Distribution of t is approximated by a Gaussian distribution. Given that the cognate TF of gene i is present with a probability γ_i ($i \in (1, M)$, where M is the total number of genes), the distribution of available transcription factor species in the system follows Poisson-binomial distribution. This is the probability distribution of a sum of independent Bernoulli trials with probabilities γ_i, that are not necessarily identically distributed. Its mean and variance are:

$$\langle t \rangle = \frac{1}{M} \sum_{i=1}^{M} \gamma_i = \langle \gamma_i \rangle, \tag{2.8}$$

$$\mathrm{var}(t) = \frac{1}{M} \sum_{i=1}^{M} \gamma_i (1 - \gamma_i) = \langle \gamma_i (1 - \gamma_i) \rangle. \tag{2.9}$$

As this distribution is difficult to compute for large values of M, we follow the central limit theorem and approximate it by a Gaussian distribution with the same mean and variance.

Exact solution of the probability distribution of X^*. For a function $X^*(t)$, where t is a random variable with probability distribution $f_t(t)$, the probability distribution of X^*, $f_{X^*}(X^*)$ is:

$$f_{X^*}(X^*) = \sum_i f_t(g_i^{-1}(X^*)) \left| \frac{\mathrm{d}g_i^{-1}(X^*)}{\mathrm{d}X^*} \right|, \tag{2.10}$$

where $g_i^{-1}(X^*) = t_i$ represents the inverse function of the $i-$th branch. In our case it has two branches:

$$f_{X^*}(X^*) = f_t(g_1^{-1}(X^*)) \left| \frac{\mathrm{d}g_1^{-1}(X^*)}{\mathrm{d}X^*} \right| + f_t(g_2^{-1}(X^*)) \left| \frac{\mathrm{d}g_2^{-1}(X^*)}{\mathrm{d}X^*} \right|. \tag{2.11}$$

The solutions for $g_i^{-1}(X^*)$ and their derivatives exist for crosstalk $X^*(t)$ and can be analytically computed. Therefore, there is a known analytical solution for the distribution of minimal crosstalk $f_{X^*}(X^*)$.

For regime I, the lower limit on crosstalk is $X^*(t) = t$. Its inverse is $g^{-1}(X^*) = t(X^*) = X^*$, while the derivative $\mathrm{d}g^{-1}(X^*)/\mathrm{d}X^* = 1$. Similarly, in regime II, the lower limit

on crosstalk equals $X^*(t) = 1 - t/(1 + \alpha t)$, the inverse function $g^{-1}(X^*) = t(X^*) = (1 - X^*)/(1 - \alpha + \alpha X^*)$, and its derivative $\mathrm{d}g^{-1}(X^*)/\mathrm{d}X^* = -(1 - \alpha - \alpha X^*)^{-2}$. The analytical solution for regime III was computed using Mathematica and the solution can be found in S3 Appendix.
Using these values, one can compute $f_{X^*}(X^*)$ for X^* in all three regimes.

Stochastic semi-analytical solution of crosstalk for a random number of present TFs. For each gene i, we randomly draw, with probability γ_i, whether its cognate TF is available. We then obtain the proportion t of available TFs. As this process is stochastic, the proportion t differs between different realizations. Next, we compute the lower limit on crosstalk $X^*(t)$ for this t value using the analytical solution in the relevant regime (I, II or III). Using multiple realizations (=10^6) of t, we numerically obtain the distribution of crosstalk values for values of $t \in (0, 1)$.

Obtaining the energy matrices from position count matrices (PCMs). Position count matrices (PCMs) document the summary statistics of TF binding site sequences. Each element c_{ij} designates the number of known TF binding site sequences with nucleotide i in position j. We obtained the PCMs from the scerTF database for *S. cerevisiae*. Given these, we calculated the energy matrices which are needed to compute the similarity measure, in the following way: for a position j and nucleotide $i \in \{A, C, G, T\}$, we computed the energy mismatch value as $\epsilon_{ij} = \log(\frac{c_{mj}}{c_{ij}})$, where $c_{mj} = \max_i c_{ij}$ is the maximal count at position j. To avoid divergence of the energy ϵ_{ij} in case of zero counts, $c_{ij} = 0$, we added a constant pseudocount $\delta = 0.1$ to all matrix entries.

Some technicalities and concerns regarding PCM usage. When computing the energy matrices using PCMs, certain issues arise that could strongly bias the results if not properly addressed:

- *Inequality of total counts between positions* in PCM data. The sum of counts over all 4 nucleotides in a given PCM should be equal for all positions, but occasionally, positions with different total counts are found. As they bias our

occurrence statistics (and hence our energy calculation), we used only PCMs in which the total count was equal throughout.

- *Zero counts* in the PCMs. Many PCMs include zero counts for certain nucleotides at specific positions, rendering that element of the energy matrix undefined. Here, we applied a commonly used practice of adding a pseudocount δ to all PCM entries. Following a previous work [Friedlander *et al.*, 2016], where various δ values were compared to an information method (where pseudocount is not needed), we set $\delta = 0.1$.

- *Count number sufficiency*. To achieve a reliable estimation of energies in the energy matrix, we only used PCMs with at least $p_{\text{counts}} = 5$ counts per position.

In total, we found 196 TF PCMs, but due to the above concerns, we considered only 23 of them in our calculations.

Numerical computation of similarity measure using PCMs. To compute the similarity measure between binding site k and a transcription factor l, we first substituted the sequence of BS k by the *consensus* sequence of its cognate TF k. The consensus sequence is obtained by taking the most common nucleotide in each position j. As the given binding site and TF consensus sequence are not necessarily of the same length, we distinguished between the following cases:

- If the TF consensus sequence l was shorter than the binding site sequence k, we computed the energies for all possible overlaps of the shorter sequence with respect to the longer one. We took the minimal value to be the binding energy.

- If the TF consensus sequence l was longer than the binding site sequence k, the TF energy matrix was again slid along the binding site and energies were calculated again for every relative positioning of the two sequences. The only difference from the previous case was that energetic contributions from positions where the TF binds outside the binding site, were taken into account by averaging energies over all four nucleotides. The total binding

energy $E = E_1 + E_2$ is the sum of contributions from nucleotides inside (E_1) and outside (E_2) the binding site. The energy contribution of positions j outside the BS equals $E_2 = \sum_j E_{2j}$, with $E_{2j} = \sum_{i=1}^{4} \epsilon_{ij}/4$ being the average binding energy at position j. Here too, we computed the binding energy for all possible overlaps between the BS and TF and took the lowest value as the binding energy E^{kl}.

This provides the matrix of binding energies E^{kl} between every binding site k and every TF l. Importantly, this binding energy is asymmetric, namely $E^{kl} \neq E^{lk}$. The similarity measure between binding site k and all other binding sites was computed as the average Boltzmann weight, taken over all non-cognate TF binding to binding site k:

$$S_k = \frac{1}{T} \sum_{l=1,\ l\neq k}^{M} C_l e^{-E^{kl}}, \tag{2.12}$$

with C_l being the concentration of TF species l, and T the number of present TF species.

Numerical computation of crosstalk given PCMs. For the numerical computation of crosstalk, we used the matrix of binding energies E^{kl} between binding site k and TF l, using the following algorithm:

1. randomly choose a subset of genes that should be regulated by their cognate TF. At each realization, a different subset is chosen. All subsets form a proportion t of the genes.

2. For gene k, obtain the similarity measure $S_k = \frac{1}{T} \sum_{l\neq k} C_l e^{-E^{kl}}$. Set the concentration of the absent TFs to zero, and set equal concentrations ($C_l = C$) to all present TFs, as in the analytical calculation.

3. Compute the probabilities that a crosstalk state occurs at any given gene, using the thermodynamic model. Other parameters include the energy difference between unbound and cognate state E_a which does not affect the final crosstalk result, and the concentration of the transcription factors, C.

4. Obtain the total crosstalk X by summing over the contributions of all individual genes.

5. Average over a large number of realizations (we used several hundred realizations for which the average crosstalk had already converged).

6. Repeat this procedure (each with multiple realizations) using a different concentration value C each time. Then, pick the one that yields the lowest crosstalk value to be $X^*(t)$.

Numerical computation of crosstalk where a gene could be regulated by multiple TFs. In an actual gene regulatory network, many TFs regulate multiple genes and many genes are regulated by multiple TFs rather than the one-to-one TF-gene association we considered so far. Specifically, in our data, around 96% of the TFs regulate more than one gene. To account for that, we obtained the list of genes that are regulated by the given *S. cerevisiae* transcription factors (from SGD). Numerical crosstalk calculation for this network closely followed the previous procedure. The only difference was the computation of the similarity measure of genes regulated by multiple cognate TFs. Such genes have multiple binding site sequences (one for each cognate TF) and consequently, multiple binding energies and similarity measures. We then calculated a unified similarity measure per gene as follows:

1. For a given gene k, find all the TFs that regulate it.

2. Obtain the consensus sequences of these TFs.

3. Assume each such consensus sequence represents a potential binding site sequence of gene k (same as in the case of only one TF regulating each gene).

4. Compute the similarity measure S_{k^i} between each potential binding site sequence i of gene k and all other TFs; this is done in the same way as for one TF regulating one gene using Eq. 2.12.

5. Use the mean of the computed S_{k^i} similarity measures taken over the various binding sites of gene k as the unified similarity of that gene.

Simulating synthetic data. To simulate synthetic data of TF binding preferences, we constructed artificial PCMs, using the data of the 23 yeast energy matrices, as follows. We first created the nucleotide abundance distribution of the yeast TFs consensus sequence and then drew random realizations from this distribution to obtain a consensus sequence for each synthetic TF. This distribution was non-uniform and biased towards excess of A and T nucleotides. We allowed for a variety of consensus sequence lengths, using the same length distribution as in the yeast data. Similarly, we created the distribution of the non-consensus energy values of the 23 TFs energy matrices and drew random realizations from this distribution to construct the energy matrices for the synthetic TFs.

Computing the subnetworks of synthetic data and their crosstalk. To construct a full network, we fabricated data for 300 TFs, as described above. We then computed the network's matrix of binding energies $E^{kl}_{\text{full network}}$ of the l-th TF to the k-th binding site, where the each binding site sequence was taken as the consensus sequence of its cognate TF, as in the yeast data. We next formed subnetworks of this full network, by choosing a subset of TFs and taking the corresponding subset of binding energy entries, to obtain $E^{kl}_{\text{subnetwork}}$. We used either randomly chosen subsets of TFs ("random networks") or deterministically picked the subset of TFs having the highest similarity measure $S_i^{\text{full network}}$ with respect to the full network. We then numerically computed minimal crosstalk X^* for each subnetwork, following the same procedure as for the yeast data. We repeated this procedure for 100 randomly drawn full networks.

Comparison of the numerical results to the analytical expression. We fit the analytical expression for $X^*(t)$ to the numerically calculated crosstalk. The main difference between the two approaches is that the analytical expression assumes uniform S_k values for all TFs, whereas the numerical approach allows for diverse S_k values. We assumed a single representative $s_{\text{effective}}$ value that would best fit the numerical result. For this, we minimized the sum of squared differences over various values of t to find the best $s_{\text{effective}}$. Distributions of $s_{\text{effective}}$ values were based on

100 randomly drawn full networks from which subnetworks were sampled. For each subnetwork size, we sampled each of the full networks just once, to avoid correlations between the random subnetworks.

2.5 Supporting Information

2.5.1 Model description

We consider a cell that has a total of M transciptionally regulated genes, which can be either active or inactive. We assume that each gene is regulated by a single unique transcription factor (TF) - its cognate TF. Each gene has a short DNA binding site to which a TF can bind to affect its regulatory state. A fraction $0 \leq p \leq 1$ of the genes is regulated by activators and the remaining $(1-p)M$ genes are regulated by repressors. When no activator is bound, activator-regulated genes are inactive (or active at a low basal level) and only become active once an activator TF binds their binding site. In contrast, repressor-regulated genes are by default active, unless a repressor TF binds their binding site and inhibits their activity. We assume that different environmental conditions require the activity of different subsets of proportion $0 \leq q \leq 1$ of these genes, while the remaining fraction $1 - q$ should be inactive. As both activity and inactivity of genes can be attained by means of either activator or repressor regulation, our model distinguishes between four sets of genes: (i) $a \leq q, p$ activator-regulated genes which are active, (ii) $q - a$ repressor-regulated genes which are active, (iii) $p - a$ activator-regulated genes which are inactive, and (iv) $(1 - p) - q + a$ repressor-regulated genes which are inactive.

The special cases in which all the genes have the same form of regulation, either repression or activation (namely $p = 0$ or $p = 1$), were studied in a previous work [Friedlander *et al.*, 2016].

We assume the system is generally at steady state, such that the required gene expression pattern does not change in time and all molecular concentrations are fixed. We then consider the average crosstalk over different gene sets of the same size. This represents a series of different gene expression patterns required in different external conditions. We assume that the system only seldom shifts from one steady state to another, such that the transient time needed for gene regulation to equilibrate following each transition is negligible. We do not consider any form of feedback exerted by the products of these genes. Rather, we assume an idealized situation in

which all necessary regulators are present exactly at the time and quantity needed. Any deviation from these conditions is expected to increase crosstalk levels. Hence, our analysis refers to a lower bound of crosstalk levels.

Each gene is associated with a short regulatory DNA sequence (binding-site), to which its specialized cognate TF preferentially binds to affect its regulatory state, either positively or negatively. Although the regulatory sequences of different genes differ from each other and we assume that each TF is specific to the regulation of only one unique gene, TFs are known to have limited specificity to their DNA targets and can occasionally bind slightly different sequences, albeit with lower probability [Maerkl and Quake, 2007]. We define cases when a TF binds a non-cognate binding site or when a binding site that should have been bound remains unoccupied, as 'crosstalk', potentially leading to an undesired regulatory outcome. To quantitate the probability of these events, we use the thermodynamic model of gene regulation [Shea and Ackers, 1985; Von Hippel and Berg, 1986; Gerland *et al.*, 2002; Bintu *et al.*, 2005a], which asserts that the occupancy of regulatory binding sites by TFs determines the expression level of the genes associated with these binding sites. The probability of this occupancy depends on the copy number of active TF molecules available to bind and on the binding energy between the binding site and TF. This binding energy is determined by the number of mismatches between the particular binding site sequence and the consensus sequence of that TF. We assume full symmetry in the biophysical properties of the binding sites associated with different genes: all have the same sequence length and equal binding energy to their cognate TFs, and all genes have the same dynamic range of expression. Each binding site can occupy different energy levels, depending on its binding state. It is in its lowest energy level $E = 0$ if it is bound by its cognate TF. Higher energy levels are obtained if it is bound by a non-cognate TF, such that there is a mismatch between the consensus sequence of the TF and the DNA sequence of the binding site. We assume additive and equal energetic contributions of size ϵ to all nucleotides in the binding site, such that the binding energy of a TF to a sequence which differs in d positions from the consensus sequence equals $\epsilon \cdot d$. Under constant external conditions, only a subset of TFs (activators and repressors) are available to bind.

These TFs are needed to maintain the activity of the q proportion that should be active and, simultaneously, the inactivity of the remaining $1 - q$. Unavailability means either that the TF molecules are physically absent from the cell at that time, because they were degraded, or that they are present in an inactive state and only become active in response to an external signal (e.g., via phosphorylation or other modifications). The probability that a particular gene i is in either of the crosstalk states depends on the copy number of competing non-cognate TFs, C_j, $j \neq i$ and on the number of mismatches, d_{ij} between each competing TF j and the regulatory binding site of gene i. We distinguish crosstalk states of genes whose desired state of activity requires that their binding site remains unoccupied and those for which it should be occupied by a cognate regulator. The binding site of an activator-regulated gene that should remain inactive as well as that of a repressor-regulated gene that should be active, must all remain unoccupied. For these genes, the cognate TF is not available to bind and any binding event by another (non-cognate) regulator is considered crosstalk. For genes whose binding sites should be occupied by their cognate regulator (an activator-regulated gene that should be active and a repressor-regulated gene that should be inactive), crosstalk states occur either if the binding site remains unbound or if it is occupied by a non-cognate regulator, in which case, the regulatory state is not guaranteed. Using equilibrium statistical mechanics, the crosstalk probabilities for a single gene i are [Von Hippel and Berg, 1986; Gerland *et al.*, 2002; Friedlander *et al.*, 2016]:

$$x_{\text{bound}} = \frac{e^{-E_a} + \sum_{j \neq i} C_j e^{-\epsilon d_{ij}}}{C_i + e^{-E_a} + \sum_{j \neq i} C_j e^{-\epsilon d_{ij}}} \tag{2.13a}$$

$$x_{\text{unbound}} = \frac{\sum_{j \neq i} C_j e^{-\epsilon d_{ij}}}{e^{-E_a} + \sum_{j \neq i} C_j e^{-\epsilon d_{ij}}} \tag{2.13b}$$

x_{bound} refers to crosstalk when the binding site should be bound (by either activator or repressor) but is either unbound or bound by a non-cognate molecule. x_{unbound} refers to crosstalk when the binding site should remain unbound and no cognate binder is available, but is still bound by some non-cognate molecule. E_a is the energy difference between cognate bound and unbound states. The expression $\sum_{j \neq i} C_j e^{-\epsilon d_{ij}}$ then captures the sum of all interactions with foreign regulators that

binding site i might receive.

To account for the sum of all non-cognate interactions received by a particular binding site in a model of multiple TF species, we define an average measure of similarity between binding site i and all other binding sites $j \neq i$ [Friedlander *et al.*, 2016]:

$$S_i \equiv \langle e^{-\epsilon d_{ij}} \rangle_{P(d)} = \frac{1}{T} \sum_{j \neq i} e^{-\epsilon d_{ij}} = \frac{1}{C} \sum_{j \neq i} C_j e^{-\epsilon d_{ij}}. \tag{2.14}$$

S_i is defined as the average of the Boltzmann factors taken over the distribution of mismatch values $P(d)$ between binding sites i and j, $\forall j$. In the last equality in (2.14), we assume that all available TFs are found in equal concentrations $C_j = C/T, \forall j$, where C is the total TF concentration and T is the total number of available TF species. We assume full symmetry between binding sites i, such that each binding site i has the same distribution of mismatches d_{ij} with respect to all the other genes, hence $S_i = S \ \forall i$. The value of S can either be estimated using binding site data (see example in Fig 2.4) or analytically calculated under different assumptions on the pairwise mismatch distribution $P(d)$. Following our symmetry assumptions, the crosstalk probabilities in (2.13) are independent of the gene identity i, such that we only need to distinguish between the four different regulatory states. In the following, we use rescaled similarity defined as $s = S \cdot M$, which represents a sum of all non-cognate interactions at a binding site.

For which TF usage t^* is crosstalk maximized?

For a fixed s $X^*(t, s)$ has a maximum at a certain t value, which we denote t^* (marked with a black circle on Fig 2.2A). We find that $t^* = t^*(s)$ and its value monotonically increases with s. For low similarity values ($s \to 0$), it asymptotically approaches the value of $2/3$, with the limit

$$\lim_{s \to 0} t^* = 2/3. \tag{2.15}$$

For $s > 0$, $t^* > 2/3$ and approaches $t \to 1$ for high s. See Fig 2.5 for illustration. See S1 Appendix for more formal formulation.

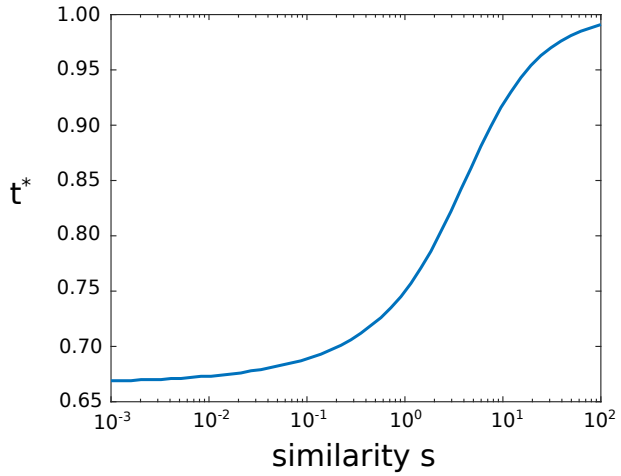

Figure 2.5: **Fraction of TFs used which maximizes crosstalk X^*, increases with similarity s.** t^* - the fraction of TFs used which maximizes crosstalk X^*, increases with similarity s. For $s \to 0$, it asymptotically approaches the value of $2/3$. For large s, it approaches 1, such that $X^*(t)$ is an increasing function of t for all but very high t values.

Optimal TF concentrations in the two strategies

The optimal concentration which minimizes crosstalk is $c^* = 0$ in regime I, $c^* = \infty$ in regime II, and

$$c^* = \frac{C^*}{M} = \frac{te^{-E_a}\left(s\left(st - t(st+2)\right) - \sqrt{s(1-t)}\right)}{s\left(-(st+1)^2 + st^2(st+3) + t\right)} \tag{2.16}$$

in regime III. The concentration in each strategy is obtained by choosing the corresponding value of t, i.e., $t_{\text{busy}} = (1-p) + 2\min(p,q) - q$ and $t_{\text{idle}} = (1-p) - 2\min(1-p,q) + q$.

Relaxation of basic model assumptions

Unequal TF concentrations So far, our model assumed equal concentrations for all present TFs. What happens if we introduce different concentrations for activators and repressors? We expanded the model to allow for two concentrations, one for activators and one for repressors, while keeping the total concentrations constant. This means:

$$A \cdot C_1 + (T - A) \cdot C_1 = C, \tag{2.17}$$

where C_1 and C_2 are the per species concentration of activators/repressors, A and $(T - A)$ the number of activator/repressor regulated genes, and C the total concentration of TFs as used in the main calculation of our model. We numerically tested which combination of concentrations C_1 and C_2 leads to lowest X^* value. Surprisingly, we found that minimal crosstalk is achieved by equal concentrations for all transcription factors, activators and repressors, i.e., $C_1 = C_2$. In other words, adding an additional degree of freedom of second concentration can only increase the crosstalk values.

The intuitive explanation is that it only matters if a gene is regulated or not, but not which type of regulation it employs. This is in similarity to our previous result, where minimal crosstalk only depends on the number of available TFs, such that regulation by activators crosstalk is a mirror image of regulation by repressors alone [Friedlander *et al.*, 2016].

Non-uniform similarity values In the basic model, we assumed uniform similarity values $s_i = s$ for all genes, which allowed us to obtain analytical solutions for crosstalk. As data show (see Fig 2.4), similarity values vary between genes even within the same organism. Here, we relax this simplifying assumption to test its significance. We analyze a special case with two subsets of genes, each with a different similarity value. The two subsets are of relative size r_1 and r_2 ($r_1 + r_2 = 1$), and similarity values s_1 and s_2, correspondingly. In each subset, there is a weighted proportion of regulated genes, $t_i = r_i t$ for $i \in \{1, 2\}$. We use fixed values for $s_{1,2}$ and then calculate $s = r_1 s_1 + r_2 s_2$ for each (r_1, r_2) combination. As before, the total crosstalk X^* is computed by summing crosstalk contributions of all individual genes and then numerically minimizing X^* with respect to the TF concentration. We still allow only equal concentrations for all available TFs. In Fig 2.6, we plot X^* vs. the proportion r_1 for different fractions of available TFs, t. We compare X^* values obtained for uniform and non-uniform s. We find that non-uniform s provides lower crosstalk than uniform s. This is obtained, however, at the cost of higher TF concentration C^* needed for the non-uniform similarity.

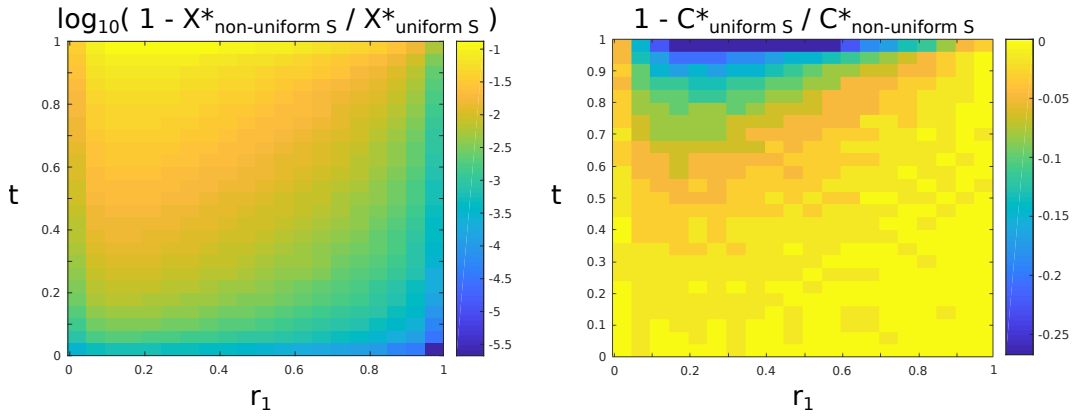

Figure 2.6: **Non-uniform similarity yields lower crosstalk than uniform similarity.** We plot relative change of crosstalk values (left) and concentration of TFs (right) between models of non-uniform and uniform similarities as a function of subset size r_1 and proportion of regulated TFs, t. The values of X^* for non-uniform s are up to 10% lower compared to uniform s. However, this strongly correlates with the increase in concentration. Values used: $s_1 = 5 \cdot 10^{-3}$, $s_2 = 5 \cdot 10^{-4}$.

2.5.2 Achievement of minimal crosstalk

Minimal crosstalk is always obtained by one of the two extreme regulation strategies

The 'busy' and 'idle' modes are the two extreme regulation strategies. Intermediate strategies, where some genes follow the first strategy and others follow the second, are possible. However, minimal crosstalk is always obtained by one of the two extremes.

We denote the proportion of TF species following 'idle' and 'busy' modes by t_1 and t_2, respectively (see Fig 2.7). A combination of the two modes would lead to a linear combination of the fraction of TF species, $t_{\text{mixed}} = \alpha t_1 + (1 - \alpha) t_2$, with $\alpha \in [0, 1]$. Since $X^*(t)$ is a concave function of t with a single maximum, minimal X^* will always be obtained at the edges of the t domain, $\alpha = 1$ or $\alpha = 0$. Thus, any mixed strategy would always bring about higher crosstalk than the extreme ones $X^*(t_{\text{mixed}}) \geq \min\left(X^*(t_1), X^*(t_2)\right)$ (see Fig 2.7).

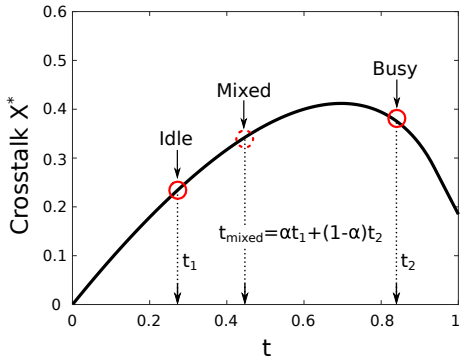

Figure 2.7: **Due to concavity of $X^*(t)$, minimal crosstalk is always obtained by one of the two extreme modes.** Crosstalk of the mixed strategy (dashed red circle) equals $X^*(t_{\text{mixed}})$, where t_{mixed} is the linear combination $t_{1,2}$ - the proportions of TF species involved in the extreme strategies (red solid circles).

Relation between t and the choice of regulatory strategy yielding lowest crosstalk

The non-monotonic dependence of crosstalk on TF usage, t, can explain the non-trivial transition between the 'busy' and 'idle' modes in the (p, q) phase space, as shown on Fig 2.3B. There we illustrate for each (p, q) which of the two modes yields lower crosstalk. Recall that $X^*(t)$ has a maximum at $t = t^*(s)$, such that it is an increasing function of t for $t < t^*$ and a decreasing function of t for $t > t^*$. The relationship of t_{idle} and t_{busy} in regards to t^* determines which strategy is more advantageous. As $t_{\text{idle}} < t_{\text{busy}}\ \forall t$ (see Eq 2.24 below), it follows that if X^* is increasing with t, idle mode is more advantageous (lower $t \rightarrow$ lower $X^*(t)$). Conversely, if X^* is decreasing function of t, busy mode leads to lower crosstalk X^* (higher $t \rightarrow$ lower $X^*(t)$). To address which mode is more advantageous, we examine $t_{\text{idle}}(p, q)$ and $t_{\text{busy}}(p, q)$ at each (p, q) value. We distinguish three cases:

1. For $t_{\text{idle}}, t_{\text{busy}} < t^* \Rightarrow$ idle mode is the most advantageous strategy.
2. For $t^* < t_{\text{idle}}, t_{\text{busy}} \Rightarrow$ busy mode is the most advantageous strategy.
3. For $t_{\text{idle}} < t^* < t_{\text{busy}} \Rightarrow$ which strategy is optimal depends on exact values of (p, q).

We summarize these results in Fig 2.8 where we show where the 3 cases lie in the phase space. The first case, where idle mode is more advantageous, occurs in

the top left and bottom right corner of the phase space (white area). Conversely, $t^* < t_{\text{idle}}, t_{\text{busy}}$ holds in the bottom left and top right corner where busy mode leads to lower crosstalk (black area). The rest (gray area) belongs to the third case where it cannot be easily determined which mode is more beneficial. The boundary between idle and busy mode (red dashed line) lies entirely in the last case and can be obtained analytically by solving the equation $X^*(t_{\text{idle}}, s) = X^*(t_{\text{busy}}, s)$ for (p, q). This result also intuitively explains the expansion of the region where 'busy' is advantageous when s becomes smaller. Since $t^*(s)$ is a decreasing function of s, for smaller s there is a larger (p, q) region where both $t^* < t_{\text{idle}}, t_{\text{busy}}$.

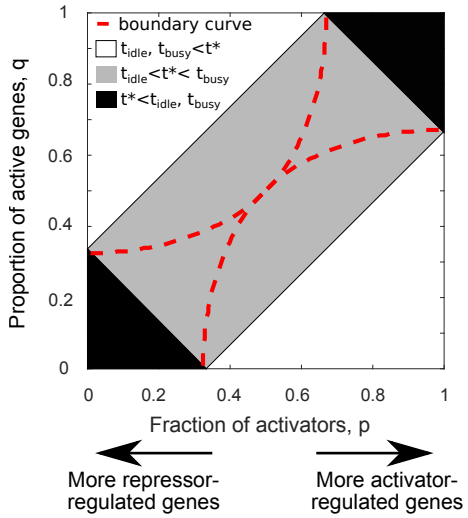

Figure 2.8: **Relation between t and the choice of regulatory strategy yielding lowest crosstalk.** The three cases for all combinations between $t_{\text{idle}}(p, q)$, $t_{\text{busy}}(p, q)$, and $t^*(s)$ in the (p, q) space. Each case is shown in different color. For $t_{\text{idle}}, t_{\text{busy}} < t^*$ (white area), idle mode leads to lower crosstalk; for $t^* < t_{\text{idle}}, t_{\text{busy}}$, busy mode is more advantageous; the third region, for which $t_{\text{idle}} < t^* < t_{\text{busy}}$, is partitioned between the two strategies. The boundary between the optimal two modes is shown in red dashed line. For this plot we used $s = 0.01$.

2.5.3 Maximization and minimization of TF usage: the four combinations

The number of genes that can be associated with activators and repressors is restricted by the number of regulators of each type. When we require that a proportion q of

the genes is active and a proportion p of the regulators are activators, we distinguish four cases depending on the relative magnitudes of these variables:

- $q < p$ and $q < 1 - p$,
- $q > p$ and $q < 1 - p$,
- $q < p$ and $q > 1 - p$,
- $q > p$ and $q > 1 - p$.

In Fig 2.9, we illustrate how TF usage is maximized and minimized in each of these cases (one of them appeared as Fig 2.2D).

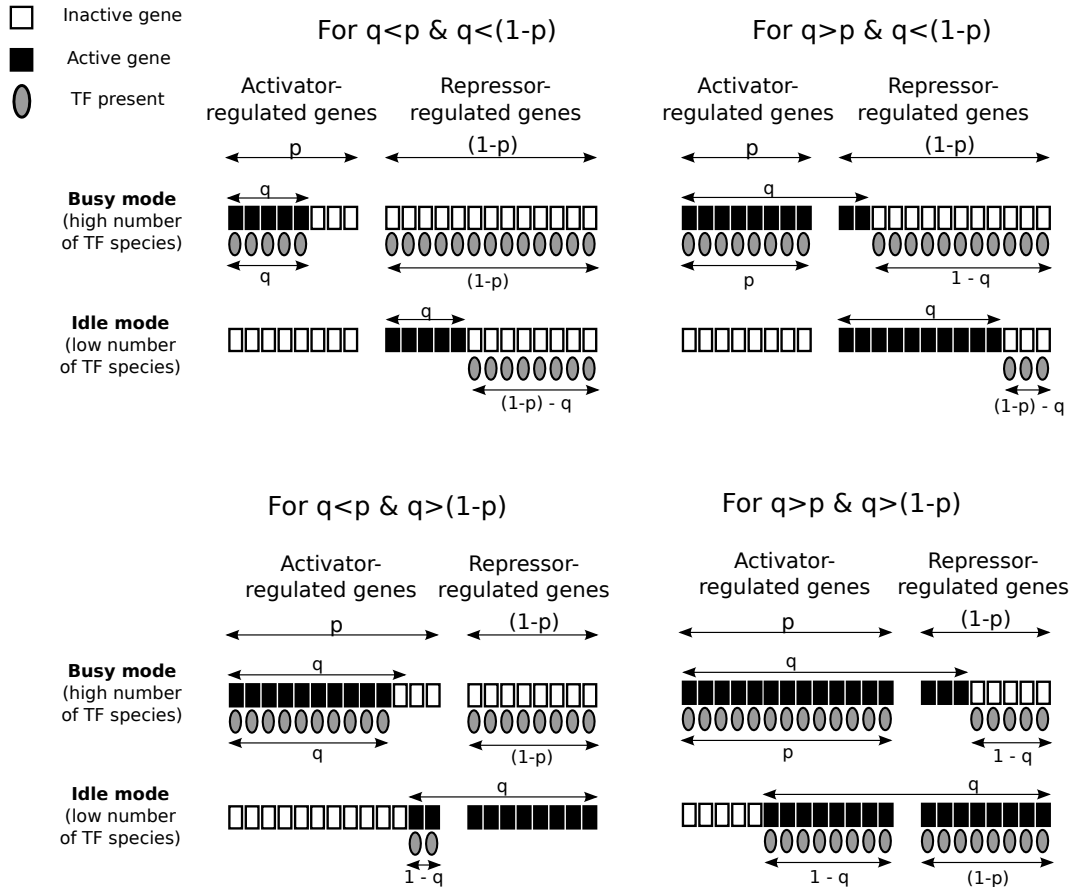

Figure 2.9: **Minimization and maximization for four different combinations of active genes q and activator regulated genes p.**

2.5.4 Crosstalk expression of the two strategies

Regime III

Crosstalk expression for 'busy' strategy In the busy mode, the proportion of active genes that are regulated by activators is $a = \min(p, q)$. The proportion of TFs involved in the busy strategy equals:

$$t_{\text{busy}} = (1 - p) + 2\min(p, q) - q. \tag{2.18}$$

To obtain the lower limit on crosstalk, we use t_{busy} in the equation for $X^*(t)$ and obtain:

$$X^*_{\text{busy}} = \left(1 - p + 2\min(p, q) - q\right) \cdot \left(-s(p + q - 2\min(p, q)) + 2\sqrt{s\left(p + q - 2\min(p, q)\right)}\right) \tag{2.19}$$

Crosstalk expression of 'idle' strategy Similarly, in the 'idle' mode, the proportion of active genes that are regulated by activators is $a = q - \min(1 - p, q)$. The proportion of involved TFs is then:

$$t_{\text{idle}} = (1 - p) - 2\min(1 - p, q) + q, \tag{2.20}$$

and the lower bound on crosstalk in the idle mode equals:

$$X^*_{\text{idle}} = \left(1 - p - 2\min(1 - p, q) + q\right) \cdot \left(-s(p - q + 2\min(1 - p, q)) + 2\sqrt{s\left(p - 1 + 2\min(1 - p, q)\right)}\right) \tag{2.21}$$

Regimes I and II

Crosstalk expression of both strategies in regime I The lower limit on crosstalk in regime I is described by $X^*(t) = t$ and:

$$X^*_{\text{busy}} = (1 - p) + 2\min(p, q) - q, \tag{2.22a}$$

$$X^*_{\text{idle}} = (1 - p) - 2\min(p, q) + q. \tag{2.22b}$$

Crosstalk expression of both strategies in regime II The lower limit on crosstalk in regime II is described by $X^*(t) = 1 - t/(1+st)$ and:

$$X^*_{\text{busy}} = 1 - \frac{(1-p) + 2\min(p,q) - q}{1 + s\left[(1-p) + 2\min(p,q) - q\right]}, \tag{2.23a}$$

$$X^*_{\text{idle}} = 1 - \frac{(1-p) - 2\min(p,q) + q}{1 + s\left[(1-p) - 2\min(p,q) + q\right]}. \tag{2.23b}$$

2.5.5 Proportion of TFs is always higher in busy mode

The proportion of TFs in busy mode is always higher than in idle mode. This can be easily shown by:

$$\begin{aligned} \Delta t &= t_{\text{busy}} - t_{\text{idle}} \\ &= \big((1-p) + 2\min(p,q) - q\big) - \big((1-p) - 2\min(1-p,q) + q\big) \\ &= -2q + 2\min(1-p,q) + 2\min(p,q), \end{aligned} \tag{2.24}$$

We distinguish between four cases:

- $q < p$ and $q < 1-p \Rightarrow \Delta t = -2q + 2q + 2q = 2q \geq 0$,
- $p < q$ and $1-p < q \Rightarrow \Delta t = -2q + 2(1-p) + 2p = 2(1-q) \geq 0$,
- $p < q < 1-p \Rightarrow \Delta t = -2q + 2q + 2p = 2p \geq 0$,
- $1-p < q < p \Rightarrow \Delta t = -2q + 2(1-p) + 2q = 2(1-p) \geq 0$.

In all cases, the difference $\Delta t > 0$, which shows that for any value of parameters, the proportion of TFs is always larger (or equal in the extreme case of $p \in \{0,1\}$) in 'busy' mode compared to 'idle'.

Idle mode minimizes, busy mode maximizes the number of transcription factor species t.

The busy and the idle mode maximize and minimize the fraction of TF species, respectively. This can be shown by taking a system, where a proportion of p genes is activator-regulated, while the rest $(1-p)$ is repressor-regulated. Within the activator-regulated genes, k_1 are active and $k_2 = p - k1$ are inactive. Moreover,

within $(1-p)$ repressor-regulated genes, k_3 are active genes and $k_4 = (1-p) - k_3$ inactive genes. Within these, several constraints exist:

- $k_1 + k_2 = p \quad \rightarrow \quad k_2 = p - k_1$,
- $k_1 + k_3 = q \quad \rightarrow \quad k_3 = q - k_1$,
- $k_3 + k_4 = 1 - p \quad \rightarrow \quad k_4 = (1-p) - k_3 = (1-p) - q + k_1$,

where q is the proportion of active genes. The total proportion of TF species is $t = k_1 + k_4 = 2k_1 + (1-p) - q$. Of course, due to the definition of the system, it holds:

- $k_1, k_2 \leq p$,
- $k_1, k_3 \leq q$,
- $q - k_1 \leq 1 - p$.

In attempt to see how the total proportion of TF species changes if we change different parameters of the system (k_i) with fixed q and p, we compute the derivative of t:

$$\frac{\partial t}{\partial k_1} = 2. \tag{2.25}$$

Therefore, the change of TFs with increasing k_1 is linear and 4 different scenarios exist. For each, the constraints described above must be met. Therefore:

1. if $q < 1 - p$ and $q > p$:
    - t is minimized by $k_1 \rightarrow 0 \Rightarrow k_2 = p, k_3 = q, k_4 = (1-p) - q$, which is the *idle mode*,
    - t is maximized by $k_1 \rightarrow p \Rightarrow k_2 = 0, k_3 = q - p, k_4 = 1 - q$, which is the *busy mode*.
2. if $q < 1 - p$ and $q < p$:
    - t is minimized by $k_1 \rightarrow 0 \Rightarrow k_2 = p, k_3 = q, k_4 = (1-p) - q$, which is the *idle mode*,

- t is maximized by $k_1 \to q \Rightarrow k_2 = p - q, k_3 = 0, k_4 = 1 - p$, which is the *busy mode*.

3. if $q > 1 - p$ and $q > p$:

   - t is minimized by $k_1 \to q - (1 - p) \Rightarrow k_2 = 1 - q, k_3 = 1 - p, k_4 = 0$, which is the *idle mode*,
   - t is maximized by $k_1 \to p \Rightarrow k_2 = 0, k_3 = q - p, k_4 = 1 - q$, which is the *busy mode*.

4. if $q > 1 - p$ and $q < p$:

   - t is minimized by $k_1 \to q - (1 - p) \Rightarrow k_2 = 1 - q, k_3 = 1 - p, k_4 = 0$, which is the *idle mode*,
   - t is maximized by $k_1 \to q \Rightarrow k_2 = p - q, k_3 = 0, k_4 = 1 - p$, which is the *busy mode*.

This formally proves what was graphically shown on Fig 2.9: minimization of TF proportion is achieved in idle mode while the maximization is obtained in busy mode.

2.5.6 For sufficiently high similarity measure, idle strategy always leads to a lower crosstalk limit X^*

For some parameter combinations of similarity s and fraction of regulated genes t, the mathematical result of the lower bound on crosstalk X^* has no biological relevance: (i) for sufficiently high similarity measure, regulation is ineffective and the lower bound on crosstalk X^* is obtained by no regulation (zero concentration of TFs, $C^* = 0$), and (ii) for high TF usage, the optimal concentration which minimizes the lower bound on crosstalk X^* diverges. Therefore, when only considering the biologically relevant regime, where ($0 < C^* < \infty$), the size of area where busy mode leads to lower crosstalk limit (Fig 2.3B red area) decreases with increasing similarity measure (Fig 2.3E).

In Fig 2.10, we plot the difference in optimal crosstalk $\Delta X^* = X^*_{\text{idle}} - X^*_{\text{busy}}$, where

black areas denote the anomalous regime. With increasing similarity, the anomalous regime grows and covers an increasingly larger portion of the phase space. For sufficiently high similarity values, the idle strategy will always lead to the most optimal crosstalk for any value of (p, q).
Fig 2.11 shows both the minimal (blue) and maximal (red) value of $\Delta X^* = X^*_{\text{idle}} - X^*_{\text{busy}}$ over the whole space of $p \in \{0, 1\}$ and $q \in \{0, 1\}$ as a function of similarity s. Values of $\Delta X^* < 0$ mean that the idle mode leads to a lower crosstalk limit and vice versa for $\Delta X > 0$. Therefore, when the busy mode completely vanishes and the idle mode is the one that always yields lower crosstalk ($\Delta X^* < 0$ for all (p, q) values), the maximal value of ΔX^* will be negative; when $\max_{p,q} \Delta X^* = 0$, the similarity value is such that the busy mode completely vanishes. That happens at $s_{\text{vanishing}} \approx 5$, which is far above the values of real organisms – Fig 2.4A shows that similarity values of *S. cerevisiae* range between $s \approx 10^{-5} - 10^{0}$.

2.5.7 Probabilistic gene activity model

Probabilistic model description

So far, we considered a deterministic model in which the numbers of active genes and available TF species were fixed, resulting in a single crosstalk value per (p, q) configuration. In reality, these numbers can temporally fluctuate, for example, because of the burst-like nature of gene expression [Golding *et al.*, 2005; Wang *et al.*, 2009]. In the deterministic model, we also assumed uniform gene usage, such that all genes are equally likely to be active. In reality, some genes are active more frequently than others.

To account for this, we study the following crosstalk in a probabilistic gene activity model. We assume independence between activities of different genes, where each gene i has demand (probability to be active) D_i. We then numerically calculate crosstalk for a set of genes. This approach enables us to incorporate a varying number of active genes and a non-uniform gene demand and compare our results to the deterministic model studied above.

Assume that to comply with its demand D_i, each gene i is regulated with

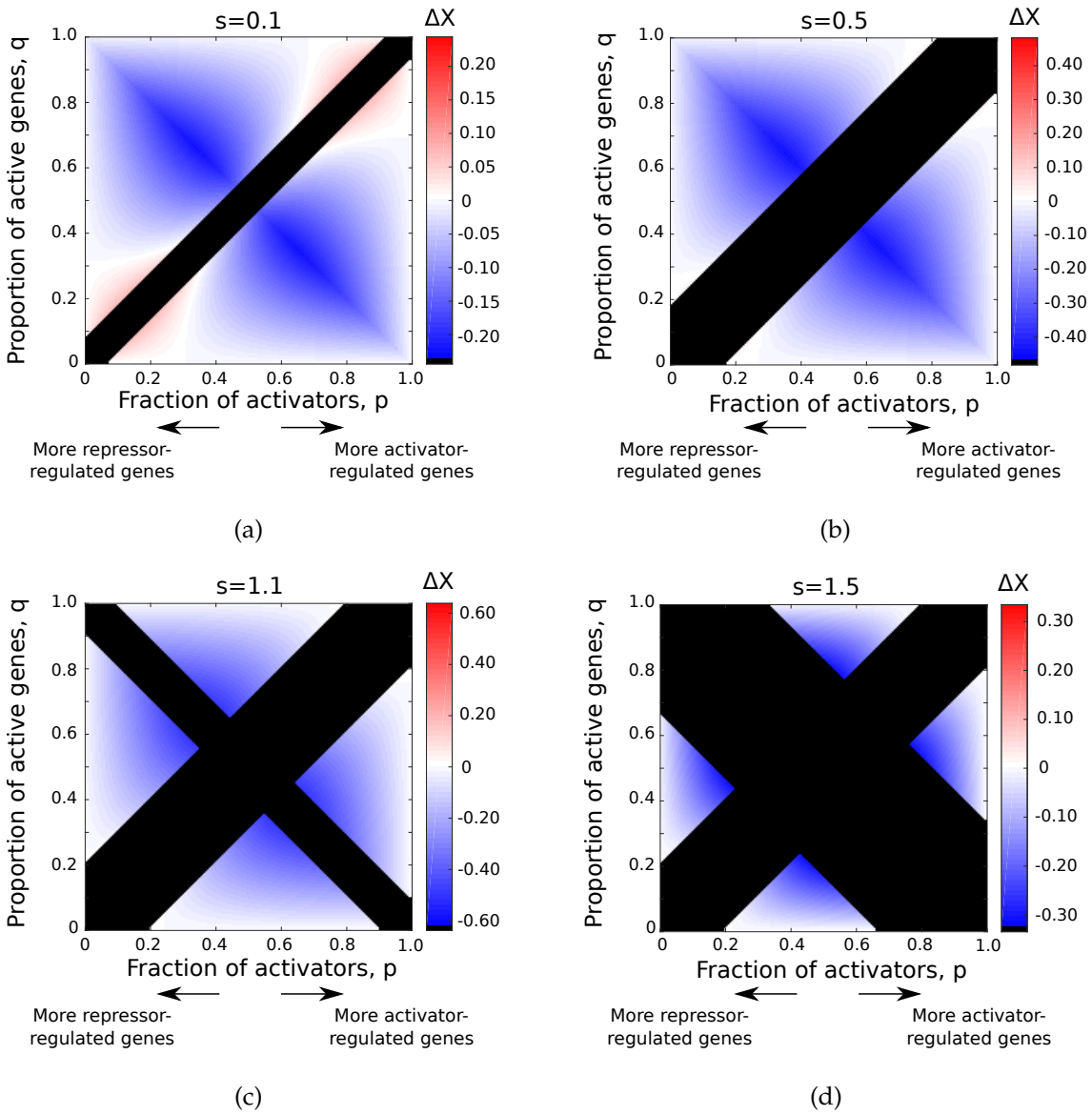

Figure 2.10: **The anomalous regions where crosstalk cannot be minimized, grow as the similarity s increases.** Difference in optimal crosstalk $\Delta X^* = X^*_{\text{idle}} - X^*_{\text{busy}}$, where black areas denote the anomalous regime. Different values of rescaled similarity were used: (**a**) $s = 0.1$, (**b**) $s = 0.5$, (**c**) $s = 1.1$, (**d**) $s = 1.5$.

probability γ_i, $i = 1...M$, where $\gamma_i = D_i$ if regulation is positive and $\gamma_i = 1 - D_i$ if it is negative. We assume that the TF species needed for these genes are available in the cell. The distribution $f_t(t)$ of the fraction of TF species follows Poisson-Binomial

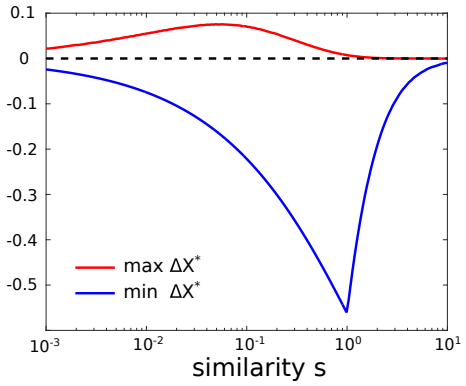

Figure 2.11: **The differences between regulatory strategies depend on** s**.** We plot the minimal (blue) and maximal (red) value of $\Delta X^* = X^*_{\text{idle}} - X^*_{\text{busy}}$ over the entire region of $p \in \{0,1\}$ and $q \in \{0,1\}$ as a function of similarity, s. The point where the maximal value (red) becomes negative (at $s \approx 5$), is where the busy mode completely vanishes and idle mode leads to lowest crosstalk (for any (p,q)).

distribution [Rice, 2006] with mean and variance of:

$$\langle t \rangle = \frac{1}{M} \sum_{i=1}^{M} \gamma_i \tag{2.26a}$$

$$\text{var}(t) = \frac{1}{M} \sum_{i=1}^{M} \gamma_i (1 - \gamma_i). \tag{2.26b}$$

Assuming that the number of genes is large, $M \gg 1$, the central limit theorem applies here: we approximate the probability distribution of t, $f_t(t)$, by a Gaussian distribution with the mean value and variance as given with Poisson-Binomial distribution (mean value $\langle \gamma_i \rangle_i$ and standard deviation $\sigma_t = \langle \gamma_i (1 - \gamma_i) \rangle_i$, with $\langle \cdot \rangle_i$ representing the average over all genes) [Rice, 2006].

Since minimal crosstalk X^* is a function of t (Eq 2.5), X^* becomes a random variable and its distribution reads:

$$f_{X^*}(X^*) = \sum_l f_t(g_l^{-1}(X^*)) \left| \frac{\mathrm{d} g_l^{-1}(X^*)}{\mathrm{d} X^*} \right|, \tag{2.27}$$

where $g_l^{-1}(X^*) = t_l$ represents the $l-$th branch of the inverse function (for some X^* values there exist two solutions t_l that satisfy the inverse equation) [Rice, 2006]. The solutions for $g_l^{-1}(X^*)$ and its derivative exist and can be analytically computed, which enables us to solve for the distribution of crosstalk $f_{X^*}(X^*)$:

1. Region I: $f^*_X(X^*) = f_t(X^*)$

2. Region II: $f_X^*(X^*) = f_t\left(1 - \frac{X^*}{1-\alpha+\alpha X^*}\right) \cdot (\alpha - 1 - \alpha X^*)$

3. Region III: see S2 Appendix,

where $f_t(t)$ is the distribution of TF usage values t.
As the exact calculation of the distribution $f_{X^*}(X^*)$ is difficult, we can often use the following useful approximation. If the distribution is narrow enough, such that $\sigma_T/\langle T\rangle \ll 1$, we can approximate the expected value of crosstalk by the deterministic value of crosstalk for an expected value of available TF fraction:

$$\langle X^*(t)\rangle \approx X^*(\langle t\rangle), \tag{2.28}$$

where the computation of both $\langle t\rangle$ and $X^*(\langle t\rangle)$ is straightforward given γ_i. We discuss below the conditions under which this approximation holds. The distribution of X^* is typically narrow, such that for practical purposes the distribution mean provides a very good estimator of crosstalk values.

Approximations

In our stochastic model, a gene i is regulated with probability $\gamma_i, i = 1, \ldots, M$. Above, we stated (i) that the distribution of TFs in use, t, can be well approximated by a Gaussian distribution. Furthermore, if (ii) in the regime where X^* is linear in t and $\sqrt{var(t)}/\langle t\rangle \ll 1$, one can approximate the expected value of crosstalk by the deterministic value of crosstalk for an expected value of total number of TF: $\langle X^*(t)\rangle \approx X^*(\langle t\rangle)$. The third claim is that if X^* is linear in t ($\partial X^*/\partial t \approx$ const.), one can also approximate well the distribution of crosstalk with a Gaussian distribution having the same mean and variance as those of the t distribution, just rescaled and translated by the slope and constant factor of the linear transformation of $X^*(t)$.
The Gaussian approximation of distribution of t follows from the central limit theorem. A numerical example is shown in Fig 2.12 (a) and (c).
The approximation (ii) uses linearity to show that $\langle X^*(t)\rangle \approx X^*(\langle t\rangle)$ holds:

$$\text{If } X^*(t) \approx \alpha t + \beta \implies \langle X^*(t)\rangle = \langle \alpha t + \beta\rangle = \alpha\langle t\rangle + \beta = X^*(\langle t\rangle). \tag{2.29}$$

The linearity assumption is fulfilled for t values that are much lower than t^* (t^* being the value at which X^* reaches maximum). For visual example of linearity of

$X^*(t)$ for $t < t^*$, see Fig 2.2A. Moreover, by using the Taylor expansion of crosstalk $X^*(t)$ for a small deviation of t around its mean $\langle t \rangle$, we show that deviations around the expected value are small and often negligible. Therefore, if $\sqrt{var(t)}/\langle t \rangle \ll 1$ holds, one can look at a representative deviation from the mean value and write (assuming the solution for $X^*(t)$ is in biologically plausible regime III);

$$\begin{aligned} X^*(t) = X^*(\langle t \rangle) + \delta X^*(t) &\approx X^*(\langle t \rangle) + \frac{\partial X^*(\langle t \rangle)}{\partial t}\delta t \\ &= X^*(\langle t \rangle) + \left[2\langle t \rangle s - s + 2\sqrt{s(1-\langle t \rangle)} - \frac{s\langle t \rangle}{\sqrt{s(1-\langle t \rangle)}} \right] \delta t \\ &= X^*(\langle t \rangle) + \underbrace{\left[\frac{X^*(\langle t \rangle)}{\langle t \rangle} - \langle t \rangle s \left(\frac{1}{\sqrt{s(1-\langle t \rangle)}} - 1 \right) \right] \delta t}_{\delta X^*} \end{aligned} \tag{2.30}$$

Fig 2.13 shows the relative error $\delta X^*(\langle t \rangle)\delta t / X^*(\langle t \rangle)$ of our approximation. We use a representative values of $M = 2500$ and $\gamma_i = 0.5$, leading to the standard deviation of t being $\sqrt{var(t)} = 10^{-2}$. We take this number to also be a variation in the number of TF species present: $\delta t = 10^{-2}$. Any larger values of M or any other values of γ_i will lead to lower error. The relative error is indeed very small. The exceptions are the values close to $t = 1$, which fall out of regime III into anomalous regime II, and values close to $t_0 = 0$, which still take a small relative error of $\delta X^*(\langle t \rangle)\delta t / X^*(\langle t \rangle) = 20\%$ for $\langle t \rangle = 0.05$. Furthermore, if the third claim of $X^*(t)$ linearity with respect to t holds, we can approximate the distribution of $X^*(t)$ by a Gaussian. As the distribution of t is Gaussian, to a good approximation, a linear transformation of a Gaussian distribution also leads to a Gaussian distribution of $X^*(t)$. Fig 2.12 (a-b) shows an example of distribution of t and X^*. There, the probabilities γ_i give the average proportion of TF species $\langle t \rangle = 0.5$, which gives values of crosstalk that are far from the maximum of X^*. The assumption of linearity is justified on the example shown and Gaussian approximation for $f_{X^*}(X^*)$ gives good results. The expected values of crosstalk $\langle X^*(T) \rangle$ and the crosstalk of the expected value of TF species $X^*(\langle T \rangle)$ have a very small relative difference (in the order of 0.01%), which is the consequence of small ratio $\sqrt{var(t)}/\langle t \rangle = 2\% \ll 1$.

On the other hand, Fig 2.12 (c-d) shows the distribution of t and X^*, where the

probabilities γ_i give the average proportion of TF species $\langle t \rangle = 0.66$ close to the maximum of X^* around $t \approx 2/3$, where the linearity assumption does not hold. We see that Gaussian approximation for $f_{X^*}(X^*)$ is not valid anymore. Even though the linearity assumption does not hold, the expected value of crosstalk $\langle X^* \rangle$ and the crosstalk of the expected value of the relative number of TF species $X^*(\langle T \rangle)$ are again very close (relative difference of 0.03%).

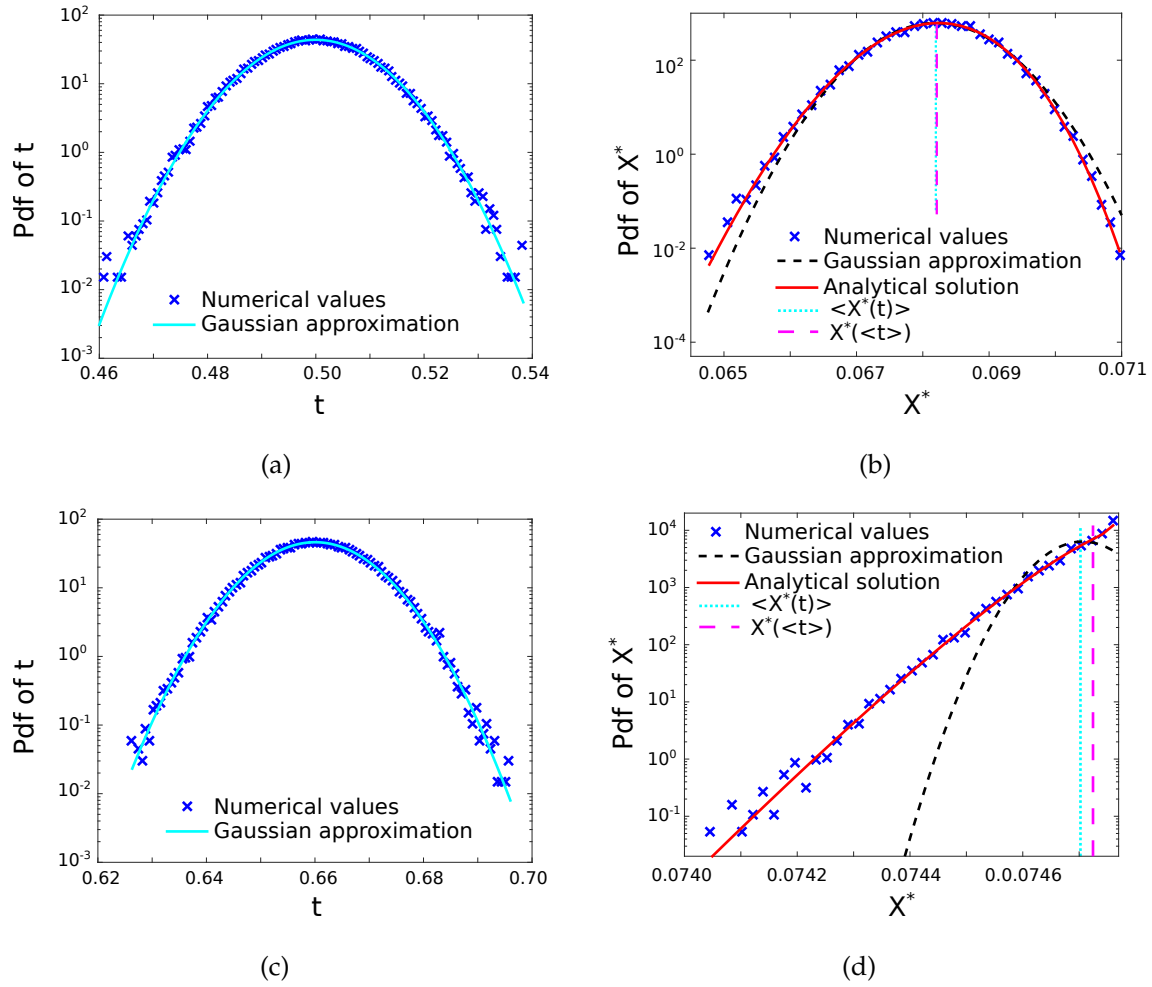

Figure 2.12: t **distribution is always well-approximated by a Gaussian. If** $\langle t \rangle$ **is far apart from** t^***,** X^* **distribution is also well approximated by a Gaussian.** We plot the distributions of t ((**a**), (**c**)) and $X^*(t)$ ((**b**), (**d**)) in two cases. The vertical lines in (**b**) and (**d**) (dotted and dash-dotted) represent $X^*(\langle t \rangle)$ and $\langle X^*(t) \rangle$, correspondingly. We find an excellent match between their values, even in the worst case scenario that $X^*(t)$ is far from being Gaussian (d). Small discrepancies between the analytical solution and numerical simulation is due to the finite number of iterations in the simulation. Parameter values: $s = 0.01$, $M = 3000$, $p = 1/3$, in (**a**) and (**b**) $\gamma_{\mathbf{i}} = \mathbf{0.66}$; in (**c**) and (**d**) $\gamma_{\mathbf{i}} = \mathbf{0.66}$.

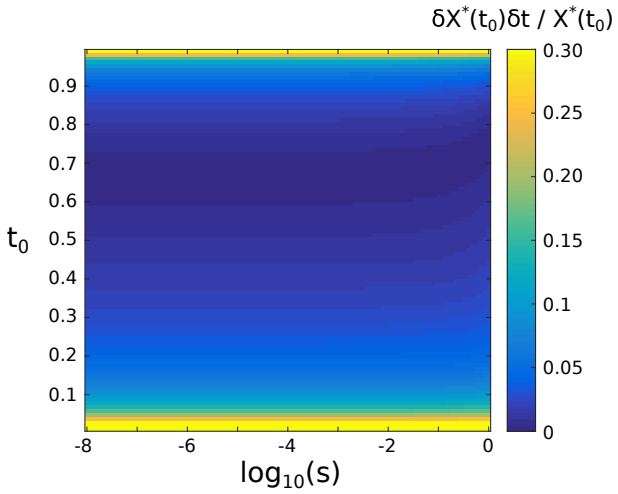

Figure 2.13: **Relative error of our approximation of distribution of crosstalk.** The relative error $\delta X^*(t_0)\delta t / X^*(t_0)$, shown in color, as a function of similarity, s, and the expected value of distribution t, $t_0 = \langle t \rangle$. Parameter values: $M = 2500$ and $\gamma_i = 1/2$, leading to $\delta t = 10^{-2}$. All other values of γ_i would lead to lower values of δt and therefore to lower values of the relative error.

The probabilistic gene activity model leads to a distribution of the number of active genes - Example

In the probabilistic model, the fraction of active genes q becomes a random variable, rather than being fixed, as we assumed before. We demonstrate this in an example below. The crosstalk behavior in the probabilistic case can be obtained as a superposition of the relevant deterministic cases taken with their corresponding weights.

Assume we have 3 genes, active with probabilities $p_1 = 1/2$ and $p_2 = p_3 = 1/4$, correspondingly. We can then enumerate all active gene combinations and active state probabilities:

- all genes inactive: $p(\text{all inactive}) = \frac{1}{2}\left(\frac{3}{4}\right)^2 = 9/32$
- first gene active: $p(\text{gene 1 active, genes 2,3 inactive}) = \frac{1}{2}\frac{3}{4}\frac{3}{3} = 9/32$
- second gene active: $p(\text{gene 2 active, genes 1,3 inactive}) = \frac{1}{2}\frac{1}{4}\frac{3}{4} = 3/32$
- third gene active: $p(\text{gene 3 active, genes 1,2 inactive}) = \frac{1}{2}\frac{3}{4}\frac{1}{4} = 3/32$

- first&second genes active: p(genes 1,2 active, gene 3 inactive) = $\frac{1}{2}\frac{1}{4}\frac{3}{4} = 3/32$
- first&third genes active: p(genes 1,3 active, gene 2 inactive) = $\frac{1}{2}\frac{3}{4}\frac{1}{4} = 3/32$
- second&third genes active: p(genes 2,3 active, gene 1 inactive) = $\frac{1}{2}\frac{1}{4}\frac{1}{4} = 1/32$
- all genes active: p(all active) = $\frac{1}{2}\frac{1}{4}\frac{1}{4} = 1/32$.

The mean number of active genes Q is then:

$$\langle Q \rangle = 0 \times p(\text{all inactive}) + 1 \times p(\text{gene 1 active, genes 2,3 inactive}) + \cdots + 3 \times p(\text{all active}) = 1. \tag{2.31}$$

Therefore, in this example, on average, one gene is active.
However, for $\langle Q \rangle = 1$ and fixed p (fraction of activator among the existing regulators), there are several possible q values (proportion of active genes). This explains why we have a distribution of crosstalk values if genes are active with some probability.

2.5.8 Data-based crosstalk calculations

Distribution of similarity measures for *S. cerevisiae* genes is relatively wide

To obtain similarity and crosstalk values of real organisms, one needs to take several aspects into consideration. First, the exact consensus sequences of different TFs are not known and position count matrices (PCMs) are used to infer them. Second, the length of binding sites and consensus sequences between different BSs and TFs can differ. Third, in a more realistic case, each TF can be cognate for multiple genes. All these concerns (and others, for more details, see Section 2.4) complicate a calculation of lower bound on crosstalk in a real organism. However, there are ways to solve these issues and obtain estimations to be compared with our analytical solutions.

We define similarity between a binding site k and transcription factor l as $S_{kl} = \exp(-E^{kl})$, where E^{kl} represents the mismatch energy of binding of the transcription factor on the binding site. The similarities between all pairs of consensus binding sites for *S. cerevisiae* are shown in Fig 2.14. The results are not symmetric between transcription factors and binding sites. A simple example with two transcription factors and their cognate binding sites can be presented to understand this intuitively:

imagine the first transcription factor with a shorter consensus sequence while the consensus sequence of the second one is longer. Let us assume that a consensus of the shorter TF is included in the consensus of the longer TF. The TF with the shorter consensus sequence will bind easily to the binding site of the second TF. Therefore, the similarity between the transcription factor with the shorter consensus sequence and longer binding site will be high. However, the transcription factor with the longer consensus sequence and a shorter binding site would have a lower similarity, as it is less likely that the long transcription factor binds to the short binding site. Indeed, the matrix of pairwise similarity values is asymmetric. Clearly, we observe many vertical lines of similar value (TFs that easily bind many binding sites - yellow strips, or that are very unique to only few binding sites - blue strips), but much weaker signatures of rows (binding sites that are very similar or very dissimilar to all others). This demonstrates that the similarity value between a transcription factor and a binding site is dominated by the transcription factor properties, and much less by the binding site's. High similarity between a transcription factor and other binding sites is highly correlated with short consensus sequences of that transcription factor.

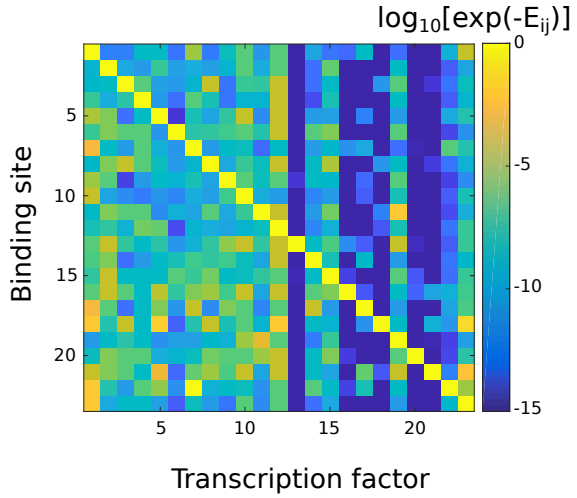

Figure 2.14: **Similarity values between all pairs of consensus binding sites and transcription factors for *S. cerevisiae*.** Columns represent TFs, rows represent binding sites.

The similarity measure of a gene i, S_i, is determined as the contribution of all

non-cognate transcription factors:

$$S_i = \sum_{\text{j} \equiv \text{over all BSs, } j \neq i} C_j e^{-E_{ij}}, \tag{2.32}$$

with C_j being the concentration of TF species j.

TFs that bind shorter DNA stretches are more promiscuous

TFs with shorter consensus BS can fit more binding sites. Subsequently, their similarity value s_i is higher. This is indeed what we find in our yeast data – see Fig. 2.15.

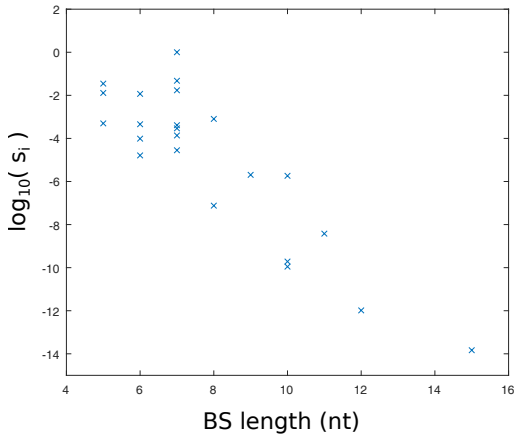

Figure 2.15: **TFs with shorter consensus BSs tend to have higher similarity values.** The data points show TF similarity values calculated for the *S. cerevisiae* dataset.

Alternative calculations of similarity values and crosstalk from data

Similarity values and crosstalk of *S. cerevisiae* and other organisms were estimated [Friedlander *et al.*, 2016] based on PCM data of the TFs in our previous work, but using a different computational approach. The main difference between these two calculations is that in the first approach, the similarity was calculated between a consensus sequence of a particular TF and an ensemble of binding sequences of the same length randomly drawn from a uniform distribution. It was shown there analytically that the average similarity between random sequences of length L and uniform energy mismatch per position ϵ is simply $S = \left(1/4 + 3/4 \exp\left(-\epsilon\right)\right)^L$,

hence only this effective ϵ needs to be calculated. In contrast, in the current work, we calculate similarity between actual pairs of TF and non-cognate binding sites.

Another major difference is the calculation of mismatch energy penalties. By using the PCMs, we obtain an energy matrix which gives energy penalties for every position and every nucleotide separately. In the previous work, only an effective ϵ uniform for all positions was calculated using either of two approaches: (i) information method [Wunderlich and Mirny, 2009] or, (ii) pseudo-count method (also used here) [Schneider *et al.*, 1986; Berg and von Hippel, 1987]. In the information method, the total information of the motif was calculated and then an effective ϵ_{eff} which evenly distributed this information between all L positions, was calculated. The advantage of this method is the avoidance of pseudo-count usage, which could bias the results. Its major drawback is the lack of position-specific energy information which we need to calculate similarity between actual pairs of binding sites. In the pseudo-count method, a pseudo-count is added to all positions in order to avoid zero counts, which result in infinite energy penalty. While this method can provide position-specific energy values ϵ_j, in the previous work, only an average of all positions $\epsilon_{\text{eff}} = \sum_j \epsilon_j$ was taken as the effective value for the similarity with respect to random sequences. In the current work, the pseudo-count method was used differently, computing the similarity measure of a gene j by directly following the definition and summing the Boltzmann weights over all TFs (i.e., sum over exponents of energies, $\sum_i \exp(-E_{ij})$). Since binding sites and TFs can have different lengths, there could be different relative positions with respect to each other, which could have different binding energies. Here, we chose the relative position with highest match (lowest energy penalty) between the binding TF i and binding site j. States with lower energy are energetically more favorable and therefore physically more likely to occur.

This difference in estimating energy penalties leads to a different approach for computing similarity measures. In our approach, we use energy matrices to compute the energy of binding for every pair of TF-BS, i.e., E_{ij}.

The two distinct approaches lead to different, but similar, distributions of similarity measures for a given gene j, s_j – Fig 2.16. The main difference are long tails of the

current approach. The median value of the similarity with the previous approach in *S. cerevisiae* was median($s^{\text{consensus-random}}$) = $0.8 \cdot 10^{-4}$, while in the current approach, we obtain median($s^{\text{TF-BS}}$) = $1.4 \cdot 10^{-4}$.

Differences between s values obtained in these two approaches can emanate from various sources:

- sequences of actual binding sites are not well captured by a uniform distribution because of biases in favor of AT-content. For example in our *S. cerevisiae* dataset, we have 31% of nucleotide A, 21% of C, 22% of G, and 26% of T.
- actual binding sites can vary in length; taking the relative position with best match is clearly non-random. See in Fig 2.16 a comparison to s values calculated when the relative position is randomly selected.
- equally partitioning the total energy of the motif between all its positions consistently under-estimates the similarity.
- if actual TF-BS are considered, insufficient data can lead to biases in similarity estimates.

Using real transcription factor copy numbers

So far, we assumed that all transcription factors are present in equal concentrations (Eq 2.1), to simplify the calculations and enable the analytical derivation of the lower bound on crosstalk. We then minimized crosstalk with respect to the TF concentrations, such that these concentrations and the energy gap E_a between cognate bound and unbound states were all left out of the minimal crosstalk expression (Eq 2.5). In general crosstalk does depend on the TF concentrations and can be numerically calculated for general concentration values. Such calculations require an extension of the crosstalk minimization procedure and additional parameter values which were not necessary for the lower bound. In this section we demonstrate this calculation using experimentally measured proteome data of *S. cerevisiae*. We follow the model described in [Landman *et al.*, 2017] for a single gene and generalize it for our case.

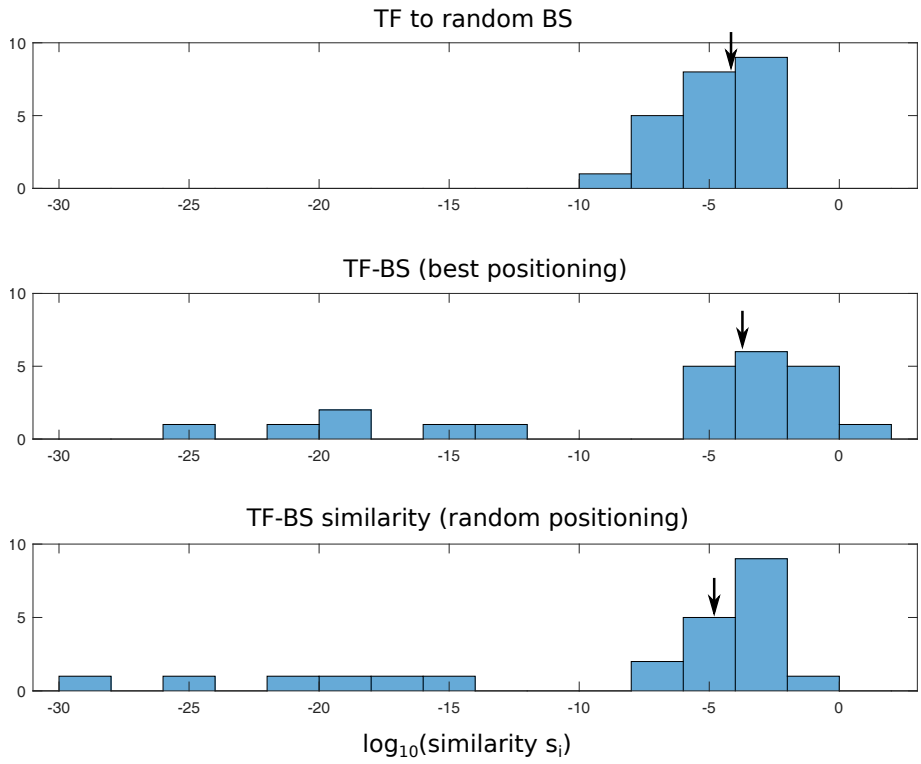

Figure 2.16: **Comparison of similarity distribution between the different approaches**; alternative to previous work (top) versus the current work with best positioning (middle) and random positioning (bottom). Random positioning takes a random binding location instead of the one with the highest match (the distribution shown is from one representative realization). The median value (averaged over many realizations) of median($s^{\text{TF-BS-random}}$) = $0.2 \cdot 10^{-4}$.

We obtain the copy number values for all 23 of TFs measured in [Ghaemmaghami *et al.*, 2003]. For proteins which were below the detection level of 50 molecules/cell, we take half the detection level of 25 molecules/cell. Generally speaking, a transcription factor molecule can be in one of three reservoirs: bound to its cognate site, bound to any non-cognate site, free in the cytoplasm or bound non-specifically to the DNA. Non-specific binding is independent of the DNA sequence and has no effect on crosstalk, but only effectively reduces the TF availability. As until now, we focused on minimal crosstalk assuming the TF availability is optimized we did not include non-specific binding in our expressions. Now, in order to properly account for the available TF copy numbers we add it to the expressions. We use the grand-canonical ensemble formulation to estimate the fugacities, i.e., the available number of TF molecules.

We write down the mass balance equation for each of the 23 TF molecules (compare to Eq 40 in [Landman *et al.*, 2017]):

$$A_j = N^{ns}\Theta_j^{ns} + \sum_{k\neq j} N_k\Theta_{jk}^{nc} + N_j\Theta_j, \tag{2.33}$$

where A_j is the total number of molecules of the j-th TF, and N^{ns} is the total number of available non-specific binding sites, N_j is the total number of cognate binding sites for j-th TF and N_k is the total number of non-cognate binding sites (which are cognate for the other TFs). Θ_j^{ns} is the occupancy of non-specific binding site by the j-th TF:

$$\Theta_j^{ns} = \frac{\lambda_j e^{-E_j^{ns}}}{e^{-E_a} + \sum_i \lambda_i e^{-E_i^{ns}}}, \tag{2.34}$$

where λ_j represents the fugacity of j-th TF, E_j^{ns} the binding energy of j-th TF to non-specific site, E_a the energy of the unoccupied state, and the sum in the denominator goes over all TFs. The product $N^{ns}\Theta_j^{ns}$ gives the total number of molecules of j-th TF, bound to all non-specific sites.

The second term in Eq 2.33 represents the total number of j-th TF molecules that are bound to non-cognate binding sites. These are cognate sites for all $k \neq j$-th TFs. Each term in the sum represents the number of j-th TF molecules bound to cognate binding sites of k-th TF. Similarly, Θ_{jk}^{nc} is the occupancy of one cognate binding site of k-th TF by the j-th TF:

$$\Theta_{jk}^{nc} = \frac{\lambda_j e^{-E_{jk}}}{e^{-E_a} + \sum_i \lambda_i e^{-E_{ik}}}, \tag{2.35}$$

where E_{jk} represents binding energy of j-th TF to cognate binding site of k-th TF. E_{jj} represents energy of cognate binding of j-th TF.

The last term in Eq 2.33 represents the number of j-th TF molecules bound to one of its N_j cognate sites. Again, the occupancy of a cognate site by the j-th TF equals:

$$\Theta_j = \frac{\lambda_j e^{-E_j}}{e^{-E_a} + \sum_i \lambda_i e^{-E_{ij}}}. \tag{2.36}$$

This gives us a set of 23 coupled non-linear equations with 23 variables λ_j (the fugacity values of all TFs). To set the energy scale we define for every binding site

that its energy level when bound by its cognate factor is zero. The parameters we need to specify in to numerically solve these equations are:

- A_j the total number of molecules of the j-th TF species per cell. We obtain these values from [Ghaemmaghami *et al.*, 2003].

- E_j^{ns} the binding energy of the j-th TF to a non-specific site. We assume that non-specific sites are random DNA sequences [Gerland *et al.*, 2002]. Hence, we calculate these energies by averaging energy contributions at each position using the TF energy matrix and sum the individual contributions.

- N^{ns}, the total number of non-specific sites. The yeast genome length is roughly 10^7 nucleotides, but we assume that only 10% of it is accessible [Wunderlich and Mirny, 2009]. Hence we assume that 10^6 non-specific binding sites are available.

- N_j, the number of cognate binding sites of j-th TF. Here we simply take the number of genes regulated by the j-th TF.

- E_{jk}, the binding energy of the j-th TF to a cognate binding site of the k-th TF. These values are used in calculation of similarity matrix S_{ij} – see Fig. 2.14.

- E_a the energy gap of a binding site between its state when occupied by its cognate factor (which we set as zero) and its unoccupied state. We assume the same value for all TFs. Unfortunately, measurement of this parameter are rare. We found estimates of this energy gap only for a few bacterial [Gerland *et al.*, 2002] and yeast [Maerkl and Quake, 2007] TFs. In the following we show crosstalk calculations for several different values of this parameter.

The binding energy is usually sequence dependent. However, when the binding becomes very unfavorable, other contributions come into play, thus effectively setting a bound on the binding energy. Therefore, we have used this bound on all binding energies E_j^{ns} and E_{jk}. Using all the details described above, we can numerically obtain all fugacity values λ_j, for all j and then calculate the total crosstalk.

However, the solutions of crosstalk using the correct fugacities seem to be relatively sensitive to the threshold values we set. Here we show solutions for a range of realistic values (Fig 2.17 left).

Furthermore, as the energy of a non-occupied state (E_a) differs between different TF, we investigated a range of realistic values to see their effect. As each individual position contributes $1.5 - 3.5 k_B T$ [Wunderlich and Mirny, 2009] and the average binding site size is between 6 to 10 bp, we estimate that $E_a \in (10, 16)$. Fig. 2.17 right shows that while results are robust for $t < 1/2$, they differ for values $t \approx 1$.

All results using real concentrations exhibit higher crosstalk estimates compared to our lower bound where optimal concentrations were used (Fig 2.17).

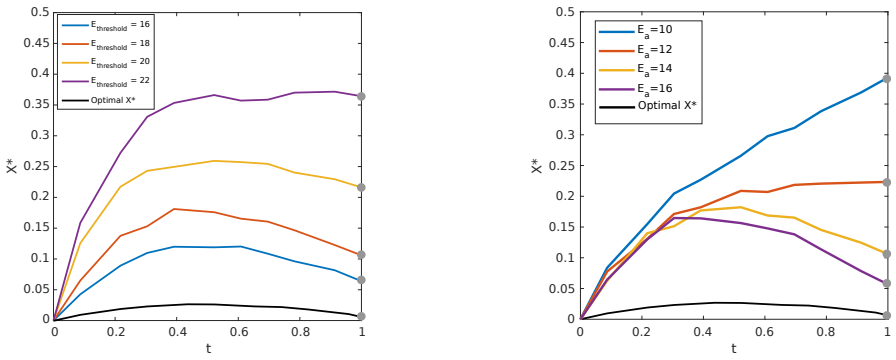

Figure 2.17: **Estimated crosstalk of *S. cerevisiae* 23 TFs using measured TF concentrations [Ghaemmaghami *et al.*, 2003] is higher than the minimal crosstalk calculated for optimal TF concentrations.** We illustrate crosstalk values as a function of the proportion of available TF species t, using measured TF concentrations [Ghaemmaghami *et al.*, 2003], for various energy thresholds (left) and energies of unoccupied binding sites E_a (right). For comparison, optimal crosstalk curve is added (black). We defined $E = 0$ as the state with of the binding site when a cognate TF is bound. In the left figure we used $E_a = 14$, and in the right figure we used $E_{\text{threshold}} = 18$. The gray dots at $t = 1$ represent points where all measured TFs copies were included in the crosstalk calculation. Otherwise, for each $t < 1$ we drew 300 times by random a subset of TFs to be present, such that only t proportion of the genes are regulated. Fugacities of TFs that were not chosen were set to zero.

2.5.9 Comparison of our global crosstalk model to single-gene models

Previous models of regulatory crosstalk [Shinar *et al.*, 2006; Sasson *et al.*, 2012] studied crosstalk at the single gene level and focused on the choice between activation and repression as the leading mode of regulation. The main innovation of our model is the consideration of multiple genes and multiple regulators simultaneously. As we demonstrate in this section, this leads to a completely different optimum of the system. For every single gene its crosstalk is minimized if its regulator is present in the highest possible level, when it should be regulated. Yet, such high levels of regulators increase crosstalk probability for all other genes - an effect which is overlooked in single gene models. Alternatively, when the gene should not be regulated (its binding site should be unoccupied), its crosstalk level is minimized if the TF concentration is zero. Only a multiple-gene model can correctly account for the trade-off between increasing TF concentration for the genes that should be regulated, but simultaneously keeping total TF levels as low as possible to reduce crosstalk of all other genes that should not be regulated at that time. Then, there is an optimal intermediate TF concentration which is neither zero nor the maximal. This optimal concentration does *not necessarily* minimize crosstalk for any one particular gene, but rather minimizes the *total* crosstalk, of all genes together.

This section shows a comparison of our model which minimizes *global* crosstalk, and a model that minimizes *local* crosstalk - it minimizes crosstalk of each individual gene, disregarding the potential interactions between every TF and all other non-cognate binding sites. If crosstalk were minimized with respect to concentration for individual genes, the lower bound on crosstalk of individual genes would be obtained for concentration $c \to \infty$. Note that this represents concentration of TFs that are regulating while by construction of our model, non-regulating TFs have zero concentration. This result can be easily seen by looking at minimum of x_{bound}

of Eq 2.2. Using this concentration, the contributions of individual genes become

$$x_{\text{bound}} = \frac{ts}{1+ts}, \tag{2.37}$$

$$x_{\text{unbound}} = 1, \tag{2.38}$$

where the limit $c \to \infty$ was taken in Eq 2.13. The total crosstalk, namely the average fraction of genes in any crosstalk state, is then

$$X = t\ x_{\text{bound}} + (1-t)\ x_{\text{unbound}} = \frac{t^2 s}{1+ts} + (1-t), \tag{2.39}$$

which, for $s \ll 1$, leads to $X \approx (1-t)$, which is the proportion of non-regulated genes. The large increase of crosstalk comes from optimization of crosstalk values of individual genes. We minimize it for the case when gene is regulated (x_{bound}), obtaining large concentrations of all TFs that are required to regulate. This will enforce that each of these individual genes will suffer barely any crosstalk when it is being regulated. However, by doing this we overlook the inevitable large TF concentrations needed that are likely to cause crosstalk to other genes that are not being regulated. Therefore, the main contribution to crosstalk comes from genes that should be unregulated (i.e., binding sites that should be unoccupied) but are instead bound by the ample non-cognate TFs.

See Fig. 2.18, for a comparison with X^* obtained by *global* minimization (Eq 2.5), which exhibits significantly lower crosstalk values. As TF concentrations are limited by biophysical constraints such as cell volume and protein production costs, concentrations are finite. It is also known experimentally, that large proportions of genes can be left unregulated, rather than being constantly induced by non-cognate binding of TFs. Hence, we conclude that single-gene crosstalk models present only a partial picture and are inadequate to study gene regulatory networks. The following table shows the main features of global vs. local minimization:

2.5.10 Complex regulatory architectures

We studied a simple regulatory architecture, where every gene is regulated by a single TF, which is either an activator or a repressor. Here we analytically study two more complex regulatory architectures and compare them to the basic model.

	Local minimization	Global minimization
Minimization on	individual genes	all genes
Conc. (present TFs)	$c = \infty$	$c = c^*$ (see Eq 2.16)
Conc. (absent TFs)	$c = 0$	$c = 0$
Main contribution to crosstalk	Nonregulated genes	All genes
$X^*(t,s)$	$(1-t) + t^2 s/(1+ts)$	$t\left(-s(1-t) + 2\sqrt{s(1-t)}\right)$
Monotonic in t?	yes, decreasing	no, has a maximum
$X^*(t,s)$ for $s \ll 1$	$(1-t)$	$\sqrt{4t^2(1-t)}\sqrt{s} \quad \propto \quad \sqrt{s}$

Table 2.2: **Main features of local vs global minimization of crosstalk.**

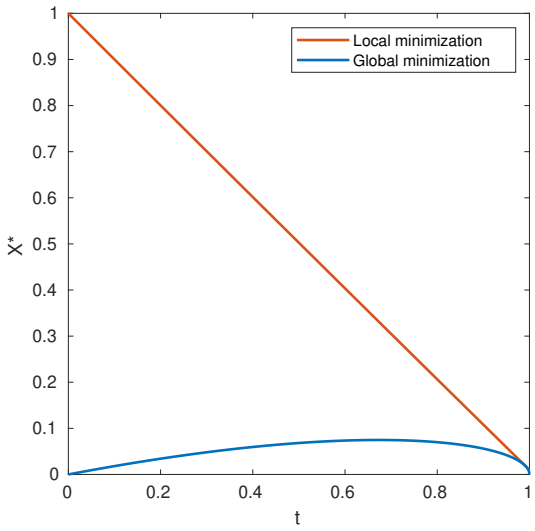

Figure 2.18: **Comparison of total crosstalk for minimization of individual genes** (blue) and global crosstalk minimization (red) as obtained in Eq 2.5. We find that local minimization requires extremely high TF concentrations, which are likely to cause crosstalk to all genes that should be left unregulated. Hence total crosstalk (local minimization) is approximately $X \approx 1 - t$ (the proportion of available TF species). In contrast, global minimization does not optimize crosstalk for every individual genes, but provides much lower total crosstalk values. Similarity $s = 10^{-2}$.

A TF which is both an activator and a repressor

In this section we explore the case that every TF has a dual role: it serves as an activator for one gene and as a repressor for another gene. This means that the genes

are now organized in pairs, such that each pair of genes shares a common TF where one of the genes is positively regulated and the other is negatively regulated by that TF. The total number of TFs is then half of their number in the basic model, in which every gene was regulated by a unique TF. That however, also constrains the ability to individually determine the regulatory state of each gene, in contrast to the basic model case. We begin by examining the two extreme scenarios:

1. If the two genes in every pair are always in opposite regulatory states, namely when one needs to be active the other one needs to be inactive, both require that the TF is present (absent) at the same time. We call this the *non-conflict* scenario. Then the number of regulated genes is twice the number of TFs in use in the basic model. The expressions for single-gene crosstalk and total crosstalk probabilities (Eq 2.13, Eq 2.4) are then slightly modified:

$$x_{\text{bound}} = \frac{e^{-Ea} + cs/2}{c/t + e^{-Ea} + cs/2}, \tag{2.40}$$

$$x_{\text{unbound}} = \frac{cs/2}{e^{-Ea} + cs/2}, \tag{2.41}$$

$$X = t \cdot x_{\text{bound}} + (1-t) \cdot x_{\text{unbound}}, \tag{2.42}$$

   where the only difference with respect to Eq 2.13 is the factor two in $cs/2$. The factor two is present as the rescaled similarity s is rescaled by half of the total TF species compared to the basic model. We have already shown [Friedlander *et al.*, 2016] (SI, p. 12) that the case that each TF regulates Θ genes is equivalent to an effective similarity scaling $s \rightarrow s/\Theta$. As to first order in s, $X^* \sim \sqrt{s}$ (Eq 2.5), this yields an effective reduction of crosstalk by a factor of $\sqrt{2}$.

2. Alternatively, if the two genes in each pair always need to be in the same regulatory state, one of them should be regulated by its cognate TF and the other should be left with unoccupied binding site. This means that only one of the genes requires the TF to be present, but the other one favors its absence to reduce crosstalk. Here we assume that if at least one gene requires the TF, then the TF is present. We call this the *conflict* scenario.

The single-gene and total crosstalk probabilities now read:

$$x_{\text{bound}} = \frac{e^{-Ea} + cs/2}{c/t + e^{-Ea} + cs/2}, \tag{2.43}$$

$$x_{\text{unbound}} = \frac{cs/2 + c/t}{c/t + e^{-Ea} + cs/2}, \tag{2.44}$$

$$X = \frac{t}{2} \cdot x_{\text{bound}} + (1 - \frac{t}{2}) \cdot x_{\text{unbound}}. \tag{2.45}$$

The change with respect to Eq 2.13 is the additional term c/t in the x_{unbound} expression as the cognate TF is present, even though it is not required. We define binding of cognate TF when not required as crosstalk. Furthermore, the weights in crosstalk X represent the proportion of regulated (unregulated) genes, i.e., $t/2$ and $(1 - t/2)$, respectively. As before, $cs/2$ is also adjusted.

As exactly one of the two genes should be regulated, for all pairs of genes, half of the genes are regulated. That represent all TF speces, i.e., $t = 1$. We then obtain that the lower bound on crosstalk in this scenario is $X^* = 0.5$.

3. Taking any general combination of the two extreme scenarios, with p proportion of conflict genes and $1 - p$ non-conflict genes the crosstalk probabilities read:

$$x^a_{\text{bound}} = \frac{e^{-Ea} + cs/2}{c/t + e^{-Ea} + cs/2}, \tag{2.46}$$

$$x^a_{\text{unbound}} = \frac{cs/2}{c/t + e^{-Ea} + cs/2}, \tag{2.47}$$

$$x^b_{\text{bound}} = \frac{e^{-Ea} + cs/2}{c/t + e^{-Ea} + cs/2}, \tag{2.48}$$

$$x^b_{\text{unbound}} = \frac{cs/2 + c/t}{c/t + e^{-Ea} + cs/2}, \tag{2.49}$$

$$\begin{aligned} X = (1-p) \cdot \left(t_G \cdot x^a_{\text{bound}} + (1 - t_G) \cdot x^a_{\text{unbound}}\right) \\ + p \cdot \left(t_G \cdot x^b_{\text{bound}} + (1 - t_G) \cdot x^b_{\text{unbound}}\right) \end{aligned} \tag{2.50}$$

where $t_{\text{G}} = \left(p/2 + (1-p)t\right)$ is the proportion of genes that are regulated, and a and b denote contributions of non-conflict and conflict genes, respectively. p constrains the possible t values to be in the range $t \in (p, 1)$. As mentioned, in the limit of $p = 1$, all gene pairs are opposite in the regulation demand,

leading to $t = 1$. In the opposite scenario, for $p = 0$, any value of $0 \leq t \leq 1$ is possible. For example, for $p = 1/2$, the system can explore values in the range $t \in (1/2, 1)$, as only half of the system is fully constrained.

Fig. 2.19 shows the total minimal crosstalk X^* for various p values, and the basic model as a reference. We find a decrease in crosstalk if all (or nearly all) genes are in non-conflict pairs. However, for conflict pairs, there is a much higher crosstalk now because of the presence of the cognate TF for genes that should be left unregulated.

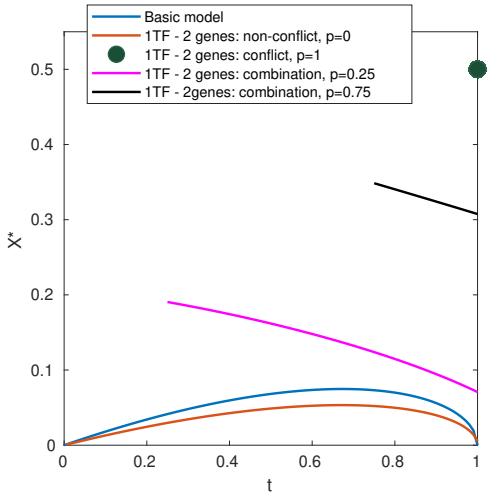

Figure 2.19: **Minimal crosstalk when every TF regulates two genes: one as an activator and the other as a repressor couples between the regulatory states of the genes sharing a common TF.** When both genes in all pairs require their TF to be either present or absent (non-conflict), crosstalk is lowered (red curve) compared to the basic model where every TF regulates only one gene (blue). In contrast, if all paired genes have opposite demands for the TF presence, crosstalk is considerably larger (green dot). We also illustrate combinations of these two extremes (magenta for p=0.25 and black for p=0.75), both still showing much higher crosstalk compared to the basic model. Similarity $s = 10^{-2}$.

Combinatorial regulation

So far we studied regulatory architectures in which each gene is regulated by a single TF (although each TF could regulate multiple genes, as in the previous section). There are however known cases in which a particular gene is regulated by a combination of distinct TF species. We study a model for such an architecture in

this section. We assume that every gene is regulated by two distinct and unique TF species having separate and non-overlapping binding sites. Binding to one site does not affect binding to the other in any way: neither inhibits it nor makes it more favorable by any form of cooperativity. We assume an AND-gate logic, such that only if both binding sites are bound by the cognate factors the gene is regulated, but not if only one of them (or neither) is bound and the other is unoccupied or bound by a non-cognate.

As opposed to the basic model where energy levels referred to a single binding site, we now refer to the energy levels of a pair of binding sites regulating a common gene. They can now be different binding states with the following statistical weights:

- $w_1 = e^{-2E_a}$ if both binding sites are unoccupied,
- $w_2 = e^{-E_a} \cdot c \cdot s$ if one binding site is unoccupied while the second one is occupied by non-cognate TF.
- $w_3 = (c \cdot s)^2$ if both binding sites are occupied by non-cognate TFs.
- $w_4 = e^{-E_a} \cdot \frac{c}{t}$ if one binding site is unoccupied while the second one is occupied by the cognate TF.
- $w_5 = \left(\frac{c}{t}\right)^2$ if both binding sites are occupied by the cognate TFs.
- $w_6 = \left(\frac{c}{t}\right) \cdot (c \cdot s)$ if both binding sites are occupied, one by the cognate and the other by a non-cognate TF.

The single gene crosstalk probabilities now read:

$$x_{\text{bound}} = \frac{w_1 + 2w_2 + 2w_4 + w_3 + 2w_6}{w_1 + 2w_2 + 2w_4 + w_3 + w_5 + 2w_6} = \frac{e^{-2Ea} + 2e^{-Ea}\,(cs + c/t) + (cs)^2 + 2c^2 s/t}{e^{-2Ea} + 2e^{-Ea}\,(cs + c/t) + (cs)^2 + (c/t)^2 + 2c^2 s/t}, \tag{2.51}$$

$$x_{\text{unbound}} = \frac{w_3}{w_1 + 2w_2 + w_3} = \frac{(cs)^2}{e^{-2Ea} + 2e^{-Ea} cs + (cs)^2}, \tag{2.52}$$

$$X = t\; x_{\text{bound}} + (1 - t)\; x_{\text{unbound}} \tag{2.53}$$

w_2, w_4, and w_6 have the pre-factor 2 because they can apply to either of the two binding sites.

In Fig 2.20 we illustrate the minimal crosstalk in this case, compared to the basic model. We find that the AND-gate configuration leads to lower minimal crosstalk for low t, but higher crosstalk for high t, compared to the basic model. The difference between these models comes from the double number of TFs used in the AND-gate, which lead to higher crosstalk for high t, but also from the stricter definition of what is considered crosstalk for genes that should be left unregulated. To demonstrate this, we can define a more lenient definition of x_{unbound}:

$$x_{\text{unbound}}^{\text{lenient}} = \frac{2w_2 + w_3}{w_1 + 2w_2 + w_3} = \frac{(cs)^2 + 2e^{-Ea}cs}{e^{-2Ea} + 2e^{-Ea}cs + (cs)^2}, \tag{2.54}$$

$$X^{\text{lenient}} = t\ x_{\text{bound}} + (1-t)\ x_{\text{unbound}}^{\text{lenient}}, \tag{2.55}$$

where a state that should be unbound and is only partially occupied by one non-cognate TF, is already considered crosstalk. This leads to an elevation of X^* for all t, compared to the basic model.

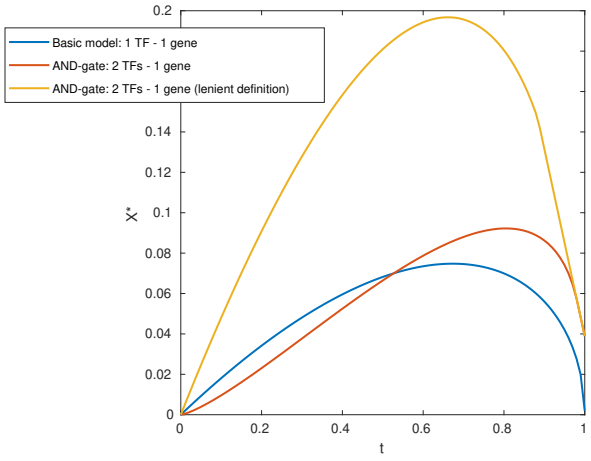

Figure 2.20: **Minimal crosstalk for combinatorial regulation**, where each gene is regulated by an AND-gate with 2 distinct TF species (red and yellow, stricter and more lenient definition of crosstalk, respectively), compared to the basic model, where each gene is regulated by a single TF (blue). Similarity $s = 10^{-2}$.

3 Normative models of enhancer function

In prokaryotes, thermodynamic models of gene regulation provide a highly quantitative mapping from promoter sequences to gene expression levels that is compatible with *in vivo* and *in vitro* biophysical measurements. Such concordance has not been achieved for models of enhancer function in eukaryotes. In equilibrium models, it is difficult to reconcile the reported short transcription factor (TF) residence times on the DNA with the high specificity of regulation. In non-equilibrium models, progress is difficult due to an explosion in the number of parameters. Here, we navigate this complexity by looking for minimal non-equilibrium enhancer models that yield desired regulatory phenotypes: low TF residence time, high specificity and tunable cooperativity. We find that a single extra parameter, interpretable as the "linking rate" by which bound TFs interact with Mediator components, enables our models to escape equilibrium bounds and access optimal regulatory phenotypes, while remaining consistent with the reported phenomenology and simple enough to be inferred from upcoming experiments. We further find that high specificity in non-equilibrium models is in a tradeoff with gene expression noise, predicting bursty dynamics — an experimentally-observed hallmark of eukaryotic transcription. By drastically reducing the vast parameter space to a much smaller subspace that optimally realizes biological function prior to inference from data, our normative approach holds promise for mathematical models in systems biology.

Published as **Grah R**, Zoller B, Tkačik G. Normative models of enhancer function. *bioRxiv*. 2020. doi: 10.1101/2020.04.08.029405.

Contributions: Grah R has computed results on regulatory phenotypes, has done stochastic simulations, and has computed analytical limits. B Zoller has done calculations and derivations on residence time distributions, noise propagation, correlation time, and helped with the optimization of the algorithm.

Some changes have been made to the text in order to integrate it into this thesis.

3.1 Introduction

An essential step in the control of eukaryotic gene expression is the interaction between transcription factors (TFs), various necessary co-factors, and TF binding sites (BSs) on the regulatory segments of DNA known as enhancers [Coulon *et al.*, 2013]. While we are far from having either a complete parts list for this extraordinarily complex regulatory machine or an insight into the dynamical interactions between its components, experimental observations have established a number of constraints on its operation: *(i)* TFs individually only recognize short, 6–10bp long binding site motifs [Wunderlich and Mirny, 2009]; *(ii)* TF residence times on the cognate binding sites can be as short as a few seconds and only 2–3 orders of magnitude longer than residence times on non-specific DNA [Gebhardt *et al.*, 2013; Chen *et al.*, 2014; Thomas *et al.*, 2019]; *(iii)* the order of arrival of TFs to their binding sites can affect gene activation [Chen *et al.*, 2014]; *(iv)* TFs do not activate transcription by RNA polymerase directly, but interact first with various co-activators, essential amongst which is the Mediator complex; *(v)* binding of multiple TFs is typically required within the same enhancer for its activation [Shlyueva *et al.*, 2014], which can lead to very precise downstream gene expression only in the presence of a specific combination of TF concentrations [Petkova *et al.*, 2019]; *(vi)* when activated, gene expression can be highly stochastic and bursty [Nicolas *et al.*, 2018; Molina *et al.*, 2013; Bartman *et al.*, 2016]; *(vii)* gene induction curves show varying degrees of steepness, suggesting tunable amounts of cooperativity among TFs [Park *et al.*, 2019]. Here we look for biophysical models of enhancer function consistent with these observations.

Mathematical modeling of gene regulation traces its origins to the paradigmatic examples of the λ bacteriophage switch [Ptashne, 1986] and the *lac* operon [Kuhlman *et al.*, 2007]. In prokaryotes, biophysical models have proven very successful [Berg and von Hippel, 1987; Kinney *et al.*, 2010; Belliveau *et al.*, 2018], assuming gene expression to be proportional to the fraction of time RNA polymerase is bound to the promoter in thermodynamic equilibrium; TFs modulate this fraction via steric or energetic interactions with the polymerase. Crucially, these models are very compact: they are fully specified by enumerating all bound configurations and energies of the TFs and

the polymerase on the promoter. While some open questions remain [Garcia *et al.*, 2012; Hammar *et al.*, 2014; Forcier *et al.*, 2018], the thermodynamic framework has provided a quantitative explanation for combinatorial regulation, cooperativity, and regulation by DNA looping [Bintu *et al.*, 2005a; Bintu *et al.*, 2005b], while remaining consistent with experiments that also probe the kinetic rates [Maerkl and Quake, 2007; Jones *et al.*, 2014].

No such consensus framework exists for eukaryotic transcriptional control. Limited specificity of individual TFs *(i)* is hard to reconcile with the high specificity of regulation *(v)* and the suppression of regulatory crosstalk [Friedlander *et al.*, 2016], suggesting non-equilibrium kinetic-proofreading schemes [Cepeda-Humerez *et al.*, 2015]. Likewise, short TF residence times *(ii)* and the importance of TF arrival ordering *(iii)* contradict the conceptual picture where stable enhanceosomes are assembled in equilibrium [Chen *et al.*, 2014]. Kinetic schemes may be required to match the reported characteristics of bursty gene expression *(vi)* [Donovan *et al.*, 2019], or realize high cooperativity *(vii)* [Estrada *et al.*, 2016]. Thermodynamic models undisputedly have statistical power to predict expression from regulatory sequence even in eukaryotes [Gertz *et al.*, 2009], yet this does not resolve their biophysical inconsistencies or rule out non-equilibrium models. Unfortunately, mechanistically detailed non-equilibrium models entail an explosion in the complexity of the corresponding reaction schemes and the number of associated parameters: on the one hand, such models are intractable to infer from data, while on the other, it is difficult to understand which details are essential for the emergence of regulatory function.

To deal with this complexity, we systematically simplify the space of enhancer models. We adopt the normative approach, commonly encountered in the applications of optimality ideas in neuroscience and elsewhere [Tkačik and Walczak, 2011; Rieckh and Tkačik, 2014; Tkačik and Bialek, 2016]: we theoretically identify those models for which various performance measures of gene regulation, which we call "regulatory phenotypes", are maximized. Such optimal model classes are our candidates that could subsequently be refined for particular biological systems and confronted with data. Thus, rather than inferring a single model from experimen-

tal data or constructing a complex, molecularly-detailed model for some specific enhancer, we find the simplest generalizations of the classic equilibrium regulatory schemes, such as Hill-type [Phillips *et al.*, 2012] or Monod-Wyman-Changeux regulation [Mirny, 2010; Walczak *et al.*, 2010; Changeux, 2012], to non-equilibrium processes, which drastically improves their regulatory performance while leaving the models simple to analyze, simulate, and fit to data.

3.2 Results

Model. Multiple lines of evidence suggest that eukaryotic transcription is a two-state process which switches between active (`ON`) and inactive (`OFF`) states, with rates dependent on the transcription factor (TF) concentrations [Larson *et al.*, 2013; Senecal *et al.*, 2014; Zoller *et al.*, 2018]. We sought to generalize classic regulatory schemes that can describe the balance between `ON` and `OFF` transcriptional states in equilibrium: a Hill-like scheme of "thermodynamic models" (discussed in SI Section 1.3), and a Monod-Wyman-Changeux-like (MWC) scheme introduced below.

Figure 3.1A shows a schematic of the proposed functional enhancer model (Section 3.4.1, see also Fig 3.10). A complex of transcriptional co-factors that we refer to as a "Mediator"[1] can interact with TFs that bind and unbind from their DNA binding sites with baseline rates k_+ and k_- (Fig 3.1B.i). Mediator – and thus the whole enhancer – can switch between its functional `ON`/`OFF` states with baseline rates κ_+ and κ_- (Fig 3.1B.ii). Enhancer `ON` state and TF bound state are both stabilized (by a factor α relative to baseline rates) when a bound TF establishes a "link" with the Mediator (Fig 3.1B.iii). The molecular identity of such links can remain unspecified: it could, for example, correspond to an enzymatic creation of chemical marks (e.g., methylation, phosphorylation) on the TFs or Mediator proteins conditional on their physical proximity or interaction. Crucially, the links can be established and removed in processes that can break detailed balance and are thus out of equilibrium. Here, we consider that a link is established at a rate k_{link} between a bound TF and the

[1] Our nomenclature is simply a shorthand for all co-factors necessary for eukaryotic transcriptional activation at an enhancer, which can include proteins not strictly a part of the Mediator family.

Mediator complex; for simplicity, we assume that the links break when the TFs dissociate or upon the switch into OFF state (this assumption can be relaxed, see Fig 3.11).

An important thrust of our investigations will concern the role of limited specificity of individual TFs to recognize their cognate sequences on the DNA. If sequence specificity arises primarily through TF binding – a strong, but relatively unchallenged assumption (that can also be relaxed within our framework, see Fig 3.12) – then we should ask how likely it is for the Mediator complex to form and activate at specific sites contained within functional enhancers (with low off-rates characteristic of strong eukaryotic TF binding sites, k_-^{S}) versus at random, non-specific sites on the DNA (with ~ 2 orders-of-magnitude higher individual TF off-rates, k_-^{NS}) from which expression should not occur.

Given the number of TF binding sites (n) and the various rate parameters ($k_+, k_-^{\mathrm{S/NS}}, \kappa_+, \kappa_-, \alpha, k_{\mathrm{link}}$) the full state of the system—i.e., the probability to observe any number of bound and/or linked TFs jointly with the ON/OFF state of the enhancer—evolves according to a Chemical Master Equation (SI Section 1.1) that can be solved exactly [Sanchez and Kondev, 2008; Lestas *et al.*, 2008; Walczak *et al.*, 2012] or simulated using the Stochastic Simulation Algorithm [Gillespie, 2007]. Importantly, we show analytically that our scheme reduces to the true equilibrium MWC model in the limit $k_{\mathrm{link}} \to \infty$: in this limit, there can be no distinction between a bound TF and a TF that is both bound and linked, and one can define a free energy F that governs the probability of enhancer being ON, which in our model is equal to (a normalized) mean expression level, $E = P_{\mathtt{ON}} = \left(1 + \exp(F)\right)^{-1}$, with

$$F = n \log \frac{1 + c/K}{1 + \alpha \cdot c/K} - L, \tag{3.1}$$

where $K = k_-/k_+^0$, $k_+ = k_+^0 c$ (see also Fig 3.1 caption), and $L = \log(\kappa_+/\kappa_-)$. The k_{link} parameter thus interpolates between the equilibrium limit in Eq (3.1), corresponding to a textbook MWC model, and various non-equilibrium (kinetic) schemes which we will explore next. A similar generalization with an equilibrium limit exists for thermodynamic Hill-type models, where, furthermore, α can be directly identified with cooperativity between DNA-bound TFs (see SI Section 1.3); we will see that

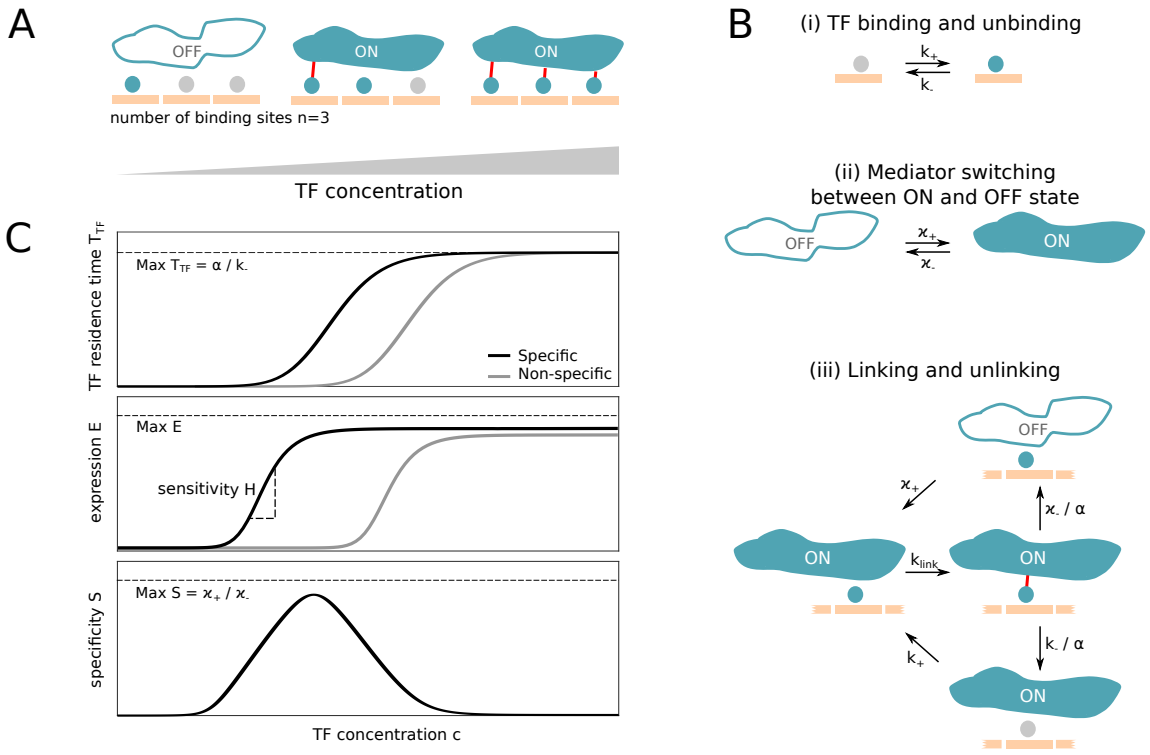

Figure 3.1: **A non-equilibrium MWC-like model of enhancer function. (A)** Schematic representation of transcription factors (TFs; tael circles) interacting with binding sites (BSs, here $n = 3$ orange slots) and the putative Mediator complex via links (red lines). The Mediator complex can be in two conformational states (`OFF` or `ON`), with the `ON` state enabling productive transcription of the regulated gene. Increasing TF concentration, c, facilitates TF binding and the switch into `ON` state (left-to-right). **(B)** Key reactions and rates of the non-equilibrium model. TFs can bind with concentration-dependent on rate ($k_+ = k_+^0 c$) and unbind with basal rate k_- that is in principle sequence dependent (i). The Mediator state switches between the conformational states with basal rates κ_+ and κ_- (ii). Linking and unlinking of TFs to Mediator (iii) can move the system out of equilibrium: links are established with rate k_{link}, and the link stabilizes both TF residence and the `ON` state of the Mediator by a factor α per established link. **(C)** Regulatory phenotypes. Mean TF residence time, T_{TF}, on specific sites in functional enhancers (black) vs random site on the DNA (gray) increases with concentration (top), as does mean expression, E (the fraction of time the Mediator is `ON`; induction curve, middle, with sensitivity, H, defined at mid-point expression). Specificity, S, is defined as the ratio of expression from the specific sites in the enhancer relative to the expression from random piece of DNA.

this qualitative role of α will hold also for the MWC case.

Phenotype	Symbol	Value	Ref
TF residence time (specific BS)	T_{TF}	$\sim 1 - 10$ s	[Gebhardt *et al.*, 2013; Morisaki *et al.*, 2014]
Expression (fraction of time `ON`)	E	$0.01 - 0.9$	[Zenklusen *et al.*, 2008; Suter *et al.*, 2011; Zoller *et al.*, 2018]
Sensitivity (apparent Hill coef.)	H	$1 - 10$	[Park *et al.*, 2019]
Specificity	S	—	—
Noise (std / mean protein exp.)	N	$\sim 0.1 - 1$	[Zoller *et al.*, 2015]

Table 3.1: **Regulatory phenotypes.**

Regulatory phenotypes. How does the regulatory performance depend on the enhancer parameters and, in particular, on moving away from the equilibrium limit? To assess this question systematically, we define a number of "regulatory phenotypes", enumerated in Table 3.1 and illustrated in Fig 3.1C. As a function of TF concentration, we compute: **(i)** individual TF residence time, T_{TF}, on specific sites in functional enhancers, as well as on random, non-specific DNA, because these quantities have been experimentally reported in single-molecule experiments and provide strong constraints on enhancer function; **(ii)** average expression, E, for functional enhancers as well as random, non-specific DNA; we require E to be in the middle (~ 0.5) of the wide range reported for functional enhancers; **(iii)** sensitivity of the induction curve at half-maximal induction, H, an observable quantity often interpreted as a signature of cooperativity in equilibrium models; **(iv)** specificity, S, as the ratio between expression E from functional enhancers vs from non-specific DNA, which should be as high as possible to prevent deleterious crosstalk or uncontrolled expression [Friedlander *et al.*, 2016]; **(v)** expression noise, N, defined more precisely later, originating in stochastic enhancer `ON/OFF` switching.[2]

Specificity, residence time, and expression. Figure 3.2A explores the relationship between three regulatory phenotypes for a MWC-like enhancer scheme of Fig 3.1A: the average TF residence time (T_{TF}), specificity (S), and the average expression (E),

[2]Protein noise levels in Table 3.1 are estimated from reported mRNA noise levels.

at fixed concentration c_0 of the TFs. Each point in this "phase diagram" corresponds to a particular enhancer model; points are accessible by varying α and k_{link} (Fig 3.2B) and fall into a compact region that is bounded by intuitive, analytically-derivable limits to specificity and the residence time. As α tends to large values, S approaches 1, as it must: once a TF-Mediator complex forms, large α will ensure it never dissociates and expression E will tend to 1 (see also Fig 3.2D) irrespective of whether this occurred on a functional enhancer or a random piece of DNA – in this limit, all sequence discrimination ability is lost, yielding undesirable regulatory phenotypes. In contrast, the equilibrium ("EQ") MWC limit as $k_{\text{link}} \to \infty$ (Eq 3.1) is functional and, interestingly, corresponds to a non-monotonic curve in the phase diagram that lower-bounds the specificity of non-equilibrium ("NEQ") models accessible at finite values of k_{link}.

In a wide intermediate range of TF residence times, the full space of nonequilibrium MWC-like models—which we can exhaustively explore—offers large, orders-of-magnitude improvements in specificity, essentially utilizing a stochastic variant of Hopfield's proofreading mechanism [Hopfield, 1974; Cepeda-Humerez *et al.*, 2015]. This observation is generic, even though the precise values of S depend on parameters that we explore below, and S always remains bounded from above by κ_-/κ_+ (in equilibrium, this is related to stochastic, thermal-fluctuation-driven Mediator transitions to `ON` state even in absence of bound TFs). At the same average TF residence time and TF concentration, the best non-equilibrium model (`II` in Fig 3.2) will suppress expression from non-cognate DNA by almost two orders-of-magnitude relative to the best equilibrium model (`I`). These findings remain qualitatively unchanged for enhancers with larger number of binding sites (see Fig 3.13).

A comparison of various enhancer operating regimes is perhaps biologically more relevant at fixed mean expression, allowing the TF concentration to adjust accordingly under cells' own control, as shown in Fig 3.2C for $E = 0.5$. As TF residence time lengthens with increasing α, TFs and the Mediator establish more stable complexes on the DNA and lower concentrations are needed for all models to reach the desired expression E (see also Fig 3.2D). Nevertheless, the ability of α to increase the

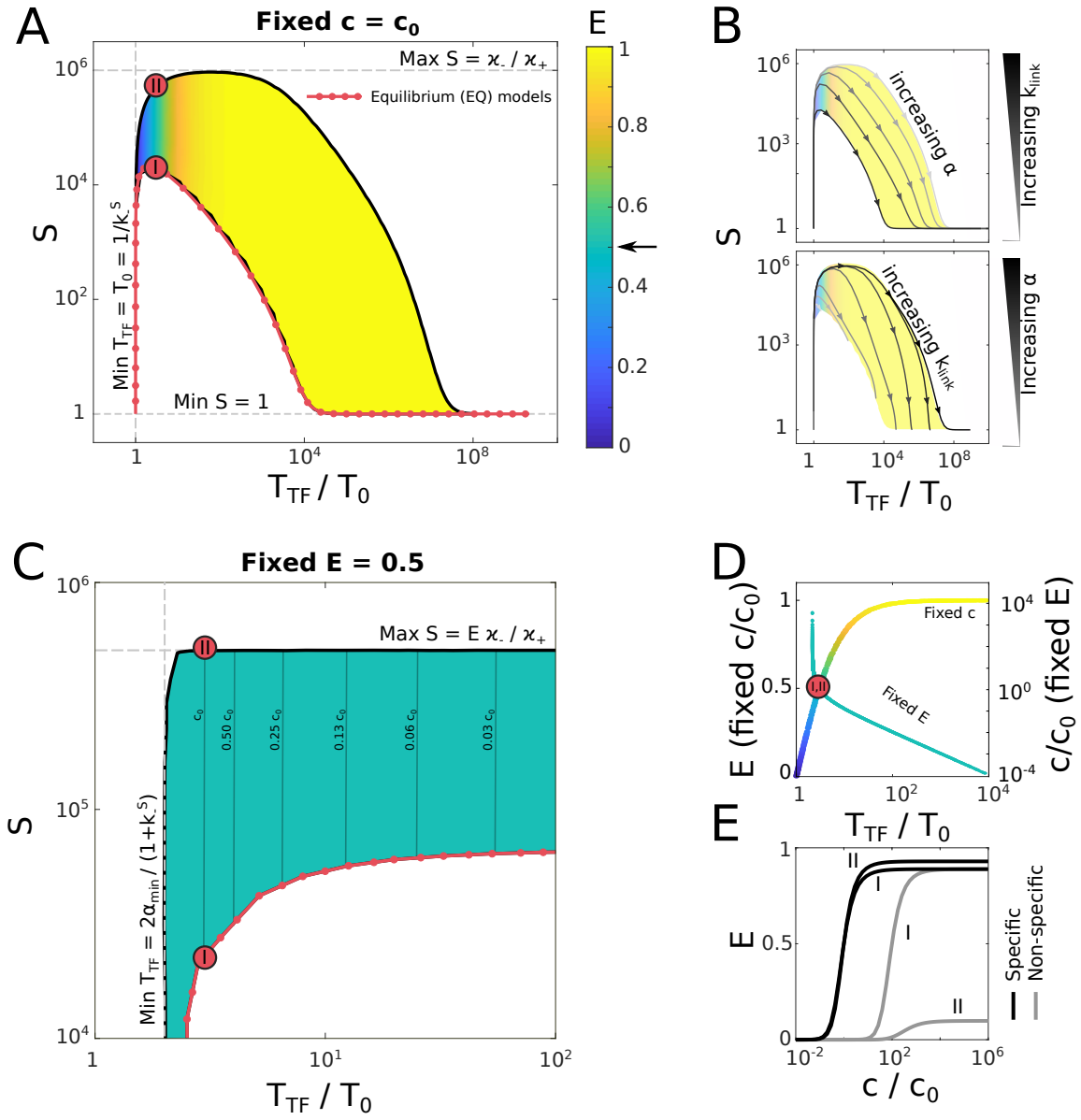

Figure 3.2: (Continued on the following page.)

specificity in equilibrium models is limited and saturates at a value substantially below the specificity reachable in nonequilibrium models at much smaller TF residence times. The observations of Fig 3.2A, C underscore an important, yet often overlooked, point: the ability to induce at low TF concentration (that is, high affinity) achieved through "cooperative interactions" at high α either has a detrimental, or, at best, a marginally beneficial effect for the ability to discriminate between cognate and random DNA sites (that is, high specificity) in equilibrium [Friedlander *et al.*, 2016].

Figure 3.2: **Accessible space of regulatory phenotypes. (A)** Specificity, S, mean TF residence time, T_{TF} (expressed in units in inverse off-rate for isolated TFs at their specific sites, $T_0 = 1/k_-^{\mathrm{S}}$), and average expression, E (color), for MWC-like models with $n = 3$ TF binding sites, obtained by varying α and k_{link} at fixed TF concentration, c_0. Equilibrium models fall onto the red line; two models with equal TF residence times, `I` (EQ) and `II` (NEQ), are marked for comparison. Dashed gray lines show analytically-derived bounds. **(B)** Phase space of regulatory phenotypes is accessed by varying α at fixed values of k_{link} (grayscale; top) or varying k_{link} at fixed values of α (grayscale; bottom). **(C)** As in (A), but the TF concentration at each point in the phase space is adjusted to hold average expression fixed at $E = 0.5$ (green color). Plotted is a smaller region of phase space of interest; nearly vertical thin lines are equi-concentration contours (Fig 3.15). **(D)** All models in the phase diagrams in (A) and (C) approximately collapse onto nearly one-dimensional manifolds ("fixed c", left axis, for (A); "fixed E", right axis, for (C)) when plotted as a function of mean TF residence time, T_{TF}, supporting the choice of this variable as a biologically-relevant observable. Color on the manifold corresponds to mean expression E using the colormap of (A). Vertical scales are chosen so that models `I` and `II` coincide. **(E)** Induction curves of equilibrium model `I` and non-equilibrium model `II` for expression from functional enhancer that contains specific sites (basal TF off-rate k_-^{S}; black curves) versus expression from random DNA containing non-specific sites (basal TF off-rate $k_-^{\mathrm{NS}} = 10^2 k_-^{\mathrm{S}}$ here; gray curves).

Figure 3.2E shows induction curves for expression from functional enhancers containing specific sites and from random DNA sites, for equilibrium (`I`) and non-equilibrium (`II`) models. Both yield essentially indistinguishable induction curves for expression from a functional enhancer (which is true generically across our phase diagram, see Fig 3.14), suggesting that it would be difficult to discriminate between the models based on induction curve measurements. In sharp contrast, the behavior of the two models is qualitatively different at non-specific DNA: with sufficiently high TF concentration (e.g., in an over-expression experiment), the EQ model `I` will fully induce even from random DNA as its binding sites get saturated by TFs; on the contrary, the nonequilibrium (NEQ) model `II` will start inducing at much higher c, and will never do so fully due to its proofreading capability. Thus, given the relatively weak individual TF preference for cognate vs non-cognate DNA, one should look at the collective response of the gene expression machinery to mutated or random enhancer sequences for signatures of equilibrium vs non-equilibrium proofreading behavior.

Sensitivity. Intuitively, sensitivity H measures the "steepness" of the induction curve. More precisely, H is proportional to the logarithmic derivative of the expression with log concentration at the point of half-maximal expression, so that for Hill-like functions, $E(c) = c^h/(c^h + K^h)$, it corresponds exactly to the Hill coefficient, $H = h$. Figure 3.3A shows that H increases monotonically with T_{TF} (and thus with α, cf. Fig 3.2B), indicating that more stable TF-Mediator complexes indeed lead to higher apparent cooperativity, which is always upper-bounded by the number of TF binding sites in the enhancer, n. The highly-cooperative "enhanceosome" concept [Arnosti and Kulkarni, 2005] would, in our framework, correspond to an equilibrium limit with very high α, and thus $H \sim n$; yet the analysis above predicts vanishingly small specificity increases as this limit is approached. In contrast, we observe that the point at which the specificity advantage of nonequilibrium models is maximized, i.e., where $S_{\mathrm{NEQ}}/S_{\mathrm{EQ}}$ is largest, occurs far away from $H = n$, at much lower H values (Fig 3.17). If high specificity is biologically favored, we should therefore not expect the "number of known binding sites" to equal the "measured Hill coefficient of the induction curve" for well-functioning eukaryotic transcriptional schemes, even on theoretical grounds.

Noise. Lastly, we turn our attention to gene expression noise. All stochastic two-state models have a steady state binomial variance of $\sigma_E^2 = E(1 - E)$ in enhancer state, where E is the probability of the enhancer to be `ON`. When `ON`, transcripts are made and subsequently translated into protein, which typically has a slow lifetime, T_P, on the order of at least a few hours. Random fluctuations in enhancer state will cause random steady-state fluctuations in protein copy number around the average, P; these fluctuations can be quantified by noise, $N = \sigma_P/P$. While there can be other contributions to noise (e.g., birth-death fluctuations due to protein production and degradation), we focus here solely on the effects of `ON/OFF` switching, since only these effects depend on the enhancer architecture [Rieckh and Tkačik, 2014].

How is noise in gene expression, N, related to the binomial variance, σ_E? Based on simple noise propagation arguments [Paulsson, 2004; Tkačik *et al.*, 2008], fractional variance in protein should be equal to fractional variance in enhancer state times

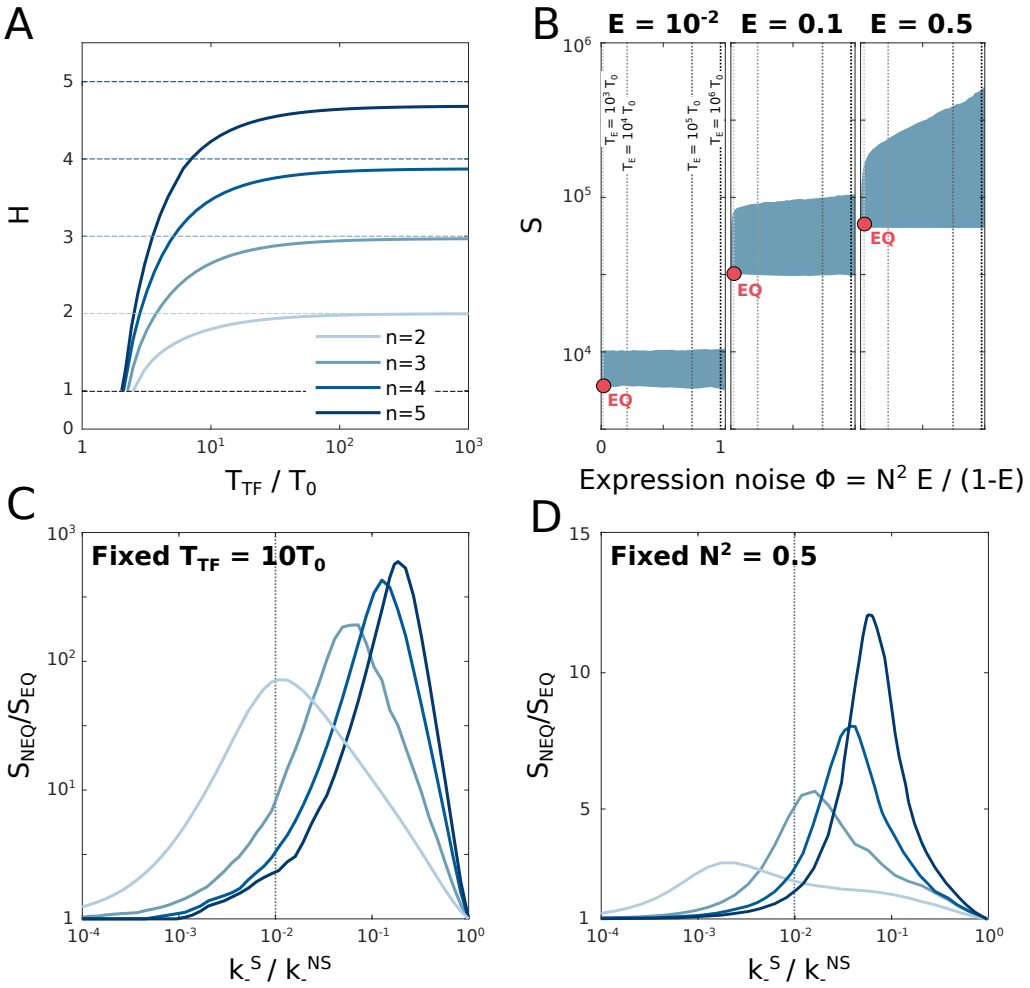

Figure 3.3: **Limits to sensitivity and specificity. (A)** Sensitivity (apparent Hill coefficient) H of enhancer models in the phase diagram of Fig 3.2C, at fixed mean expression, $E = 0.5$. All models collapse onto the manifolds shown for different number of TF binding sites, n. **(B)** Phase diagram of enhancer models for three different values of mean expression, E (columns), shows specificity S and fraction of variance in enhancer switching propagated to expression noise (see text). Compact blue region for each E shows all MWC-like models with $n = 3$ binding sites accessible by varying α and k_{link}; equilibrium model ("EQ") with lowest noise is shown as a red dot. Increase in noise is monotonically related to increase in enhancer correlation time, T_E, marked with dashed vertical lines. Largest specificity increases over EQ models occur at high T_E and thus high noise (upper right corner of the blue region). **(C)** Maximal gain in enhancer specificity for non-equilibrium vs equilibrium models for different n (legend as in A), as a function of the intrinsic specificity of individual TF binding sites, k_-^S/k_-^{NS}. Expression is fixed to $E = 0.5$ and mean TF residence time to $T_{TF}/T_0 = 10$. Typical value $k_-^S/k_-^{NS} = 10^{-2}$ used in Fig 3.2 and panels A,B is shown in vertical dashed line. **(D)** Same as in (C), but with the comparison at fixed gene expression noise, $N^2 = 0.5$.

the noise filtering that depends on the timescales of enhancer switching, T_E, and protein lifetime, T_P (here we assume $T_P = 10$ hours), so that $N^2 = (\sigma_P/P)^2 \sim (\sigma_E/E)^2 \cdot T_E/(T_E + T_P)$ (see SI Section 1.5 for exact derivation). Thus, if enhancer switches much faster than the protein lifetime, $T_E \ll T_P$, protein dynamics almost entirely averages out the enhancer state fluctuations. Since all enhancer models have the same binomial variance, the gene expression noise in various models will be entirely determined by the mean expression, E, and the correlation time, T_E, both of which we can compute analytically for any combination of enhancer model parameters in the phase diagram of Fig 3.2.

Figure 3.3B shows the phase diagram of accessible MWC-like regulatory phenotypes for the specificity (S), mean expression (E) and fraction of enhancer switching noise that propagates to gene expression, $T_E/(T_E + T_P)$, found by varying α and k_{link}. As in Fig 3.2, equilibrium models ("EQ") have the lowest specificity S, but also lowest correlation time T_E and thus lowest noise, regardless of the average expression, E. There exist NEQ models that achieve higher specificity at a small increase in noise, but the highest specificity increases always come hand-in-hand with a substantial lengthening of the correlation times in enhancer state fluctuations, and thus with the inevitable increase in noise.

To better elucidate the tradeoffs and limits to specificity in non-equilibrium vs equilibrium models, we next explore how enhancer specificity gains depend on the ability of individual TFs to discriminate cognate binding sites from random DNA in Fig 3.3C. If individual TFs permit very strong discrimination ($k_-^{\text{S}}/k_-^{\text{NS}} < 10^{-4}$; prokaryotic TF regime), NEQ models at fixed individual TF residence times, T_{TF}, do not offer appreciable specificity increases in the collective enhancer response; in contrast, for the range around $k_-^{\text{S}}/k_-^{\text{NS}} \sim 10^{-2}$ typically reported for eukaryotic TFs, the specificity increase ranges from ten to thousand-fold, with the peak depending on the number of TF binding sites, n, as well as baseline Mediator specificity limit, κ_-/κ_+ (as this increases, the peak specificity gain is higher and moves towards lower $k_-^{\text{S}}/k_-^{\text{NS}}$, see Fig 3.18). If, instead of fixing $k_-^{\text{S}}/k_-^{\text{NS}} = 10^{-2}$ as we have done until now, we pick this ratio to maximize the specificity gain ($S_{\text{NEQ}}/S_{\text{EQ}}$) and again explore the noise-specificity tradeoff as in Fig 3.3B, we find that the extreme specificity gains are

only possible when correlation times, T_E diverge (see Fig 3.19), implying high noise.

These observations are summarized in Fig 3.3D, showing the specificity gain of NEQ models relative to EQ models, if the comparison is made at fixed noise level rather than at fixed individual TF residence time as in Fig 3.3C. Specificity gains are limited to roughly ten-fold even when, as we do here, we systematically search for best NEQ models through the complete phase diagram in Fig 3.2C. The specificity-noise tradeoff thus appears unavoidable.

Experimentally observable signatures of enhancer function. To illustrate how the proposed nonequilibrium (NEQ) MWC-like scheme could function in practice, we simulated it explicitly and compared it to an equilibrium (EQ) scheme with the same mean TF residence time in Fig 3.4. The two enhancers, composed of $n = 5$ TF binding sites, respond to a simulated protocol where the TF concentration is first switched from a minimal value that drives essentially no expression to a high value giving rise to $E = 0.5$, and after a long stationary period, the concentration is switched back to the low value. Figure 3.4A shows the occupancy of the binding sites and the functional `ON/OFF` state of the enhancer. Even though the two models share the same TF mean residence time and nearly indistinguishable induction curves (with $H \sim 2.7$), their collective behaviors are markedly different: the EQ scheme appears to have significantly faster TF binding / unbinding as well as Mediator switching dynamics, whereas NEQ scheme undergoes long, "bursty" periods of sustained enhancer activation and TF binding that are punctuated by `OFF` periods. If the typical residence time of an isolated TF on its specific site were $T_0 = 1$ s, NEQ enhancer could stay active even for hour-long periods ($\sim 10^4$ s), just somewhat shorter than the protein lifetime ($\sim 4 \cdot 10^4$ s). Such enhancer-associated stable mediator clusters are consistent with recent experimental reports [Chen *et al.*, 2018; Cho *et al.*, 2018].

The detailed steady-state behavior at high TF concentration is analyzed in Fig 3.4B. Consistent with our theoretical expectations, the NEQ scheme enables ten-fold higher specificity but at the cost of substantial noise in gene expression ($N \sim 0.42$) due to strong transcriptional bursting. High noise is a direct consequence of the

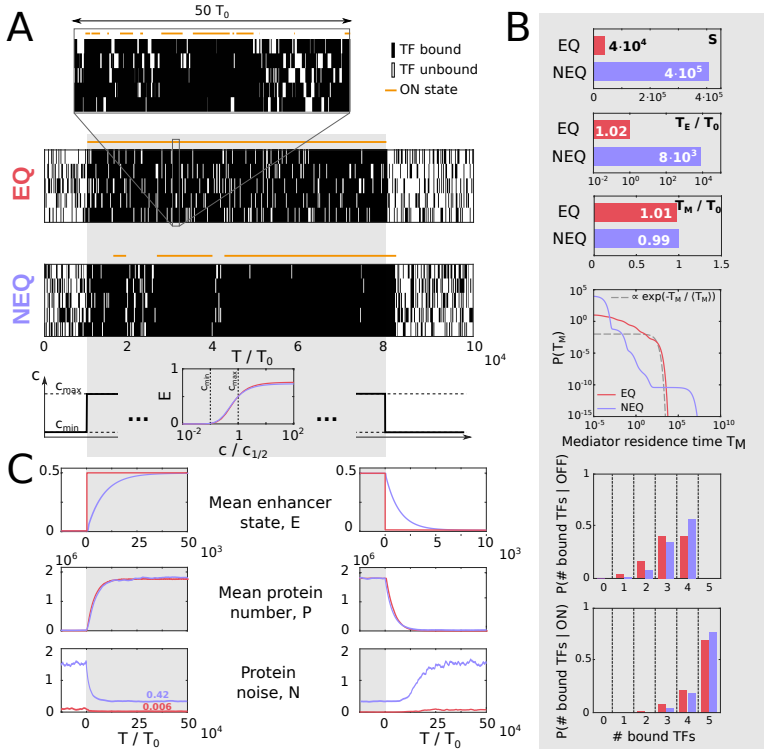

Figure 3.4: **High-specificity non-equilibrium schemes predict bursty gene expression. (A)** Stochastic simulation of an equilibrium (EQ) and a nonequilibrium (NEQ) enhancer model with $n = 5$ TF binding sites, responding to a TF concentration step (bottom-most panel). Average TF residence times are the matched between EQ and NEQ models at $2.1T_0$, $T_0 = 1/k_-^S = 1$ s, and both induction curves (scaled for half-maximal concentration) are identical, with sensitivity $H \approx 2.7$. When TF concentration is high, expression is fixed at $E = 0.5$. Parameters for NEQ model: $\alpha = 127$, $k_{\text{link}} = 2$, $c_{\text{max}} = 0.065$; for EQ model: $k_{\text{link}} \to \infty$, $\alpha = 19.8$,$c_{\text{max}} = 0.037$. Rasters show the occupancy of TF binding sites; orange line above shows the enhancer `ON/OFF` state; zoom-in for EQ model is necessary due to its fast dynamics. **(B)** Regulatory phenotypes for EQ and NEQ models during steady-state epoch (gray in A). Specificity (S) and enhancer state correlation time (T_E) are higher for the NEQ model; the Mediator mean `ON` residence time, T_M, is the same between the models, but the probability density function reveals a long tail in the NEQ scheme, and a nearly exponential distribution for the EQ scheme. Last two panels show the TF occupancy histogram during high TF concentration interval, conditional on the enhancer being `OFF` or `ON`. **(C)** Transient behavior of the mean enhancer state (E), mean protein number (P; assuming deterministic production/degradation protein dynamics given enhancer state), and gene expression noise, $N = \sigma_P/P$, for the NEQ and EQ models, upon a TF concentration low-to-high switch (left column) and high-to-low switch (right column). Traces shown are computed as averages over 1000 stochastic simulation replicates.

much longer correlation time of enhancer fluctuations, T_E, for the NEQ scheme, seen in Fig 3.4A. Interestingly, the mean residence time of the enhancer `ON` state, T_M,

is nearly unchanged between the EQ and NEQ scheme at ~ 100 s: but here, the mean turns to be a highly misleading statistic, as revealed by an in-depth exploration of the full probability density function. The NEQ scheme has a long tail of extended ON events interspersed with an excess of extremely short OFF events (due to high κ_- rate necessary for high specificity) relative to the EQ scheme (which, itself, does not deviate strongly from an exponential density function with a matched mean). The behavior of such an enhancer is highly cooperative even though the sensitivity (H) is not maximal: when the enhancer is ON, with very high probability all TFs are bound, and when OFF, often 4 out of 5 TFs are bound – yet the enhancer is not activated. In sum, a well-functioning non-equilibrium regulatory apparatus with its Mediator complex makes many short-lived attempts to switch ON, but only commits to a long, productive ON interval rarely and collectively, after insuring that activation is happening due to a sequence of valid molecular recognition events between several TFs and their cognate binding sites in a functional enhancer.

Transient behavior after a TF concentration change is analyzed in Fig 3.4C. The mean response time of the two models to the concentration change is governed by the correlation time of the enhancer state, T_E, and is thus much slower for NEQ vs EQ models; but since the protein lifetime is even longer, the mean protein levels adjust equally quickly in the equilibrium and nonequilibrium cases. This suggests that the dynamics of the mean protein level is unlikely to discriminate between EQ and NEQ models. In contrast, live imaging of the nascent mRNA could put constraints on T_E [Coulon *et al.*, 2013]. In that case, the filtering time scale is the elongation time, typically on the order of a few minutes, while the reported transcriptional response times—and thus estimates of T_E—would range from minutes to $1-2$ hours [Molina *et al.*, 2013; Donovan *et al.*, 2019].

Steady-state noise levels at high induction, as reported already, are considerably higher for the NEQ model due to transcriptional bursting; an intriguing further suggestion of our analyses is a long transient in the noise levels upon a high-to-low TF concentration switch, which finally settles to a high fractional noise level (here, $N \sim 1.6$) even at very low induction, due to sporadic transcriptional bursts.

3.3 Discussion

In this chapter, we took a normative approach to address the complexity of eukaryotic gene regulatory schemes. We proposed a minimal extension to a well-known Monod-Wyman-Changeux model that can be applied to the switching between the active and inactive states of an enhancer. The one-parameter extension is kinetic and accesses nonequilibrium system behaviors. We analyzed the parameter space of the resulting model and visualized the phase diagram of "regulatory phenotypes", quantities that are either experimentally constrained (such as mean expression, mean TF residence time, sensitivity), are likely to be optimized by evolutionary pressures (such as noise and specificity), or both. This allowed us to recognize and understand biophysical limits and trade-offs, and to identify the optimal operating regime of the proposed enhancer model that is consistent with current observations, as we summarize next.

Our analyses suggest the following: *(i)* individual TFs are limited in their ability to discriminate specific from random sites, $k_-^{\mathrm{S}}/k_-^{\mathrm{NS}} \sim 10^{-2}$, so high specificity must be a collective enhancer effect in the proofreading regime where $k_{\mathrm{link}} \sim k_-^{\mathrm{S}}$; *(ii)* mean TF residence times in an enhancer are not much higher than the typical TF residence time at an isolated specific site, $T_{\mathrm{TF}}/T_0 \lesssim 10$, enabling rapid turnover of bound TFs on the $1 - 10$ s timescale; (iii) typical sensitivities are much lower than the total number of TF binding sites, yielding a reasonable specificity/noise balance at $H \sim n/2$ (Fig 3.16, Fig 3.17); *(iv)* Mediator basal rates should maximize κ_-/κ_+, i.e., mediator switches `OFF` essentially instantaneously if not stabilized by linked TFs; *(v)* TF concentrations required to activate the enhancer in this regime are substantially higher than expected for the equivalent but highly cooperative enhanceosome (at higher α); *(vi)* optimal nonequilibrium models achieve order-of-magnitude improvements in S relative to matched equilibrium models—thereby avoiding crosstalk and spurious gene expression—by suppressing induction from non-cognate (random) DNA, while induction curves from functional enhancers bear no clear signatures of non-equilibrium operation; *(vii)* to permit large increases in specificity S, enhancer state fluctuations will develop long timescale correlations,

$T_E \gg T_{\mathrm{TF}}$ (but still be bounded by the protein lifetime, $T_E \lesssim T_P$ to enable noise averaging), leading to substantial observed noise levels; *(viii)* the enhancer `ON` residence time distribution will be non-exponential, with excess probability for very long-lived events, during which an enhancer could trigger a transcriptional burst following an interaction with the promoter; *(ix)* in our model, long correlation time, T_E, in steady state also implies long (minutes to hours) response times when TF concentration change, which would be observable with live imaging on the transcriptional, but likely not protein-concentration, level.

We find it intriguing that a single-parameter extension of a classic equilibrium model led to such richness of observed behaviors, and to a suggestion that the optimal operating regime is very different from regulation at equilibrium. Central to this qualitative change is the fact that long fluctuation and response timescales of enhancer activation appear necessary to achieve high specificity of regulation through proofreading. Such long timescales are not inconsistent with our current knowledge. Indeed, some developmental enhancers form active clusters (super-enhancers) that are rather long-lived (order of minute to hours), perhaps precisely because developmental events need to be guided with extraordinary precision [Cho *et al.*, 2018; Sabari *et al.*, 2018].

A strong objection to our model could be that it is too simple: after all, we neglected many structural and molecular details, many of which we may not even know yet. This is certainly true and was done, in part, on purpose, to permit exhaustive analysis across the complete parameter space. Such understanding would have been impossible if we explored much richer models or were concerned with quantitative fitting to a particular dataset. These are clearly the next steps, to which we contribute by highlighting the functional importance of breaking the equilibrium link between TF binding and enhancer activation state. Since our model is fully probabilistic, specializing it for a particular experimental setup, e.g., live transcriptional imaging, and doing rigorous inference is technically tractable, but beyond the scope of this chapter.

Perhaps a key simplification of our model is the link between enhancer / Mediator `ON` state and transcriptional activity. We assumed that expression is proportional to

the probability of enhancer state to be ON, yet the enhancer-promoter interaction itself is a matter of vibrant current experimentation and modeling [Bartman *et al.*, 2016; Ren *et al.*, 2017; Hnisz *et al.*, 2017; Chen *et al.*, 2018; Bialek *et al.*, 2019]. For example, long-lived activated enhancers that we predict could interact with promoters only intermittently to trigger transcriptional bursts, as suggested by the "dynamic kissing model" [Cho *et al.*, 2018], which could substantially impact the experimentally-observable quantitative noise signatures of enhancer function at the transcriptional level. Whatever the true nature of enhancer-promoter interactions might be, however, they are unlikely to be able to remove excess enhancer switching noise, due to its slow timescale, suggesting that the tradeoffs that we identify should hold generically.

One could also question whether the importance we ascribed to high specificity is really warranted. Evolutionarily, regulatory crosstalk due to lower specificity helps networks evolve during transient bouts of adaptation, even though it could be ultimately selected against [Friedlander *et al.*, 2017]. Mechanistically, molecular mechanisms such as chromatin modification or the regulated 3D structure of DNA decrease the number of possible non-cognate targets that could trigger erroneous gene expression [Adam *et al.*, 2015; Klemm *et al.*, 2019], and thus alleviate the need for the high specificity of the transcriptional control. Empirically, there is ample evidence for abortive or non-sensical transcriptional activity [Struhl, 2007; Ehrensberger *et al.*, 2013], whose products could be dealt with downstream or simply ignored by the cell. Yet it is also clear that regulatory specificity must be a collective effect, as individual TFs bind pervasively across DNA even in non-regulatory regions [Biggin, 2011], and self-consistent arguments suggest that in absence of non-equilibrium mechanisms, crosstalk could be overwhelming in eukaryotes [Friedlander *et al.*, 2016]. It is also possible that real enhancers are very diverse with large variation along the specificity axis, thereby navigating the noise-specificity tradeoff as appropriate given the biological context. Where some erroneous induction can be tolerated, expression could be quicker, less noisy, and closer to equilibrium. In contrast, where tight control is needed, enhancers could take a substantial amount of time to commit to expression correctly, perhaps benefitting additionally from extra time-averaging that could further reduce the

Berg-Purcell-type noise intrinsic to TF concentration sensing [Berg and Purcell, 1977; Bialek and Setayeshgar, 2005; Tkačik *et al.*, 2008; Kaizu *et al.*, 2014].

3.4 Supporting Information

3.4.1 Model

Model specification

We consider a class of toy models for transcriptional regulation that could plausibly be employed by eukaryotic cells. Specifically, we look for models where transcription factors (TFs) interact with special regulatory sequences on the DNA, known as binding sites (BSs) in the enhancer, to control the expression of a given gene. The emphasis here is not on devising a single scheme that has a direct molecular interpretation, but rather to ask about *possible* schemes that simultaneously achieve several properties which are desirable for efficient regulation and are consistent with metazoan observations.

Our model, schematically displayed in Fig 3.10 for a single binding site for simplicity, includes a large pool of TFs at a fixed concentration c of the same type that can bind to n binding sites in the enhancer with a concentration-dependent rate $k_+ = k_+^0 c$ and unbind with a concentration-independent rate k_-. Additionally, a complex of transcriptional co-factors that we refer to as a "Mediator" can switch between `ON` and `OFF` state with rates κ_+ and κ_-. Only when Mediator is found in an `ON` state, can a so-called link between any bound TF and Mediator be established with a rate k_{link}. In principle, the link could be removed actively at a rate k_{unlink}, but here we assume for simplicity that $k_{\text{unlink}} = 0$ (we later relax this assumption). While the links are not removed actively, they are removed automatically when the TFs dissociate or upon the Mediator switch into `OFF` state. Molecularly, the link formation and removal could be catalyzed by dedicated enzymes, and when coupled to an energy source, could be kept out-of-equilibrium. The formation of any link increases the stability of the linked complex by decreasing the rate of unbinding of the linked molecules (and the rate of Mediator switching `OFF`) by a

constant multiplicative factor; in other words, the Mediator OFF switching rate falls as a power in the number of links with TFs. We will see that this ansatz permits our model to have a clear thermodynamic-equilibrium limit.

Parameters of the model. The parameters of the model are:

- n – number of specific binding sites in the enhancer.
- α – fold-reduction in the unbinding rate of a linked TF ($k_- \rightarrow k_-/\alpha$) or Mediator OFF switching rate ($\kappa_- \rightarrow \kappa_-/\alpha^b, b \in \{0, 1, \ldots, n\}$ is number of links Mediator has with bound TFs). We focus on values $\alpha \geq 1$ which stabilize the bound complexes.
- k_+ – binding rate of TF to a BS. $k_+ = k_+^0 c$, with k_+^0 being the binding rate per unit concentration and c is the free concentration of TFs.
- k_- – unbinding rate of TF bound in the enhancer. Generally, the unbinding rate of a TF depends on the presence of a link of the TF with the Mediator. If a link is present, the unbinding rate becomes k_-/α. Importantly, k_- is the only sequence dependent quantity in the model: $k_- = k_-^{\mathrm{S}}$ for an isolated unlinked TF bound on the specific BS, and $k_- = k_-^{\mathrm{NS}}$ for an isolated unlinked TF bound on a non-specific, random site.
- k_{link} – rate of establishing new links between bound TF and the Mediator in ON state. If the link-forming reaction were catalyzed by an enzyme with own sequence specificity, we can achieve an even higher specificity of our regulatory scheme (examined later in this document). To be conservative, we set this rate to be constant and thus have no sequence specificity, i.e., once TF is bound and Mediator is in ON state, the link can be created with the same rate regardless of whether this happens in the enhancer or on a random piece of genomic sequence.
- κ_+ – switching rate into ON state of the Mediator.
- κ_- – switching rate into OFF state of the Mediator. Generally, this rate depends on the number of links a Mediator has established with TFs. Thus, this rate takes a form κ_-/α^b, where b is the number of linked bound TFs.

- k_{unlink} – rate of active link removal. In our default model, we set this parameter to $k_{\text{unlink}} = 0$. Fig 3.11 analyzes the effects of this assumption.

In our analysis we will focus on the effects of k_{link}, α, and n while taking representative values of the other parameters; if not stated otherwise, we use $k_{\text{unlink}} = 0$, $k_+ = ck_+^0$, with $c = c_0 = 0.01$ and $k_+^0 = 1$, $\kappa_- = 10^4$, $\kappa_+ = 10^{-2}$, $k_-^{\text{S}} = 10^{-2}$, and $k_-^{\text{NS}} = 1$. The latter also sets the timescale in our model, that is, we define $T_0 = 1/k_-^{\text{NS}} = 1$, i.e., the typical time a TF in isolation is bound on a random, non-specific site on the DNA, as our time unit. For exploration of the phase space we use $\alpha \in (1, 10^{10})$ and $k_{\text{link}} \in (10^{-8}, 10^8)$.

Dynamical variables and computation of the model. The dynamical variables of our model are:

- s_i – an indicator variable in $\{0, 1\}$ indicating if a TF is bound on the site $i = 1, \ldots, n$.
- b_i – an indicator variable in $\{0, 1\}$ indicating if TF at site i has a link with the Mediator. It can take a value of 1 only if the TF at site i is bound, i.e., if $s_i = 1$.
- s_M – an indicator variable in $\{0, 1\}$ indicating if the Mediator is in `ON` state.

The behavior of the system in state space of $\{s_i, b_i, s_M\}$ is a continuous-time Markov chain, with the rates fixed by our parameters. Generally, we can write down our system as a Master equation for the Markov chain:

$$\frac{\mathrm{d}\mathbf{V}}{\mathrm{d}t} = \hat{M}\mathbf{V}, \tag{3.2}$$

with $\mathbf{V}$ being a vector whose components are the probabilities of the system to be in any of the states at time t (and thus $\sum_{j=1}^{m} V_j = 1$, where the sum is taken over all m components of the vector $\mathbf{V}$) and $\hat{M}$ the transition matrix between different states. However, in practice the number of all possible microstates is large, making explicit manipulation of the Master equation feasible only for smaller values of n.

Constructing the transition matrix $\hat{M}$. In this paragraph we describe how to construct the transition matrix $\hat{M}$. First, we define a state vector $\mathbf{B} = (s_M, s_1, \ldots, s_n, b_1, \ldots, b_n)$.

With the state vector **B** we can enumerate all possible states of $s_M \in \{0,1\}$, $s_i \in \{0,1\}$, and $b_i \in \{0,1\}$. However, the b_i values are constrained by the binding state of the TFs (s_i) and cannot independently take on arbitrary values. For example, if $s_i = 0$ ($i-$th TF not bound), then there can never be any link, i.e., $b_i = 0$. Only if $s_i = 1$, then $b_i \in \{0,1\}$. If we take the three example from Fig 3.1A with $n = 3$ at increasing TF concentrations, the corresponding state vectors would be: $S_{\text{low c}} = (0,1,0,0,0,0,0)$, $S_{\text{medium c}} = (1,1,1,0,1,0,0)$, $S_{\text{high c}} = (1,1,1,1,1,1,1)$.

Altogether there are

$$m = \sum_{i=0}^{n} \binom{n}{i} 2^i + 2^n \tag{3.3}$$

different states, where the sum goes over all possible combinations of i bound and potentially linked TFs with the Mediator in `ON` state. The second part represents the number of different states when Mediator is in `OFF` state, i.e., when $s_M = 0$. Next, we order the states such that states with $s_M = 1$ come first, followed by states $s_M = 0$.

We can write the transition matrix by accounting for all possible events, and then finding states between which events cannot occur. Roughly, there are 3 types of events: (i) binding and unbinding of TFs, (ii) linking and unlinking, and (iii) switching the Mediator between `ON` and `OFF` (main text Fig 3.1B). In the following procedure, we will go over the three different types of events, finding all possible transitions between them, and assigning rates to those state-change events in the transition matrix. Due to symmetries, directly finding only a subset of events is enough. For example, starting at state j, let us assume that a linking event can lead to state k. As the unlinking events are reciprocal to linking events, the unlinking of the same TF (assuming states of all other TFs and Mediator did not change) would lead from state k to state j. However, this is not entirely correct for binding and unbinding events − the extra complication is that unbinding can also destroy a link and reciprocity between binding and unbinding does not always exist. Therefore, if unbinding of a TF leads from state j to state k, the state j can be reached from state k only if the unbinding did not destroy a link. This means that there are more unbinding transitions than there are binding transitions. Using this approach, we will find all possible state transitions. The procedure to write down the elements in

the transition matrix $\hat{M}$ is:

For each possible state vector **B** (`original state`) do

1. First we locate all linking and unlinking transitions. Therefore, for each TF i in order:

   - If the $i-$th TF is not linked (i.e., $B(i+n+1)=0$), continue to the next TF.
   - Otherwise, define a `new state` by removing the present link at $i-$th TF from the `original state`. This new state has all elements the same as original state with the exception of no link at TF i ($B_{\text{new}}(i+n+1)=0$).
   - Mark the unlinking transition with unlinking rate as $M(\texttt{new state}, \texttt{original state}) = k_{\text{unlink}}$.
   - Mark linking transition as $M(\texttt{original state}, \texttt{new state}) = k_{\text{link}}$.

2. Next, we locate all binding and unbinding transitions. As the two are not always reciprocal, we have to follow if a link is broken when unbinding occurs. For each TF i in order:

   - If the $i-$th TF is not bound ($B(i+1)=0$), continue to the next TF.
   - Otherwise, define a `new state` by removing the bound $i-$th TF from the `original state`; further, remove the link of $i-$th TF (if it existed in the `original state`).
   - Count the number of removed links b_i: 0 if the bound TF was not linked and 1 if it was.
   - Mark the unbinding transition with unbinding rate as $M(\texttt{new state}, \texttt{original state}) = k_-/\alpha_i^b$.
   - If no link was removed (i.e., $b_i = 0$), this means that binding from the new to the original state can occur. Therefore, mark binding transition as $M(\texttt{original state}, \texttt{new state}) = k_+$.

3. Lastly, locate the states that are affected by Mediator switching:

- If the Mediator is `OFF` in the `original state`, continue to the next state vector **B** and restart this processing at (1). If the Mediator is `ON` in the `original state`, define a `new state` by switching the Mediator into `OFF` state.
- If any links existed between the Mediator and any other TF in the `original state`, remove them in the `new state`.
- Count the number of removed links b.
- Mark the transition into `OFF` state as: $M(\texttt{new state}, \texttt{original state}) = \kappa_- / \alpha^b$.
- If no links were removed (i.e., $b = 0$), mark the transition into `ON` state as: $M(\texttt{original state}, \texttt{new state}) = \kappa_+$.

At the end we set the diagonal values as minus sum of the columns.

Residence time distributions

Since all the individual processes involved in our enhancer models are Poisson processes occurring either sequentially or in phase, the residence time distributions of a given TF site or Mediator being `ON` are phase-type distributions. To compute these distributions, we first need to define various subsets of states. The set I_b of states correspond to a given TF site or Mediator being bound (`ON`). The set I_u of states correspond to a given TF site or Mediator being unbound (`OFF`). The set I_{bn} of states correspond to a given TF site or Mediator being bound (`ON`) with no link attached. We can then define the following matrix from the original transition matrix $\hat{M}$ (Eq. 3.2)

$$\hat{M}_w = \hat{J}\hat{M}\hat{J}^{\mathbf{t}}, \tag{3.4}$$

where $\hat{J}$ is a diagonal matrix whose entries J_{ii} are equal to 1 if $i \in I_b$ and zero otherwise. We will also define a vector **a** describing the probability for the system to have just settled in any of the bound states

$$\mathbf{a}_i = \frac{\tilde{\mathbf{a}}_i}{\sum_{i=1}^{n} \tilde{\mathbf{a}}_i} \quad \text{with} \quad \tilde{\mathbf{a}}_i = \begin{cases} \sum_{j \in I_u} \hat{M}_{ij} \mathbf{V}_j & \text{for } i \in I_{bn} \\ 0 & \text{otherwise} \end{cases}, \tag{3.5}$$

where **V** is the vector of steady state occupancies that is computed from Eq. 3.32. The probability density function for the residence time of a given TF site or Mediator being bound is then given by

$$f(T) = -\mathbf{I}^{\mathbf{t}} \exp\left(\hat{M}_w T\right) \hat{M}_w \mathbf{a}, \tag{3.6}$$

where **I** is a vector whose entries are all equal to one. Of note, the exponential here is the matrix exponential. The mean residence time μ_T and the variance of the distribution σ_T^2 are then given by

$$\begin{aligned} \mu_T &= -\mathbf{I}^{\mathbf{t}} \hat{M}_w^{-1} \mathbf{a} \\ \sigma_T^2 &= 2\mathbf{I}^{\mathbf{t}} \hat{M}_w^{-2} \mathbf{a} - \mu_T^2. \end{aligned} \tag{3.7}$$

Equilibrium limits

Our model is a generalization of an equilibrium MWC model. In the following section, we first show that our model reduces to the MWC model in the equilibrium limit, and we then derive the different regulatory phenotypes in this limit. At the end we also address the Hill-type models.

MWC model. As a thermodynamic equilibrium model, one can fully specify the MWC model by means of a partition function Z that enumerates all the possible states of the system. The partition function of the MWC model with n TF binding sites can be written as

$$Z = \sum_{\{\sigma_M, \sigma_i\}} \exp\left[L\sigma_M + (\epsilon + \log c + \delta\sigma_M) \sum_{i=1}^{n} \sigma_i \right], \tag{3.8}$$

where $\sigma_M \in \{0, 1\}$ and $\sigma_i \in \{0, 1\}$ are the occupancy variables for Mediator and the TF binding sites i. The different energy contributions in our model are L, ϵ and δ, which represent the energy difference between the Mediator `ON` and `OFF` state, between an empty and occupied TF binding site, and the energy benefit due to the established link. Their Boltzmann states can be respectively written as e^L, e^δ, and ce^ϵ where c represents the concentration of TFs. For example, the energy and Bolztmann weight of a state with Mediator in `ON` state and two bound and linked TFs would read $L + 2\epsilon + 2\delta$ and $ce^{L+2\epsilon+2\delta}$, respectively.

MWC is an equilibrium limit of our model. It turns out that our proposed model collapses into an equilibrium MWC model when taking the limit $k_{\text{link}} \to \infty$. This can be shown without loss of generality, by examining the simplest case of our model with $n = 1$. In that case, the single irreversible step dictating the non-equilibrium nature of our model occurs between the two following states: i) the TF is bound and Mediator `ON` but no link is present, and ii) a link is established between Mediator and the TF (Fig 3.5). When increasing k_{link}, such that $k_{\text{link}} \gg k_-, \kappa_-$, the transition between these two states becomes very fast, and the dwell time in the first state becomes negligible. Thus, in the limit of $k_{\text{link}} \to \infty$, the two states with a bound TF and Mediator `ON` collapse into a single state where a link is always present (Fig 3.5). The resulting kinetic scheme exactly corresponds to an equilibrium MWC model.

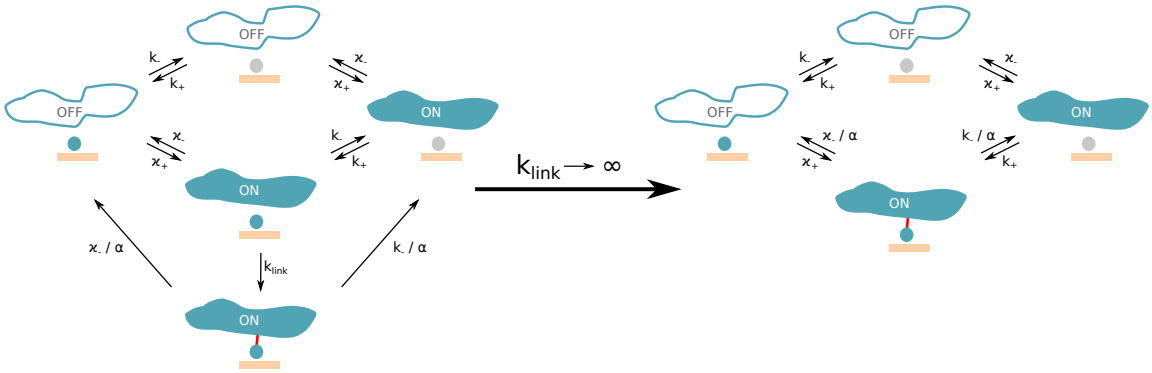

Figure 3.5: **Our model is a generalization of MWC model.** A comparison between schemes for our non-equilibrium model (left) and MWC model (right) for $n = 1$. When taking the limit $k_{\text{link}} \to \infty$, our model converges into MWC model.

To make the correspondence clear, we connect the equilibrium energies defined in the MWC partition function (Eq. 3.8) with the kinetic rates of our model. Due to equilibrium, all processes follow detailed balance:

$$\pi_i W_{ij} = \pi_j W_{ji}, \tag{3.9}$$

where W_{ij} is the transition rate from state i to state j, and π_i and π_j are the equilibrium probabilities of being in states i and j, respectively.

First, we compare two states devoid of bound TFs where Mediator is `OFF` and `ON`

respectively. Transitioning between these two states only involves the Mediator kinetic rates κ_+ and κ_-. Applying detailed balance gives $\frac{1}{Z}\kappa_+ = \frac{e^L}{Z}\kappa_-$, where $1/Z$ and e^L/Z are the equilibrium probabilities π of the `ON` and `OFF` state, respectively. It follows from this condition that $e^L = \frac{\kappa_+}{\kappa_-}$.

Similarly, we compare transitions between an empty state and a state with one bound TF. Following detailed balance we obtain $\frac{1}{Z}k_+ = \frac{ce^\epsilon}{Z}k_-$, where $1/Z$ and $\frac{ce^\epsilon}{Z}$ are the equilibrium probabilities π of the TF unbound and bound state. The transition rates k_+ and k_- are the rates of TF binding and unbinding, respectively. It thus follows that $ce^\epsilon = \frac{k_+}{k_-}$.

Lastly, we compare the two following states; i) Mediator is `ON` without any TF bound, and ii) Mediator is `ON` and a TF is bound. In the equilibrium limit, both Mediator and the TF are linked when present together. The equilibrium probabilities π of the two states are ce^L/Z and $ce^{L+\epsilon+\delta}/Z$, respectively. The transition rates between these states are k_+ and k_-/α. By applying the detailed balance condition, $\frac{ce^L}{Z}k_+ = \frac{ce^{L+\epsilon+\delta}}{Z}\frac{k_-}{\alpha}$, we obtain $e^\delta = \alpha$. This demonstrates a one-to-one correspondence between the "cooperative energy of binding" in thermodynamic models of gene regulation, and the parameter α of the non-equilibrium model.

Expression. In our model, we defined expression as the occupancy of Mediator in the `ON` state. To calculate the expected expression in the MWC model, we first separate the partition function (Eq. 3.8) in two sub-partitions Z_{ON} and Z_{OFF} such that $Z = Z_{\text{ON}} + Z_{\text{OFF}}$. Here, Z_{ON} and Z_{OFF} correspond to the sum over all the states with Mediator `ON` and `OFF` respectively. These sums can be calculated as follows:

$$\begin{aligned} Z_{\text{ON}} =& e^L \sum_{\{\sigma_i\}} \exp\left[(\epsilon + \log c + \delta)\sum_{i=1}^{n}\sigma_i\right] \\ =& e^L \sum_{\{\sigma_i\}} \prod_{i=1}^{n} ce^\epsilon e^\delta e^{\sigma_i} \\ =& e^L \sum_{k=0}^{n} \binom{n}{k} (ce^\epsilon e^\delta)^k \cdot 1^{n-k} \\ =& e^L (1 + ce^\epsilon e^\delta)^n. \end{aligned} \tag{3.10}$$

Similarly, we obtain $Z_{\mathrm{OFF}} = (1 + ce^{\epsilon})^n$. The probabilities to find the Mediator in the `ON` and `OFF` states can then be expressed as

$$P_{\mathrm{ON}} = \frac{Z_{\mathrm{ON}}}{Z} = \frac{1}{Z} e^{L} \left(1 + ce^{\epsilon} e^{\delta}\right)^n$$
$$P_{\mathrm{OFF}} = \frac{Z_{\mathrm{OFF}}}{Z} = \frac{1}{Z} \left(1 + ce^{\epsilon}\right)^n . \tag{3.11}$$

We can thus write the occupancy of Mediator in the `ON` state, which corresponds to our definition of expression:

$$E = \frac{P_{\mathrm{ON}}}{P_{\mathrm{ON}} + P_{\mathrm{OFF}}} = \frac{e^{L} \left(1 + ce^{\epsilon} e^{\delta}\right)^n}{e^{L} \left(1 + ce^{\epsilon} e^{\delta}\right)^n + \left(1 + ce^{\epsilon}\right)^n} = \left[1 + e^{-L} \left(\frac{1 + ce^{\epsilon}}{1 + ce^{\epsilon} e^{\delta}}\right)^n\right]^{-1} . \tag{3.12}$$

As in the main text, all the bounds derived and reported below come from varying k_{link} and α, while keeping other parameters constant. The only exception is concentration c which is either kept constant or adjusted to achieve fixed expression E.
As per definition of occupancy, expression is bounded from above by $E = 1$; that occurs when Mediator is always in `ON` state. The lower bound, $\min E$, occurs when binding sites are almost never occupied. In that case, the expression is solely determined by the intrinsic Mediator `ON` probability, thus $E = \kappa_+/(\kappa_+ + \kappa_-)$, which reduces to $E = \kappa_+/\kappa_-$ when $\kappa_- \gg \kappa_+$, as we assumed. Therefore, expression is limited to $E \in (\frac{\kappa_+}{\kappa_-}, 1)$.
When fixing specific expression to $E^{\mathrm{S}} = E_0$, concentration must vary to meet that requirement. By equating $E_0 = \frac{P_{\mathrm{ON}}}{P_{\mathrm{ON}} + P_{\mathrm{OFF}}}$, and solving for c, we obtain:

$$c = \frac{k_-}{k_+^0} \frac{x - 1}{1 - x\alpha} \quad \text{with} \quad x = \left(\frac{\kappa_+(1 - E_0)}{\kappa_- E_0}\right)^{1/n} . \tag{3.13}$$

Since concentration must be positive, α must satisfy $\alpha \geq \frac{1}{x}$. From this inequality we obtain a lower bound for α

$$\alpha_{\min} = \left(\frac{\kappa_- E_0}{\kappa_+(1 - E_0)}\right)^{1/n} . \tag{3.14}$$

.

Specificity. We define specificity as the ratio of expression from a functional enhancer, E^{S}, and expression from a random piece of sequence, E^{NS}. Therefore:

$$S = \frac{E^{\mathrm{S}}}{E^{\mathrm{NS}}}. \tag{3.15}$$

By definition of specific binding site, $E^{\mathrm{S}} \geq E^{\mathrm{NS}}$, leading to $\min S = 1$.
Furthermore, a general upper bound of specificity is given by the ratio of Mediator switching rates, $\max S = \kappa_-/\kappa_+$. This upper bound is attained when specific expression is maximal, $E^{\mathrm{S}} = 1$, while non-specific expression is minimal $E^{\mathrm{NS}} = \frac{\kappa_+}{\kappa_-}$. Thus, $S \in (1, \kappa_-/\kappa_+)$. However, if a system has a fixed specific expression $E \neq 1$ (as in Fig 3.2C), the upper bound is adjusted by a factor of E. Indeed, the specific expression takes value $E^{\mathrm{S}} = E$ by construction while minimal non-specific expression is again $E^{\mathrm{NS}} = \kappa_+/\kappa_-$. Taking their ratio we thus obtain $\max S_{\text{fixed E}} = E\kappa_-/\kappa_+$.

Residence time. We defined TF residence time as the average time a TF spends bound to its specific binding site. To calculate the TF residence time, we assumed that changes in the Mediator state do not happen while the TF is bound, which means that the residence time of TFs is either much shorter or longer than the time Mediator spends in `ON` state. This is a valid assumption in our model, since the Mediator `ON` state is either very short lived due to high `OFF` rate in absence of any link, or very long lived due to very small `OFF` rate in presence of stabilizing links.
Based on the assumption above, we can calculate the residence time as the weighted average of the average time spent in the two following configurations: a TF resides on a binding site with a link (Mediator `ON`), and without any link (Mediator `OFF`). These average times are given by the inverse of the escape rate, namely the inverse TF unbinding rate e^{δ}/k_- with a link and $1/k_-$ without a link. To obtain the mean residence time, one needs to properly average the two durations above. The weights to perform the average are given by the probabilities A_{ON} and A_{OFF} that the system has just settled in these configurations. The residence time of a single TF being bound is then given by

$$T_{\mathrm{TF}} = \frac{e^{\delta}}{k_-} A_{\mathrm{ON}} + \frac{1}{k_-} A_{\mathrm{OFF}}. \tag{3.16}$$

The probabilities A_{ON} and A_{OFF} are proportional to the product of i) the probability W to find the system with a given binding site unoccupied, and ii) the rate of TF binding k_+. The probabilities W for a given binding site being unoccupied can be calculated from the partition function Z for n binding sites (Eq. 3.8). After partitioning the states into Mediator `ON` and `OFF`, the resulting probabilities are proportional to the sum over all configurations of a $n-1$ system (Eq. 3.10):

$$\begin{aligned} W_{\mathrm{ON}} &= \frac{1}{Z} e^{L} \left(1 + c e^{\epsilon} e^{\delta}\right)^{n-1} \\ W_{\mathrm{OFF}} &= \frac{1}{Z} \left(1 + c e^{\epsilon}\right)^{n-1} \end{aligned} \tag{3.17}$$

We can then express A_{ON} and A_{OFF} as

$$\begin{aligned} A_{\mathrm{ON}} &= k_+ W_{\mathrm{ON}} / A \\ A_{\mathrm{OFF}} &= k_+ W_{\mathrm{OFF}} / A, \end{aligned} \tag{3.18}$$

where A is normalization constant that ensures $A_{\mathrm{ON}} + A_{\mathrm{OFF}} = 1$. It follows that $A_{\mathrm{ON}} = W_{\mathrm{ON}}/(W_{\mathrm{ON}} + W_{\mathrm{OFF}})$ and $A_{\mathrm{OFF}} = W_{\mathrm{OFF}}/(W_{\mathrm{ON}} + W_{\mathrm{OFF}})$. Finally, by plugging these expressions into Eq. 3.16, we find that the residence time of a single TF on a binding site is given by

$$T_{\mathrm{TF}} = \frac{1}{k_-} \frac{W_{\mathrm{ON}} e^{\delta} + W_{\mathrm{OFF}}}{W_{\mathrm{ON}} + W_{\mathrm{OFF}}}. \tag{3.19}$$

Numerical results show that our assumption about time-scale separation in this model is a valid approximation (Fig 3.2A,C).

Using the expression for the residence time above, we derive the lowest achievable T_{TF} in different scenarios. Since T_{TF} increases with the stability of complexes (i.e., by increasing $\alpha = e^{\delta}$), we can calculate the lower bound on T_{TF} by using the smallest possible α, namely $\alpha = 1$ at fixed concentration and $\alpha = \alpha_{\min}$ (Eq. 3.14) at fixed expression E_0. We thus obtain the following $\min T_{\mathrm{TF}}$:

$$\min T_{\mathrm{TF}} = \frac{1}{k_-^S}, \qquad \text{for fixed } c, \tag{3.20}$$

$$\min T_{\mathrm{TF}} = \min T_{\mathrm{TF}} = \frac{1}{k_-^{\mathrm{S}}} \cdot \frac{1}{1 + E_0(\alpha_{\min}^{-1} - 1)}, \quad \text{for fixed } E. \tag{3.21}$$

Hill-type models. As the number of possible equilibrium models for enhancer regulation is unlimited, let us consider at least one alternative to MWC models: Hill-type models. In these models, the presence of Mediator is not required to mediate the stabilization of TFs through links. Only the description of the TF interactions with DNA and the TF interaction with each other is necessary. In that scenario, linking could occur between neighbouring bound TFs (1D chain), between any pair of bound TFs, or via some other intermediate interaction scenario. As in the MWC model we considered, the creation of a link would lead to a multiplicative decrease in unbinding rate of TF by a factor of α.

Since in Hill-type models we no longer have a two-state Mediator that naturally dictates what "active" enhancer (and thus expression) means, we need to revise our definition of expression. There are multiple possible definitions specifying the TF binding / linking configurations that lead to productive expression, i.e., are considered as effective `ON` states. The only constraint in order to preserve the proof-reading mechanisms and the high specificity advantage is that expression has to occur from states in which TFs are not only bound but also linked. For example, it could be i) all states that have at least one link, ii) only the state where all possible links are established, or iii) some other similar combination.

Let us show an example of non-equilibrium extension of a Hill-type model. We consider a 1D chain model where links can be established only between neighbouring bound TFs. In the non-equilibrium version of this model, links are not immediately created but are established with finite rate k_{link}. As in our MWC model extension, in the limit of $k_{\text{link}} \to \infty$, the states that differ only in the presence or absence of a link collapse into a single state. This occurs because as k_{link} increases, the transitions from unlinked to linked state become much faster, until the two states are indistinguishable. Fig 3.6 shows an example for $n = 2$ binding sites. In the equilibrium limit at large k_{link}, assuming expression occurs only from the linked TF state, it is straightforward to write down the partition function, show that it predicts a Hill function with $n = 2$ for the induction curve, and that parameter α is directly related to the cooperative energy of interaction carried by the "link".

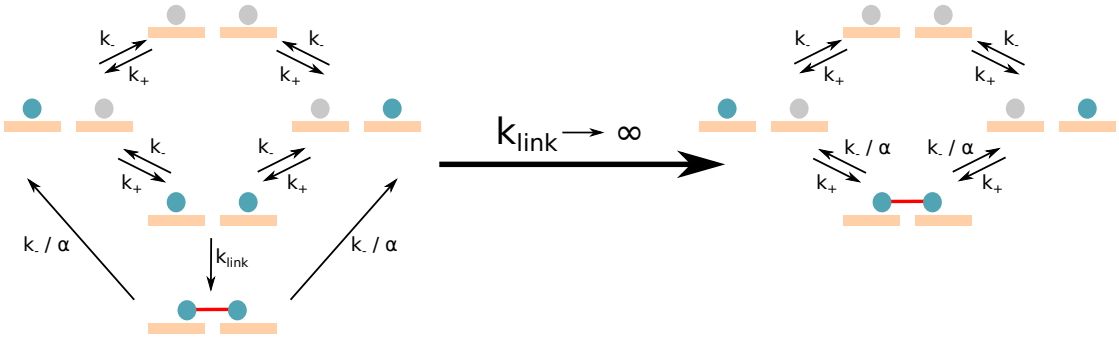

Figure 3.6: **Non-equilibrium model extension of a Hill-type model.** A comparison between schemes of non-equilibrium extension (left) and equilibrium (right) Hill-type model for $n = 2$. When $k_{\text{link}} \to \infty$, the two models collapse. In the example of this model, links can be established only between two neighbouring bound TFs (red line).

Regulatory phenotypes

Expression E is the normalized expression level of a gene expressed under the control of the modeled enhancer. We compute expression as the fraction of time the Mediator is in `ON` state, that is $E = \langle s_M \rangle$. The average is taken over the stationary distribution of the master equation (except where we study transient effects, as in main text Fig 3.4). In practice, this means that we first compute the stationary solution $\mathbf{V}$ of Eq. 3.2: $\mathrm{d}\mathbf{V}/\mathrm{d}t = \hat{M}\mathbf{V} = 0$. We then marginalize $\mathbf{V}$ to obtain $E = \sum_{j=1}^{m} V_j I(s_M = 1)$, where the sum is taken over all the states of the Markov chain and $I(s_M = 1)$ is the indicator function which is 1 if the Mediator is `ON` in state indexed with j and zero otherwise. As in the equilibrium limit, the expression is bounded: $E \in (\frac{k_+}{\kappa_-}, 1)$.

We expect that functional enhancers lead to high expression when TF concentration is high, which, in our model, should correlate with high occupancy of TFs on the specific BSs in the enhancer. We thus require the Mediator to be `ON` with high probability (typically $E \sim 0.5$, although we also consider in the main chapter scenarios where E can be smaller).

Specificity S is the ratio between the level of expression from a functional enhancer (i.e., enhancer that contains n specific BSs), and expression from a random piece of sequence, $S = E^{\mathrm{S}}/E^{\mathrm{NS}}$. High specificity of regulation ($S > 1$) is generally realized as a collective state of many bound TFs interacting with the Mediator. As in the equilibrium limit, specificity is bounded $S \in (1, \kappa_-/\kappa_+)$. In addition, when the specific expression is fixed $E \neq 1$ (as in Fig 3.2C), the upper bound is adjusted by a factor of E, such that $\max S_{\text{fixed E}} = E\kappa_-/\kappa_+$.

In our model, we estimate specificity by independently computing the expression E for specific and non-specific site (i.e., for two different values for the unbinding rate, k_-^{S} and k_-^{NS}), then taking their ratio.

We expect specificity to be as high as possible. Indeed, high specificity allows for accurate binding and control: first, it ensures that most of the TFs are not sequestered stably on random sequences; second, this further ensures that non-specific binding of TFs to non-cognate regulatory regions in the genome does not lead to erroneous

gene expression, also known as transcriptional crosstalk. Given the high relative excess of possible non-specific binding configurations in the genome that outnumber binding configuration in the cognate regulatory region by thousands or millions, specificity should numerically be as high as possible.

Residence time T_{TF} is the average time that a TF spends bound to its specific BS. As in the equilibrium limit, the shortest TF residence time is obtained in absence of any stabilizing links, i.e.,

$$\min T_{\mathrm{TF}} = \frac{1}{k_{-}^{\mathrm{S}}}. \tag{3.22}$$

In other words, the minimal TF residence time is determined by the unbinding rate of an isolated TF. Furthermore, when we consider the enhancer at a fixed specific expression $E^{\mathrm{S}} = E_0$ (as in Fig 3.2C), this value gets adjusted to

$$\min T_{\mathrm{TF}} = \frac{1}{k_{-}^{\mathrm{S}}} \cdot \frac{1}{1 + E_0(\alpha_{\mathrm{min}}^{-1} - 1)}, \tag{3.23}$$

with $\alpha_{\mathrm{min}} = \left(\frac{\kappa_-}{\kappa_+}\frac{E_0}{1-E_0}\right)^{1/n}$. This bound is obtained from the EQ model given the constraint for $E^{\mathrm{S}} = E_0$ (see Chapter 3.4.1). In our model, we computed the mean TF residence time directly from transition matrix $\hat{M}$ of the system (Eq. 3.2). More details about the residence time distributions and the moments can be found in Section 3.4.1.

Overall, we expect TF residence time to be small. Indeed, small residence time should provide better responsiveness to regulatory elements and lower the noise in gene expression. Furthermore, small residence time is consistent with recent single-molecule measurements. Since residence time is expected to increase with increasing stability of complexes (i.e., by increasing α and k_{link}), there should be some trade off residence time and specificity.

Sensitivity H refers to the effective steepness of the steady-state input/output curve that maps out the gene expression level as a function of the TF concentration. We compute sensitivity H as the slope of the induction curve (expression E vs concentration of TFs c on functional enhancers containing specific BSs with off-rate

k_-^S) at half maximum expression E, i.e., $H = 4\frac{c_{1/2}}{E_{\max}}\frac{dE}{dc}|_{c=c_{1/2}}$, where $\frac{dE}{dc}|_{c=c_{1/2}}$ represents the derivative of expression with respect to the concentration, taken at concentration $c_{1/2}$ where expression reaches half its maximum value $E_{\max}$. The normalization factor $4\frac{c_{1/2}}{E_{\max}}$ ensures that H is properly bounded between 1 and n. Indeed, for Hill-like functions, $E(c) = c^h/(c^h + K^h)$, the defined sensitivity corresponds exactly to the Hill coefficient, $H = h$.

We construe sensitivity H broadly, in terms of its functional effect on the shape of the induction curve regardless of the underlying molecular mechanism. Mechanisms giving rise to high sensitivity could be very diverse, for example: additional energy contribution due to a physical interaction of two TFs at the binding site, as in thermodynamic models of regulation; or a collective effect of competition of TF binding with nucleosomes, as in the MWC-like model proposed by Mirny et al. *PNAS* **107** (2010); or as a result of positive auto-regulation of a transcribed gene; or as a result of kinetic regulatory models out-of-equilibrium; or any other alternative that can increase the steepness of the induction curve beyond $H = 1$.

Mean protein number P represents the amount of protein, assuming protein dynamics is a deterministic consequence of the enhancer state. Its dynamics are governed by:

$$\frac{dP}{dt} = k_P R(t) - \frac{P}{T_P}, \tag{3.24}$$

where P represents the protein number, k_P and $1/T_P$ the protein production and degradation rate, respectively, and $R(t)$ the enhancer state: 1 for `ON` and 0 for `OFF`. $R(t)$ is a random variable whose stochastic realizations can be computed using stochastic simulation. To this end, we evolve the system state using a propagator, i.e., the formal solution of Eq. 3.2, which gives the conditional probability that the system will be found in some state after time Δt, given its current state. The enhancer state $R(t)$ is updated after each Δt by randomly drawing a binary random number according to the probabilities computed using the propagator. Mathematically, the vector of probabilities $\mathbf{W}$ to go from state j to any other state after time Δt can be

written as:

$$\mathbf{W}(\Delta t) = \exp(\hat{M}\Delta t)\,\mathbf{I}_j, \tag{3.25}$$

where $\hat{M}$ is the transition matrix (see Eq. 3.2), $\mathbf{I}_j$ a vector of zeros with value 1 at j-th entry, representing the j-th state, and "exp" represents matrix exponential (see the formalism in Section 3.4.1 for details). We compute $R(t)$ at fixed Δt time steps, with $\Delta t \ll T_{\mathrm{M}}$ and $\Delta t \ll T_{\mathrm{TF}}$, making sure that no representative `ON` state is missed. After generating a stochastic realization $R(t)$, we solve the Eq. 3.24 using standard ODE solvers using $k_p = 1$ and $T_P = 3.6 \cdot 10^6$. Assuming $1/k_-^S = 1$ s, then $T_P = 10$ h. Results in Fig 3.4C (middle- and bottom panel) represent the mean and standard deviation over 1000 replicates for different stochastic realizations of the enhancer state.

Noise in protein number N represents the variability in protein expression levels due to random enhancer state switching. N is defined as σ_P/P, where P and σ_P represent the mean and the standard deviation of protein number, respectively. For calculation of dynamical trace of the noise in Fig 3.4C, we followed the same procedure as for mean protein number (above).

Noise propagation

Telegraph model Here, we briefly review some general results regarding the telegraph model or 2-state model that includes protein production and degradation with constant rates k_P and $\gamma_P = 1/T_P$. The temporal evolution of the central moments can be derived from the master equation. The mean protein number P and the mean gene activity E satisfy the following equations

$$\begin{cases} \dfrac{d}{dt}P(t) = k_P E(t) - \gamma_P P(t) \\ \dfrac{d}{dt}E(t) = -(\kappa_+ + \kappa_-)E(t) + \kappa_+. \end{cases} \tag{3.26}$$

At steady state ($\frac{d}{dt}P = 0$ and $\frac{d}{dt}E = 0$), the mean protein number and the mean activity is simply given by $P = P_0 E$ with $P_0 = k_P/\gamma_P$ and $E = \kappa_+/(\kappa_+ + \kappa_-)$. Similarly,

the covariances satisfy the following set of equations

$$\begin{cases} \dfrac{d}{dt}\sigma_P^2(t) = -2\gamma_P\sigma_P^2(t) + 2k_P\sigma_{PE}(t) + \gamma_P P(t) + k_P E(t) \\ \dfrac{d}{dt}\sigma_{PE}(t) = -(\gamma_P + \kappa_+ + \kappa_-)\sigma_{PE}(t) + k_P\sigma_E^2(t), \end{cases} \tag{3.27}$$

where the gene state variance $\sigma_E^2(t)$ is directly determined from the evolution of the mean $E(t)$, i.e. $\sigma_E^2(t) = E(t)(1 - E(t))$. This follows immediately from the fact that one state being occupied (either the active or inactive one) necessarily implies that the other is empty. Thus σ_E^2 must be the binomial variance at all time. Solving the equations 3.27 at steady state leads to

$$\begin{aligned} \sigma_P^2 &= P_0 E + P_0\sigma_{PE} \\ \sigma_{PE} &= P_0\frac{\gamma_P}{\gamma_P + \kappa_+ + \kappa_-}\sigma_E^2. \end{aligned}$$

It follows that the protein variance is given by

$$\sigma_P^2 = P_0 E + P_0^2 E(1 - E)\Phi(T_P/T_E) \tag{3.28}$$

where $\Phi(x) = 1/(1+x) \in [0, 1]$ is a noise averaging/filtering function that determines the amount of propagated switching noise at the level of proteins by comparing the two relevant time scales of the system, namely the mean protein life time $T_P = 1/\gamma_P$ and the switching correlation time $T_E = 1/(\kappa_+ + \kappa_-)$. The first term $P_0 E$ in Eq. 3.28 corresponds to the Poisson variance resulting from the birth and death of proteins, while the second term stems from the propagation of the switching binomial variance

$$\left(\frac{dP}{dE}\right)^2 \sigma_E^2 \cdot \Phi(T_P/T_E) = P_0^2 \underbrace{E(1 - E)}_{\text{binomial variance}} \Phi(T_p/T_e). \tag{3.29}$$

In the limit of fast and slow gene switching respectively, the noise filtering function reduces to

$$\begin{aligned} T_P \gg T_E \quad & \lim_{x\to\infty}\Phi(x) = 0 \\ T_P \ll T_E \quad & \lim_{x\to 0}\Phi(x) = 1. \end{aligned}$$

As we will see later, Eq. 3.28 remains valid for all the considered enhancer models, although the functional form of Φ will now depends on the details of the kinetic scheme considered. The protein noise at steady state is thus generally given by

$$N^2 = \frac{\sigma_P^2}{P^2} = \frac{1}{P} + \frac{1-E}{E}\Phi(T_P/T_E) \simeq \frac{1-E}{E}\frac{T_E}{T_E+T_P}, \tag{3.30}$$

where in the last equality we use the filtering function of the 2-state model and we drop the Poisson noise term $1/P$ assuming large number of proteins. It turns out that the last expression still provides an excellent approximation for the amount of propagated noise in the case of sophisticated n-state model, provided we use a good proxy for the switching correlation time T_E, which we will address below.

General m-state enhancer model For any m-state model of gene activity[3] where protein production and degradation occur as Poisson processes with constant rates k_P and γ_P, the protein noise will satisfy the same functional form as the 2-state model (Eq. 3.30). In this general context, the gene mean activity E is defined as the total mean occupancies of all the gene states i allowing protein production, namely $E = \sum_{i=1}^{m_p} v_i$, with $1 \leq m_p < m$ the number of producing states and v_i the mean occupancy of state i. It turns out that the switching noise can still be obtained by propagation of the binomial variance $\sigma_E^2 = E(1-E)$ multiplied by some noise filtering function $\Phi_m \in [0,1]$ (Eq. 3.29). The only difference for a m-state model of gene activity lies in the noise filtering function Φ_m that depends on the kinetic rates and topology of the gene state transition network, i.e. the $m \times m$ state transition matrix M of the model. Starting from the equations for the first and second moment derived from the master equation (Eq. 3.2), we generalize the noise filtering function obtained for the 2-state model (Eq. 3.28) to an arbitrary number of gene states and transitions. The first moment equations are given by

$$\begin{cases} \dfrac{d}{dt}P(t) = k_P \mathbf{v_p^t}\mathbf{V}(t) - \gamma_P P(t) \\ \dfrac{d}{dt}\mathbf{V}(t) = \hat{M}_r \mathbf{V}(t) + \mathbf{f}, \end{cases} \tag{3.31}$$

where $\mathbf{V}(t) = (v_1(t), v_2(t), ..., v_{m-1}(t))^{\mathbf{t}}$ is the vector of mean occupancies, or equivalently the probability to find the system in each individual gene state $i \in \{1, .., m-1\}$.

[3] gene states described by a continuous time Markov process with linear propensity functions

The vector $\mathbf{v_p^t} = (1, ..., 1, 0, ..., 0)$ defines which gene states permit protein production, such that $E(t) = \mathbf{v_p^t V}(t)$. The operator $\hat{M}_r$ is obtained by reduction of the original transition matrix $\hat{M}$

$$\hat{M}_r = A_1 \hat{M} A_0,$$

where A_1 and A_0 are the following $(m-1) \times m$ and $m \times (m-1)$ matrices

$$A_1 = \begin{pmatrix} 1 & & 0 \\ & \ddots & \vdots \\ & 1 & 0 \end{pmatrix} \quad A_0 = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \\ -1 & \dots & -1 \end{pmatrix}.$$

The vector $\mathbf{f}$ is given by the first $m-1$ terms of the last column of M. The reduction above is necessary to later on invert the $\hat{M}_r$ operator. Indeed, the transition matrix $\hat{M}$ is degenerate by construction and has a single zero eigenvalue corresponding to the steady state solution (provided the system is ergodic), which follows from $\sum_i \hat{M}_{ij} = 0 \; \forall j$ (that ensures conservation of probability). The occupancy of the last state m is thus given by $v_m(t) = 1 - \sum_{i=1}^{m-1} v_i(t)$. Assuming steady-state, the gene state occupancies $\mathbf{V}$ are calculated from

$$\hat{M}_r \mathbf{V} + \mathbf{f} = \mathbf{0} \tag{3.32}$$

and the mean protein is given by $P = k_P \mathbf{v_p^t V}/\gamma_P = P_0 E$ as in the 2-state model. Similarly, the time evolution of the covariances can be derived from the master equation and are given by

$$\begin{cases} \dfrac{d}{dt}\sigma_P^2(t) = -2\gamma_P \sigma_P^2(t) + 2k_P \mathbf{v_p^t} \boldsymbol{\sigma}_{PV}(t) + \gamma_P P(t) + k_P \mathbf{v_p^t V}(t) \\ \dfrac{d}{dt}\boldsymbol{\sigma}_{PV}(t) = -(\gamma_P \hat{I} - \hat{M}_r)\boldsymbol{\sigma}_{PV}(t) + k_P \hat{S}(t)\mathbf{v_p}, \end{cases} \tag{3.33}$$

where $\boldsymbol{\sigma}_{PV}(t) = \left(\sigma_{P1}(t), \sigma_{P2}(t), ..., \sigma_{P(m-1)}(t)\right)^{\mathbf{t}}$ is the vector of covariances between the protein and each gene state, $\hat{I}$ the identity matrix, $\hat{S}(t)$ the covariance matrix of the gene states. Here again, it is important to realize that each gene state can only be occupied if all the others are empty. Thus, the covariance matrix $\hat{S}(t)$ is the multinomial covariance for a single trial, which is fully determined by the mean

occupancies $\mathbf{V}(t)\ \forall t$

$$\hat{S}_{ij}(t) = \begin{cases} v_i(t)(1 - v_j(t)) & \text{for } i = j \\ -v_i(t)v_j(t) & \text{for } i \neq j. \end{cases} \tag{3.34}$$

Solving Eqs 3.33 at steady state, we find

$$\begin{aligned} \sigma_P^2 &= P_0 E + P_0 \mathbf{v_p^t} \boldsymbol{\sigma}_{PV} \\ \boldsymbol{\sigma}_{PV} &= P_0 (\hat{I} - \hat{M}_r/\gamma_P)^{-1} \hat{S} \mathbf{v_p}. \end{aligned} \tag{3.35}$$

By rearranging the steady state solutions (Eq. 3.35), we recover an expression for the protein variance σ_P^2, which is similar to the one derived before for a simple switch (cf. Eq. 3.28):

$$\sigma_P^2 = P_0 E + P_0^2 E(1 - E)\Phi_m. \tag{3.36}$$

The main difference is the filtering function Φ_m now given by

$$\Phi_m = \frac{1}{E(1-E)} \mathbf{v_p^t} (\hat{I} - T_P \hat{M}_r)^{-1} \hat{S} \mathbf{v_p}, \tag{3.37}$$

with $T_P = 1/\gamma_P$ the mean protein lifetime as before. In Eq. 3.37, the binomial variance $\sigma_E^2 = E(1 - E)$ can be written as follows

$$E(1-E) = \sum_{i=1}^{m_p} v_i \left(1 - \sum_{i=1}^{m_p} v_i\right) = \sum_{i=1}^{m_p} v_i(1-v_i) - 2 \sum_{i<j\leq m_p} v_i v_j = \sum_{i=1}^{m_p} \sigma_i^2 + 2 \sum_{i<j\leq m_p} \sigma_{ij} = \mathbf{v_p^t} \hat{S} \mathbf{v_p}.$$

Plugging the above expression for the binomial variance back in Eq. 3.37, we finally find the following expression for the filtering function

$$\Phi_m = \frac{\mathbf{v_p^t} \hat{F} \hat{S} \mathbf{v_p}}{\mathbf{v_p^t} \hat{S} \mathbf{v_p}} \quad \text{with} \quad \hat{F} = (\hat{I} - T_P \hat{M}_r)^{-1}. \tag{3.38}$$

Of note, the $\hat{F}^{-1}$ operator is positive definite[4], which follows from the positive definiteness of $-\hat{M}_r$ and $T_P \geq 0$. In addition, the spectrum of $\hat{F}^{-1}$ is bounded from below, i.e. all its eigenvalues $\lambda_i \geq 1$. Thus, $\mathbf{v_p^t} \hat{F} \hat{S} \mathbf{v_p} \leq \mathbf{v_p^t} \hat{S} \mathbf{v_p}\ \forall T_P$ and the resulting

[4]Although $\hat{M}_r$ or $\hat{F}$ are not necessarily symmetric, $-\mathbf{x^t} \hat{M} \mathbf{x} > 0\ \forall$ non-zero vector $\mathbf{x}$ and all the eigenvalues of $-\hat{M}_r$ are positive. These properties follow from the structure of the master equation transition matrix $\hat{M}$.

filtering function Φ_m satisfies all the desired conditions, namely $\Phi_m \in [0, 1]$ and

$$\lim_{T_P \to \infty} \Phi_m(T_P) = 0$$

$$\lim_{T_P \to 0} \Phi_m(T_P) = 1.$$

In addition, we recover the correct expression for the 2-state model, where $\hat{M}_r = -(\kappa_+ + \kappa_-) = -1/T_E$. Indeed, $\hat{F} = (1 + T_P/T_E)^{-1}$ and thus $\Phi_2 = T_E/(T_E + T_P)$ as expected.

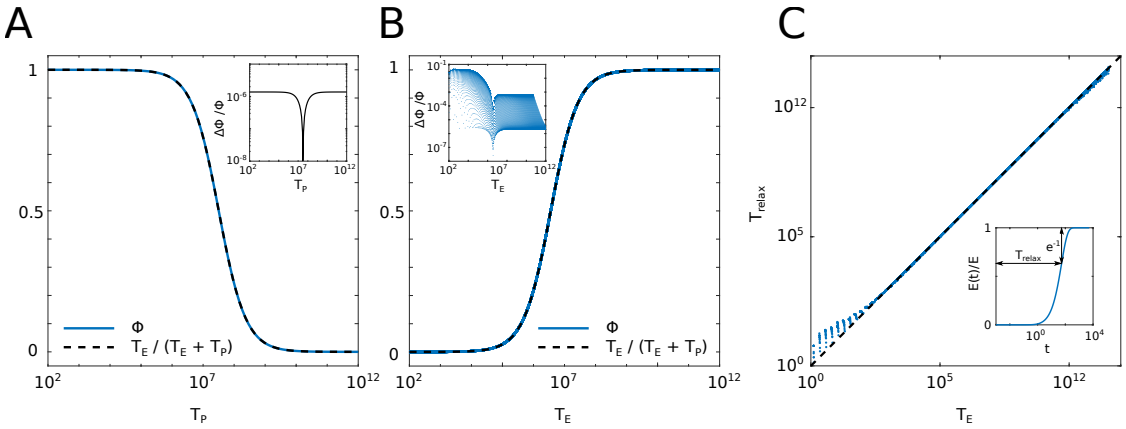

Figure 3.7: **Effective correlation time determines propagated noise and relaxation time.** (A) Propagated noise fraction as a function of protein lifetime T_P computed for parameters as model II in Fig 3.2A,C ($\alpha = 1.75 \cdot 10^4$, $k_{\text{link}} = 6.5 \cdot 10^{-2}$). The simple noise averaging function $T_E/(T_E + T_P)$, where T_E is the effective correlation time, provides an excellent approximation to the true propagated noise Φ. Indeed, the relative error (inset) remains small for the whole range of T_P. (B) Propagated noise fraction as a function of T_E computed by probing the whole parameter space $\alpha \in (1, 10^8)$ and $k_{\text{link}} \in (10^{-5}, 10^5)$, at fixed T_P. The approximation $T_E/(T_E + T_P)$ captures the true propagated noise well over the full range of sampled models, as the relative error (inset) never exceeds 10%. (C) Relaxation time T_{relax} as a function of effective correlation time T_E for the whole parameter space as in (B). We estimated T_{relax} from the temporal relaxation of the enhancer mean activity $E(t)$ to its steady state value E (inset). To this end, we first solve Eq. 3.31 with $E(0) = 0$ to obtain $E(t)$ for each model. We then estimated T_{relax} assuming $E(t)/E$ relaxes as $1 - \exp(-t/T_{\text{relax}})$, which is exact for the 2-state model. The resulting T_{relax} matches T_E well over the full range of sampled models. Thus our proposed effective correlation time T_E is a good predictor of both propagated noise and mean relaxation time. In all panels, we used $n = 3$ binding sites for the models.

Propagated noise and effective correlation time As we have shown above, all the m-state models lead to the same functional form for the mean and variance

(Eq. 3.36), differing only in the noise filtering function Φ_m. Although multiple time scales, given by the inverse spectrum of $-\hat{M}_r$, are involved in the noise filtering, we can define a single effective switching correlation time T_E that preserves as well as possible the amount of propagated noise of the n-state model. We aim for a definition of T_E that is independent of the value of T_P, such that the resulting T_E characterizes the filtering for all T_P well. One way of proceeding is to realize that in the case of the 2-state model, $T_E = T_P$ implies $\Phi = 1/2$. We can thus use this property to define T_E such that $\Phi_m(T_P = T_E) = 1/2$. Based on Eq. 3.38, we can then solve the following equation to obtain an effective T_E

$$\frac{1}{2}\mathbf{v_p^t}\hat{S}\mathbf{v_p} = \mathbf{v_p^t}(\hat{I} - T_E\hat{M}_r)^{-1}\hat{S}\mathbf{v_p}. \tag{3.39}$$

With such an effective T_E, the filtering function Φ_m is well approximated by

$$\Phi_m(T_P) \simeq \Phi(T_P/T_E) = \frac{T_E}{T_E + T_P}, \tag{3.40}$$

which is exact when $T_P = T_E$ and only slightly deviate from $\Phi_m(T_P)$ when $T_P > T_E$ or $T_P < T_E$, (Fig 3.7A,B).

Consequently, for all the enhancer models the propagated noise N^2 at the protein level is well approximated by

$$N^2 = \frac{1-E}{E}\frac{T_E}{T_E + T_P}. \tag{3.41}$$

In addition, the effective T_E provides an excellent approximation for the mean relaxation time-scale of the models (Fig 3.7C).

Effect of α and k_{link} on regulatory phenotypes

To understand how varying α and k_{link} affects regulatory phenotypes, we have a look at the phenotypes in the phase space $(\alpha, k_{\text{link}})$.

Fig 3.8 shows the dependence of main regulatory phenotypes on $(\alpha, k_{\text{link}})$ for fixed concentration (left column) and fixed expression (right column). For fixed concentration, there exists only a narrow range that gives high specificity − around the point where specific expression is already large enough (~ 1) while non-specific expression is still small ($\ll 1$). There, both residence time and sensitivity take

relatively low values.

Meanwhile, for fixed expression, specificity increases with larger α and lower k_{link}. This is due to the fact that with increasing α, the concentration required to reach fixed specific expression decreases, thus ensuring that non-specific expression stays low.

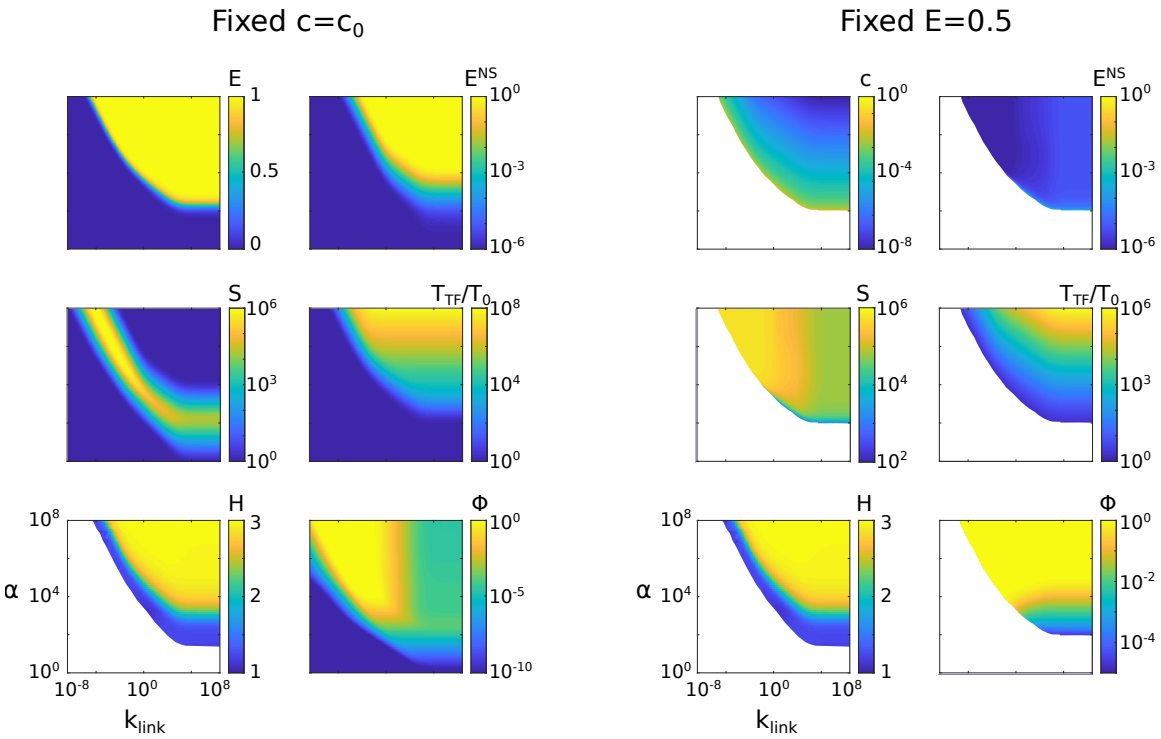

Figure 3.8: **Phase space of all regulatory phenotypes.** Expression from cognate enhancers containing $n = 3$ specific sites, E^{S}, and random DNA with nonspecific sites, E^{NS}, the TF residence time T_{TF}, specificity S, and sensitivity H (in color) as a function of two parameters, α and k_{link}. Left and right column are showing regulatory phenotypes at fixed concentration $c = c_0$ and at fixed expression $E = 0.5$, respectively. For fixed expression, regions where $E = 0.5$ cannot be satisfied are colored white. Due to numerics, area of sensitivity H where maximum expression (in the limit $c \to \infty$) is below $E < 10^{-3}$, is also colored white.

Effect of the unbinding rate ratio $k_-^{\text{S}}/k_-^{\text{NS}}$ on the specificity gain

In the main text we investigated how the maximum gain in specificity, $S_{\text{NEQ}}/S_{\text{EQ}}$, varies with the ratio of specific and non-specific unbinding rate, k_-^{S} and k_-^{NS}, (Fig 3.3C). We identified an optimal value of $k_-^{\text{S}}/k_-^{\text{NS}}$ which maximizes this gain.

For smaller values of $k_-^{\text{S}}/k_-^{\text{NS}}$, NEQ model is at the bound of specificity, S_{max}, with EQ

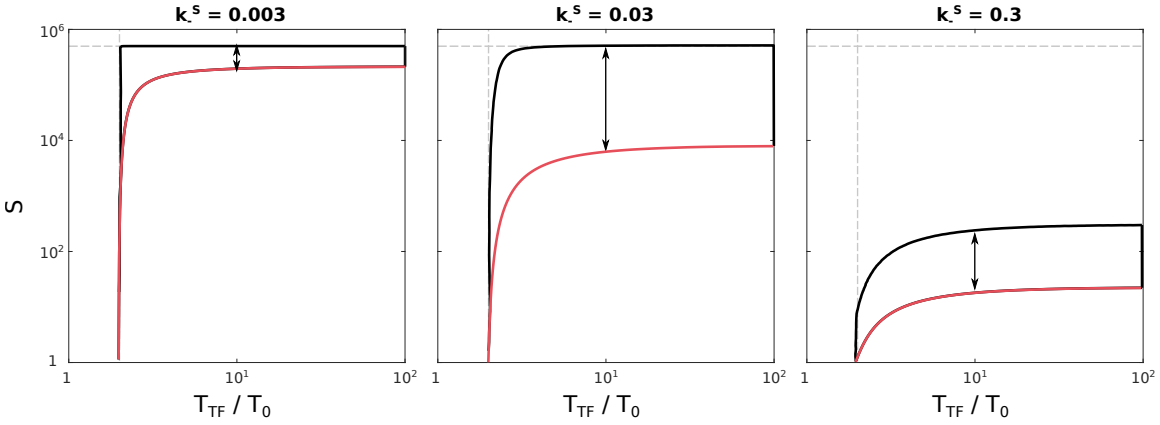

Figure 3.9: **Specificity gain is effected by unbing rates.** Specificity as a function of TF residence time at fixed $E = 0.5$, showing how specificity gain (black arrow) changes for different values of k_-^{S} with fixed $k_-^{\mathrm{NS}} = 1$ at $n = 3$. EQ model solutions lie on the red line while the black/red envelope represents the space of solutions for NEQ model. The reference time is $T_0 = 1/k_-^{\mathrm{S}}$ which varies between the three figures. Dashed lines represent minimum residence time and maximum specificity.

model being close to it as well: see Fig 3.9 left. With increasing $k_-^{\mathrm{S}}/k_-^{\mathrm{NS}}$, the specificity of EQ model decreases, leading to an increase in the specificity gain (Fig 3.9 middle). The largest specificity gain is obtained when the maximum specificity in NEQ model is not bounded anymore but very close to it. With further increasing $k_-^{\mathrm{S}}/k_-^{\mathrm{NS}}$, the difference in specificity between NEQ and EQ model starts to decrease (Fig 3.9 right).

3.4.2 SI Figures

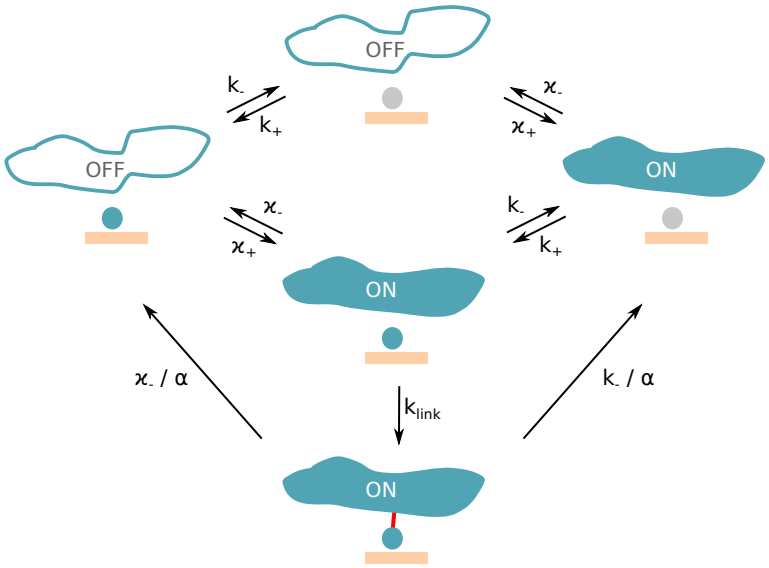

Figure 3.10: **Kinetic scheme of the non-equilibrium MWC-like model.** For simplicity, the scheme is illustrated for a single binding site, $n = 1$. TFs can bind to the specific binding site and Mediator can switch into `ON` state; when a TF is bound it can form a link only if a Mediator is found in `ON` state. The link decreases the unbinding rate of both linked TF and Mediator by a factor of α. The link is removed either when the Mediator switches `OFF` or when the linked TF unbinds. With increasing number of binding sites n, the number of possible states exponentially increases.

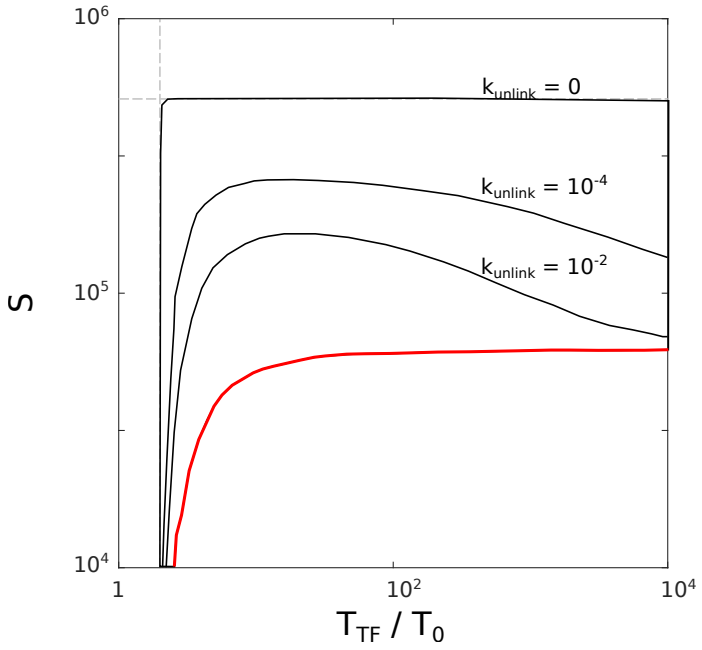

Figure 3.11: **Specificity effect of the nonzero unlinking rate.** The phase diagram of specificity, S, and mean TF residence time, T_{TF}, for $n = 3$ binding sites and fixed $E = 0.5$, demonstrating the effect of unlinking rate. With increasing unlinking rate, the maximum specificity of the nonequilibrium model decreases. Furthermore, for non-zero unlinking rate, $k_{\mathrm{unlink}} > 0$, at larger TF residence times, the maximum specificity starts to decrease with TF residence time. This is qualitatively different than in case of a zero unlinking rate where maximum specificity never decreases with T_{TF}. Each black envelope shows all solutions for varying $\alpha \in (1, 10^8)$ and $k_{\mathrm{link}} \in (10^{-5}, 10^{-8})$. Red curve represents equilibrium solutions at $k_{\mathrm{link}} \to \infty$, which do not vary with k_{unlink}. We used $k_-^S = 0.01$ and $k_-^{NS} = 1$. See also Fig 3.18D.

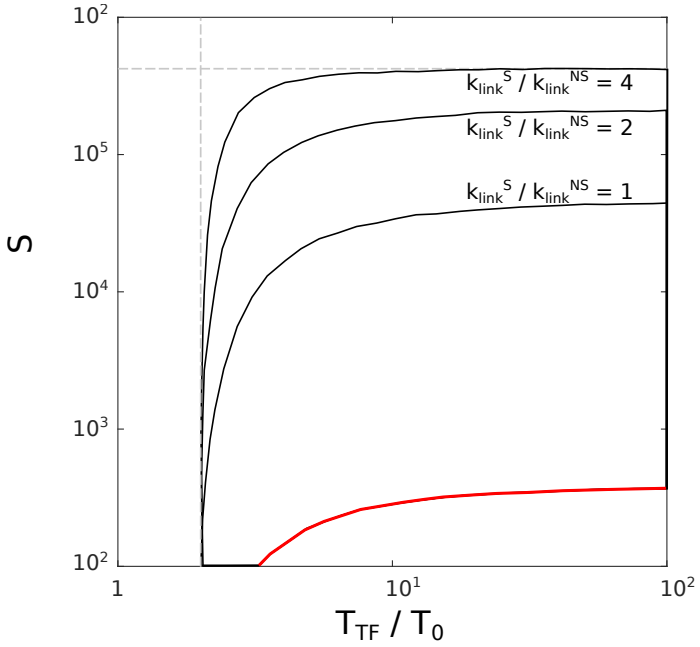

Figure 3.12: **Specificity gain due to sequence-specific linking rate.** The phase diagram of specificity, S, and mean TF residence time, T_{TF}, for $n = 3$ binding sites and fixed $E = 0.5$, demonstrating the effect of different sequence-specific linking rates. We assume that the formation of link on a TF bound to a specific site is faster, by the indicated factor, $k^{\mathrm{S}}_{\mathrm{link}}/k^{\mathrm{NS}}_{\mathrm{link}}$, relative to the link formation when the TF is bound to a nonspecific site; this could happen, for instance, if the links are created by dedicated enzymes with their own DNA sequence binding preference. Large specificity increases are possible even at $k^{\mathrm{S}}_{\mathrm{link}}/k^{\mathrm{NS}}_{\mathrm{link}}$ not much larger than 1. We used $k^{\mathrm{S}}_{-} = 0.1$ and $k^{\mathrm{NS}}_{-} = 1$ (instead of the $k^{S}_{-} = 0.01$ used elsewhere). The ratio of $k^{\mathrm{S}}_{\mathrm{link}}/k^{\mathrm{NS}}_{\mathrm{link}} = 1$ represents the case in the main chapter, without any linking sequence specificity. Each black envelope shows all solutions for varying $\alpha \in (1, 10^8)$ and $k_{\mathrm{link}} \in (10^{-5}, 10^{-8})$. Red curve represents equilibrium solutions, which do not vary with $k^{\mathrm{S}}_{\mathrm{link}}/k^{\mathrm{NS}}_{\mathrm{link}}$. Gray dashed lines show the analytically-derived bounds. The increase in S due to linking specificity is seen only for ratios of $k^{\mathrm{S}}_{-}/k^{\mathrm{NS}}_{-}$ that are smaller that the optimal value (the ratio where $S_{\mathrm{NEQ}}/S_{\mathrm{EQ}}$ reaches a maximum, see Fig 3.3C). The reason is that the linking rate specificity affects only the nonequilibrium models, and for lower values of $k^{\mathrm{S}}_{-}/k^{\mathrm{NS}}_{-}$, the NEQ models already reach the maximum possible specificity, κ_{-}/κ_{+}; see Fig 3.9. This means that for low values of $k^{\mathrm{S}}_{-}/k^{\mathrm{NS}}_{-}$, any additional linking specificity would not be able to increase NEQ enhancer specificity any further as it is already saturated. Linking specificity does not qualitatively change the overall conclusions, but can quantitatively boost specificity S.

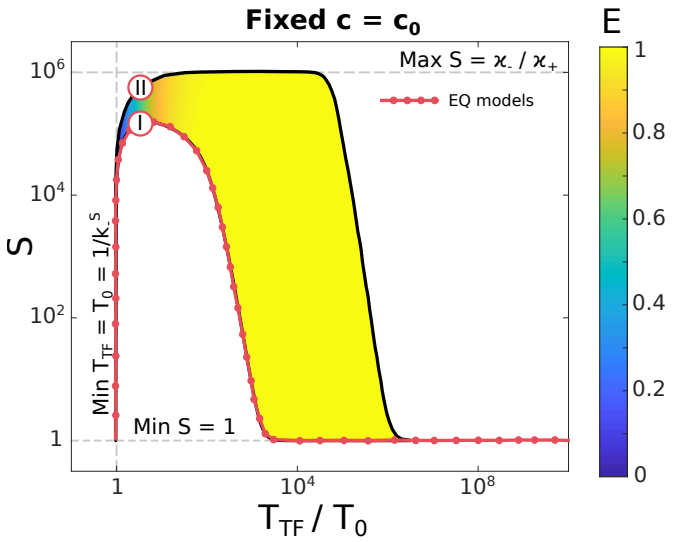

Figure 3.13: **Accessible space of regulatory phenotypes is similar for different number of binding sites.** Specificity, S, mean TF residence time, T_{TF} (expressed in units in inverse off-rate for isolated TFs at their specific sites, $T_0 = 1/k_-^S$), and average expression, E (color), for MWC-like models with $n = 5$ TF binding sites (main text Fig 3.2A showing $n = 3$), obtained by varying α and k_{link} at fixed TF concentration, c_0. Equilibrium models fall onto the red line. As in the main text, two models with equal TF residence times, `I` (EQ) and `II` (NEQ), are marked for comparison. Dashed gray lines show analytically-derived bounds. The EQ model reaches higher specificity than for $n = 3$, making the space of solutions for $E < 1$ smaller. NEQ model solutions become limited by specificity ceiling, $S_{\mathrm{max}} = \kappa_-/\kappa_+$, and $S > 1$ solutions only exist for $T_{\mathrm{TF}}/T_0 < 10^6$ (for $n = 3$ this was true for $T_{\mathrm{TF}}/T_0 < 10^8$). Nevertheless, the accessible space of regulatory phenotype is qualitatively preserved for $n > 3$.

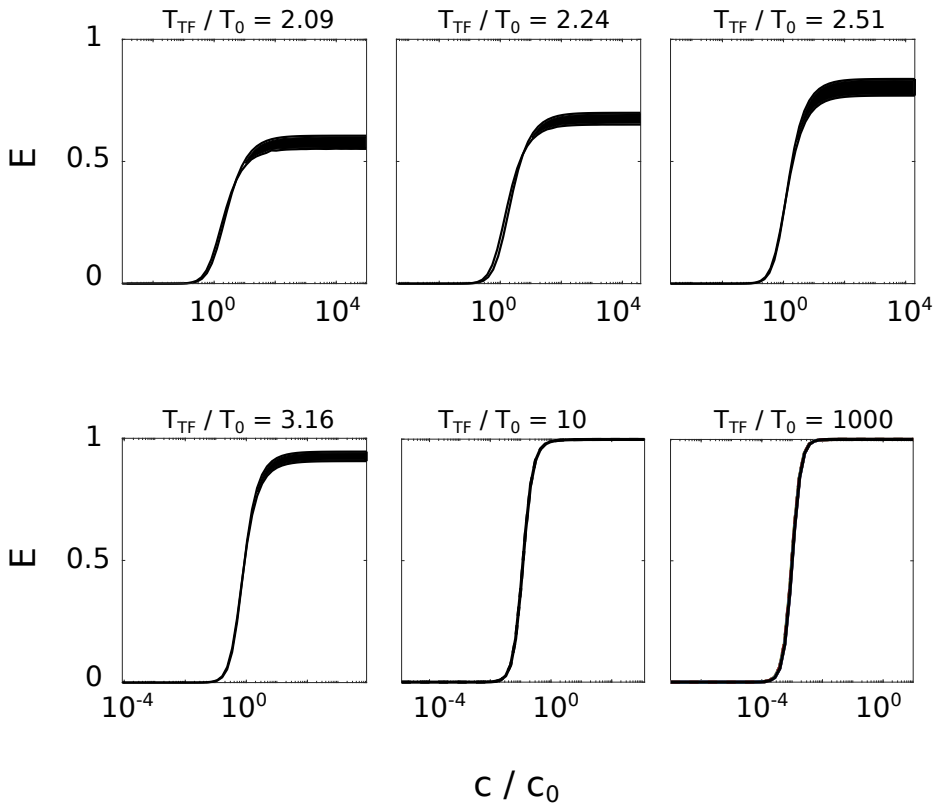

Figure 3.14: **Different models lead to indistinguishable induction curves from functional enhancers.** Induction curves for expression from functional enhancers (that contain specific binding sites) at fixed TF residence times (as indicated in the plot titles of different plots), for $n = 3$ and $E = 0.5$. In each plot with a given TF residence time, we find 20 different models with a range of specificities S, including the equilibrium model, and overlay their induction curves in black. Induction curves are nearly indistinguishable, with largest differences found for low residence times at large concentrations. The minimal achievable TF residence time is min $T_{\mathrm{TF}} = 1.98T_0$.

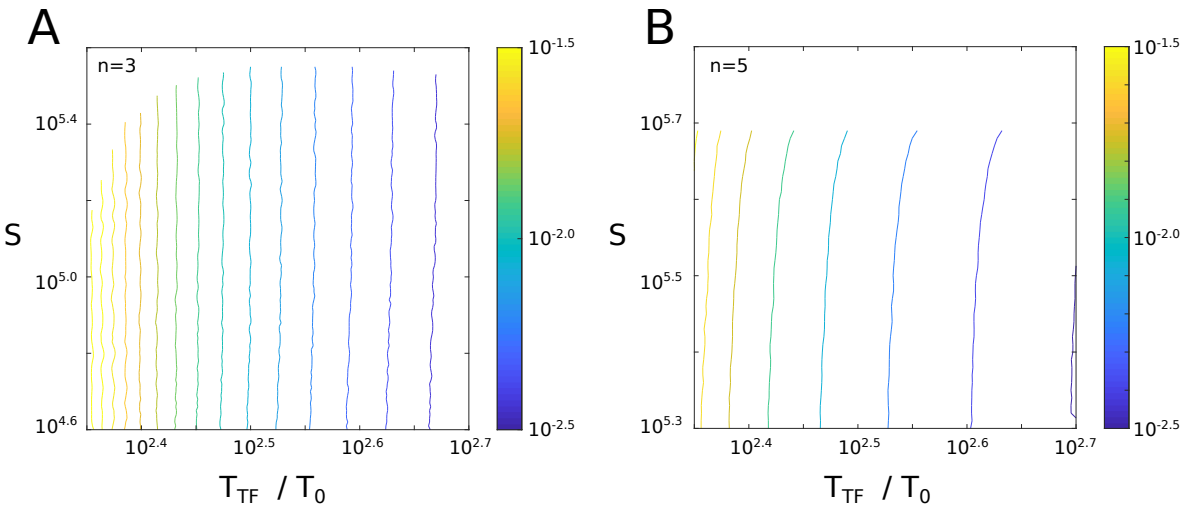

Figure 3.15: **Equi-concentration lines in enhancer phase diagrams at fixed expression are nearly vertical.** Lines of constant concentration (color) as a function of specificity and TF residence time for $n = 3$ (A) and $n = 5$ (B). Since all NEQ and EQ models have almost identical induction curves (Fig 3.2E and Fig 3.14), this implies that lines of constant concentration at fixed expression in the S vs T_{TF} space (Fig 3.2C), are nearly vertical. This assumption holds very well for smaller values of T_{TF}. With increasing residence time, lines of constant concentration start slightly tilting towards larger residence times. Additionally, for large specificity (close to the maximum specificity κ_-/κ_+) lines of constant concentration start to increase their curvature, especially at higher n.

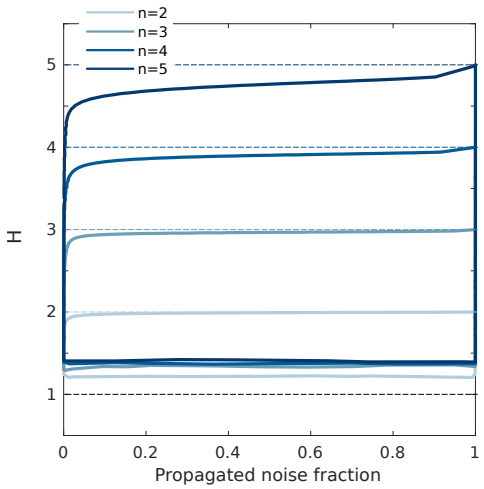

Figure 3.16: **Sensitivity and propagated noise fraction are uncorrelated.** Phase diagrams of sensitivity H and propagated noise fraction, $T_E/(T_E + T_P)$ (see main text). Each envelope represents different value of n (blue shade, legend); different models within each envelope are obtained by varying $\alpha \in (1, 10^8)$ and $k_{\mathrm{link}} \in (10^{-5}, 10^8)$, holding expression fixed at $E = 0.5$ by adjusting TF concentration. Almost all combinations of sensitivity and noise fraction are possible, indicating that these regulatory phenotypes are largely uncorrelated. The exception are highest possible sensitivities that are accessible only at higher values of propagated noise fraction.

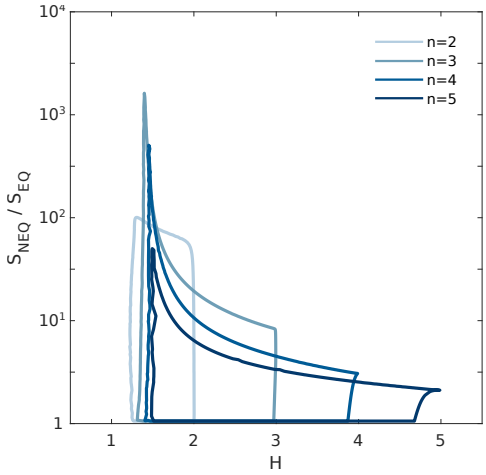

Figure 3.17: **Trade-off between optimal specificity and sensitivity.** Phase diagrams show the space of solutions of specificity gain, $S_{\mathrm{NEQ}}/S_{\mathrm{EQ}}$, and sensitivity, H. The envelopes were obtained by varying $\alpha \in (1, 10^8)$ and $k_{\mathrm{link}} \in (10^{-5}, 10^8)$ at fixed expression $E = 0.5$ and various number of binding sites n (blue shade, legend). Specificity gain is highest at lower residence times where the sensitivity is the lowest (Fig 3.3A). These results do not qualitatively vary with number of binding sites n, and show a trade-off between optimal specificity and high sensitivity gain.

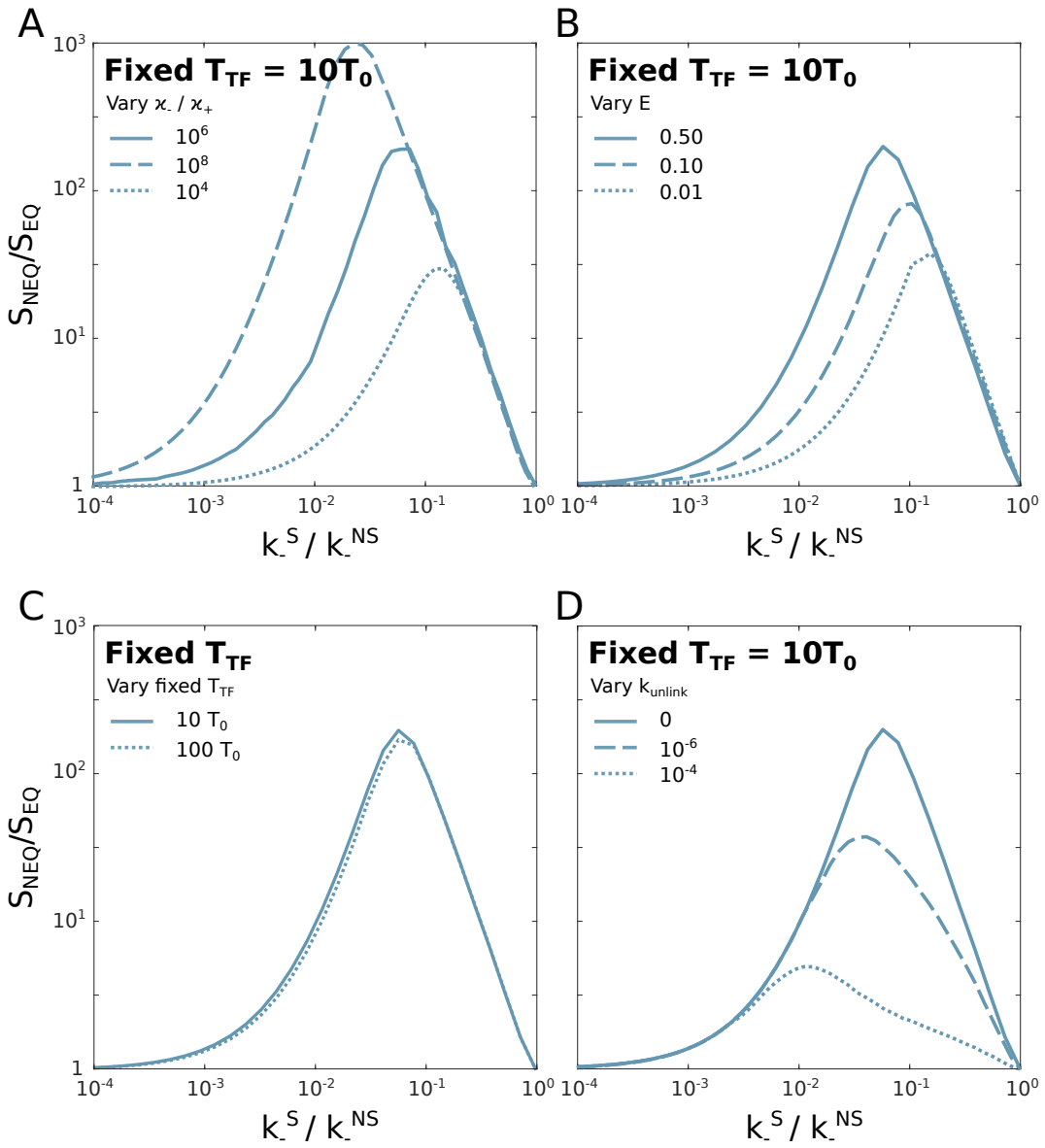

Figure 3.18: **Impact of kinetic and phenotypic parameters on optimal specificity gain.** Maximum gain in specificity as a function of a ratio of specific and non-specific unbinding rates, $k_-^{\mathrm{S}}/k_-^{\mathrm{NS}}$. Different plots show effects of different parameters that were varied: (**A**) ratio of Mediator switching rates κ_-/κ_+; (**B**) fixed expression E; (**C**) TF residence times at which the specificity is compared; and (**D**) unlinking rate k_{unlink}. The strongest dependence is on Mediator switching rates, which set the upper bound for specificity, S_{max}; when that increases, the maximum specificity gain also increases. Additionally, the value of fixed expression E has a visible impact, similar as varying κ_-/κ_+. The residence time at which we compare specificity has a negligible role; for almost any value of the residence time, specificity gain does not vary much. This is due to the fact that for $T_{\mathrm{TF}} \gg T_0$, both EQ and maximum NEQ specificity do not significantly vary with T_{TF}. Results show that for specificity gains right of the peak (i.e., larger k_-^S/k_-^{NS}), the ratio of Mediator switching rates κ_-/κ_+ does not play a visible role. The same is true for fixed expression E, and the TF residence time at which we compare specificity. When it is non-zero, k_{unlink} can strongly affect the specificity gain. For optimal gain, k_{unlink} must be much smaller than all other rates in the system.

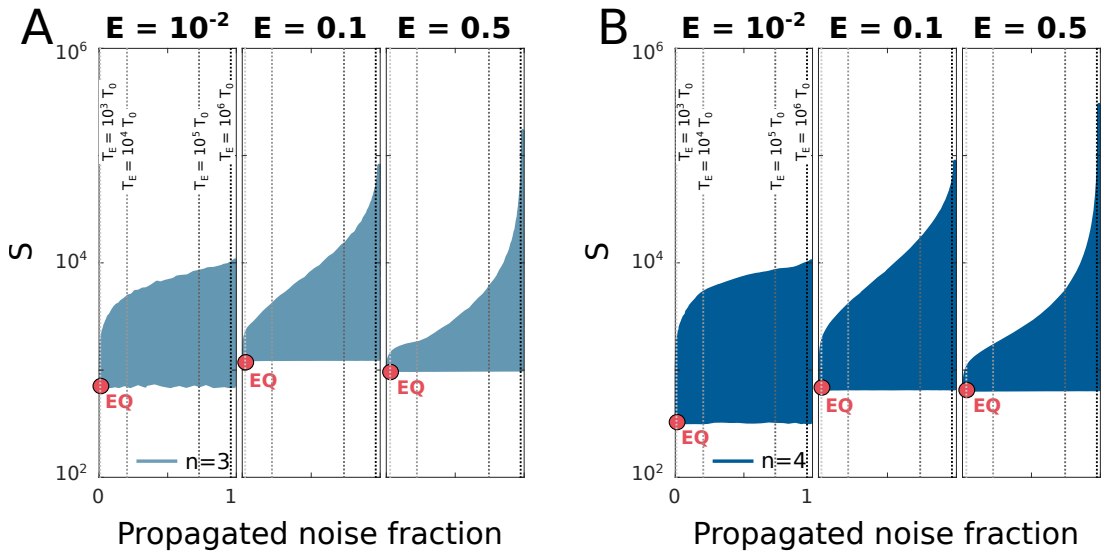

Figure 3.19: **Trade-off between optimal specificity and propagated noise.** Phase diagram of enhancer models for three different values of mean expression, E (columns), shows specificity S and fraction of variance in enhancer switching propagated to expression noise ($T_E/(T_E + T_P)$, see main text). Compact blue region for each E shows all MWC-like models with $n = 3$ (**A**) and $n = 4$ (**B**) binding sites accessible by varying $\alpha \in (1, 10^8)$ and $k_{\text{link}} \in (10^{-5}, 10^5)$; equilibrium model ("EQ") with lowest noise is shown as a red dot. We have picked the k_-^S/k_-^{NS} ratio to maximize the specificity gain S_{NEQ}/S_{EQ}: $k_-^S/k_-^{NS} \approx 0.06,\ 0.12$ for $n = 3,\ 4$, respectively (see Fig 3.3C). With these values, the specificity can increase but only at a cost of a large increase in correlation time T_E, implying high noise.

4 Evolving complex promoters for complex phenotypes

Understanding how genotype determines phenotype has been a long-standing goal in evolutionary biology. Although genotype-phenotype (GP) maps have been extensively studied, current approaches suffer from some combination of three major shortcomings: (i) a majority of experimental work focused on a neighbourhood of only a handful of mutations away, making these descriptions local; (ii) they do not go beyond toy models, to fit the data and give predictions; or (iii) they consider very simple and thus unrealistic genotypes and phenotypes. While studies focused on at most two of these points, no work includes a combination of all three. Therefore, biophysically realistic GP maps that give global predictions describing complex promoters are still lacking. Here, we investigate complex promoters and complex phenotypes in realistic setting. We go beyond the typically studied single phenotype of a constitutive promoter and study how mutations in bacterial promoters alter gene expression dynamics between different environments. We developed a biophysically realistic and mechanistic model that accurately predicts GP mapping for gene expression in a regulated bacterial promoter. Using this model, we show how promoter architecture, molecule concentrations, and transcription factor binding affinities constrain GP mapping and evolutionary trajectories of promoters. Furthermore, we show the need to account for the biophysical mechanisms that govern the GP mapping to correctly predict, and hence understand, evolution.

Grah R, Lagator M, Guet CC, Tkačik G. Evolving complex promoters for complex phenotypes. Manuscript in preparation.

Contributions: Grah R has constructed the model and has done all computations. Lagator M has done all experimental measurements. Grah R and Lagator M have done data interpretation.

4.1 Introduction

Mutations are the raw materials of evolution. They can alter the fitness of an organism, enabling selection to act on such changes in order to drive evolution. Therefore, understanding and predicting evolution requires the ability to predict how mutations alter organismal fitness, and doing so requires understanding how mutations alter specific traits [Dean and Thornton, 2007]. The effects of genetic mutations (genotype) on one or more organismal traits (phenotype) has been the central problem of evolutionary biology ever since Mendel's work questioned Darwin's notion of gradual evolution. By affecting what traits emerge in different genetic backgrounds, Genotype-Phenotype (GP) mapping impacts organismal development and function, influences the emergence of genetic disorders and diseases, and shapes how populations evolve and respond to selection [Alberch, 1991; Lehner, 2013; Houle *et al.*, 2010]. GP mapping has been extensively studied in a range of experimental and theoretical systems, most of which indicate that the mapping is complex and non-linear [Kemble *et al.*, 2019; de Visser and Krug, 2014; Wagner and Zhang, 2011; Hansen, 2006]. And yet, the wealth of experimentally determined maps has not resulted in comprehensive or generalizable understanding of the relationship between genotype and phenotype. In other words, we lack the ability to predict how genotype maps onto phenotype for most biological systems.

One major area of focus for describing GP mapping has been the regulation of gene expression, due to its central role in enabling organisms to respond to environmental change and to coordinate inter-cellular processes. Structures of numerous gene regulatory networks (GRNs) have been empirically determined, enabling the development of empirical GP maps [Payne and Wagner, 2014; Aguilar-Rodríguez *et al.*, 2017]. These maps discovered some fundamental properties of GP maps, at least as they apply to GRNs: the maps tend to be 'small world' [Watts and Strogatz, 1998] so that traversing a wide range of phenotypes is possible with only a handful of mutations, even though genotypes with similar phenotypes tend to be clustered closer together.

While offering unprecedented insights into how GRNs evolve, the existing GP maps have a fundamental limitation – they are largely based on experimental studies that focused on measuring steady-state expression levels in cells [Karlebach and Shamir, 2008; Kinney and McCandlish, 2019; Shultzaberger *et al.*, 2012; Kim *et al.*, 2009]. And yet, temporal dynamics of gene expression play an important role in determining how a biological system functions [Yosef and Regev, 2011; López-Maury *et al.*, 2008; Longo and Hasty, 2006]. For example, bistable behavior observed in various bacterial species is often enabled by having different rates at which relevant genes are turned on or off in response to a stimulus [Dubnau and Losick, 2006]. Similarly, the stress response in *S. cerevisiae* involves a temporary alteration to the global gene expression patterns, during which genes are rapidly turned on and off often without reaching steady-state expression levels. Yeast cells alter their response depending on the source of stress in order to optimize global gene expression dynamics, indicating that gene expression dynamics affect organismal fitness [Gasch *et al.*, 2000]. Cascades of genes that are critical for eukaryotic embryo development have highly optimized expression dynamics, and many transcription factors involved in embryo development never reach steady state expression [Arbeitman, 2002]. These examples highlight that not only the steady state expression levels, but also the expression dynamics (how rapidly the steady state is reached) affect organismal fitness [Bar-Joseph *et al.*, 2012; Ueda *et al.*, 2004]. In spite of this, understanding how mutations in gene regulatory elements (promoters and transcription factors) alter gene expression dynamics has received little attention.

Although previous work investigated GP mapping, they mostly addressed single phenotypes in a single environment [Lagator *et al.*, 2017b; Kinney and McCandlish, 2019; Kim *et al.*, 2009; Shultzaberger *et al.*, 2012; Karlebach and Shamir, 2008]. Alternatively, description of temporal dynamics of gene regulation represents a system which combines multiple environments (ON, OFF) and multiple phenotypes (dynamical phenotypes). Additionally, exploring such a system in a biophysically

realistic setting, allows for a connection between molecular mechanisms and key evolutionary quantities, such as distribution of mutational effects or evolutionary trajectories.

In this work, we set out to study how mutations in bacterial promoters alter gene expression dynamics. To achieve this goal, we developed a mechanistic model that can accurately predict GP mapping for gene expression dynamics in a regulated bacterial promoter. The model allowed us to understand how the mechanisms of promoter function constrain GP mapping, how those constraints changed depending on whether we considered only steady-state expression or the dynamics of expression, and how they affect the evolutionary trajectories of promoters.

4.2 Results

4.2.1 Experimental system

In order to develop a mechanistic model that can predict GP mapping for gene expression dynamics, we focused on the canonical model system in bacterial genetics – the Lambda bacteriophage promoter P_L [Ptashne, 1986]. We focused on a relatively simple promoter, as such promoters make the fundamental building blocks of GRNs and, hence, provide a relevant starting point for understanding the forces that shape gene regulatory evolution. P_L is a repressible promoter: in the absence of the transcription factor CI, σ_{70}-RNA polymerase complex (RNAP for brevity) binds to the -10 and -35 sites with high affinity and leads to strong expression; when CI is present in the system, it cooperatively binds to its two operator binding sites, O_{R1} and O_{R2}, preventing RNAP binding due to direct binding site competition and repressing the expression (Fig 4.1). In the experimental synthetic system we used, we placed the cI gene under an inducible P_{TET} promoter on the same small copy number plasmid (pZS*, with only 2-3 copies) as the P_L promoter, enabling external control of CI concentrations. The P_L promoter, which

controlled the expression of a yellow fluorescence marker (*yfp*) in our system, was also modified to exclude the OR3 site, and, with it, the P_{RM} promoter that is typically present on the reverse complement. The plasmid was placed in the MG1655 K12 strain of *Escherichia coli*, modified only to express the tetracycline repressor, TetR.

In our experiments, this system could exist in two environments, represented by two distinct states (Fig 4.1). In the ″ON″ state, cI is not present and hence only RNAP binding determines the expression levels of *yfp*. In the ″OFF″ state, cI is present at a high concentration, fully repressing the wild-type P_L promoter. In order to study the dynamics of gene expression in this system, we considered switching in both directions: from ″ON→OFF″ transition, and from ″OFF→ON″ transition. In other words, we would maintain the system under one condition (either ″ON″ or ″OFF″) for a sufficiently long time to ensure steady-state expression levels are reached. Then we induce the other state by either stimulating (″ON→OFF″) or stopping (″OFF→ON″) cI expression.

As temporal dynamics of gene expression represent a highly dimensional phenotype, a simpler representation of these trajectories is required. Since all trajectories are sigmoidal, we summarized the dynamics of gene expression through six distinct phenotypes (Fig 4.1): (i) steady state ″ON″ expression level; (ii) steady state ″OFF″ expression level; (iii) the duration of the lag when the system is switching from ″ON→OFF″, defined as the time from induction of the system to the point when the expression level is at half of the amplitude; (iv) lag when the system is switching from ″OFF→ON″; (v) the slope at half amplitude when the system is switching from ″ON→OFF″; (vi) the slope when the system is switching ″OFF→ON″. We also sometimes considered the amplitude (the difference between ″ON″ and ″OFF″ expression levels), but we did not treat it as a distinct phenotype. To avoid the obvious effect of amplitude on the slope (twice the amplitude would mean twice the slope), we rescaled the slope with the amplitude. This representation of temporal dynamics is by no means unique. However, having steady state phenotypes as a subset of all phenotypes allows us to directly compare steady state and dynamical

picture within the same setup.

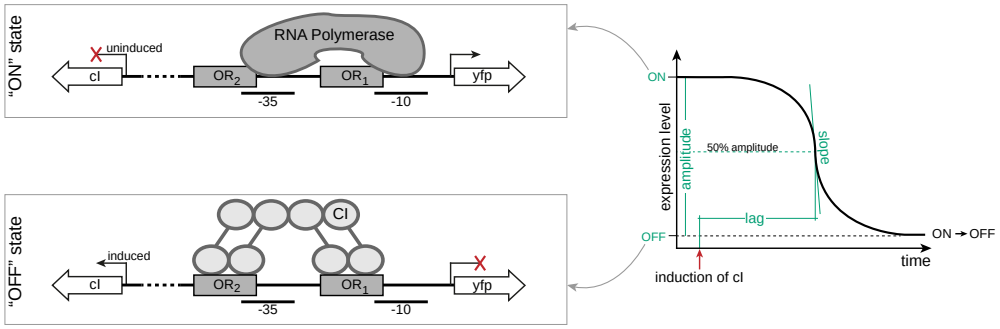

Figure 4.1: **Dynamics of gene expression are characterized with 6 phenotypes.** Binding of RNAP to -10 and -35 leads to expression of *yfp* (top) which we denote as "ON" state. When cI is induced, CI dimers can bind cooperatively to O_{R1} and O_{R2}, thus repressing the system, leading to "OFF" state (botom). Expression trajectory from "ON" into "OFF" state is shown on the right side with 2 steady-state (ON and OFF expression) and two dynamical (slope and lag) phenotypes marked. Slope is computed at half-amplitude expression, and lag time is defined as time between induction of cI and half-amplitude expression. Lag and slope for OFF to ON dynamics are defined in the same way. Steady-state expressions ON and OFF, together with lag and slope in both directions form 6 phenotypes which represent the dynamics of gene expression.

4.2.2 Combined model of gene expression dynamics

To describe the temporal dynamics of gene expression we combined two established modeling approaches – the thermodynamic model of steady state expression and the mass action kinetics (Fig 4.2A). First is the thermodynamic (TD) model, which describes the mapping between the genotype and the steady state expression [Bintu *et al.*, 2005a; Bintu *et al.*, 2005b; Shea and Ackers, 1985]. For a given promoter and the molecules that bind and regulate it, the TD model calculates the probability of finding the system in all possible states and then assumes that the steady state expression levels are proportional to the probability of finding the system in a productive state. By assumption, this is a state where RNAP is bound with mRNA being expressed at a fixed rate. In the model system used in this study (Fig 4.1), the system can be in four distinct states (Fig 4.2A): (i) no molecules bound to

the promoter; (ii) RNAP bound; (iii) repressor bound; (iv) two repressors bound cooperatively to both operators. From these four states, only (ii) is productive and leads to transcription. To calculate the energy of binding for each molecule that binds the P_L promoter, the TD model uses the energy matrices (EM) of RNAP and CI, as well as the strength of cooperative binding between two repressors (relevant only in state v). The energy matrix contains the information about how every possible point mutation in the DNA-binding site of a given molecule impacts its overall binding energy [Kinney *et al.*, 2010]. As such, each DNA-binding molecule has a unique energy matrix associated with it, which can be thought of as a unique representation of that molecule's function, much like the amino acids sequence is a two-dimensional representation of that molecules 3D structure.

The second approach, the mass action kinetics (MAK), uses standard ODEs to describe the temporal dynamics of different molecules in the system. MAK accounts for the changes in concentrations of the CI repressor and the measurable system output, YFP. While we assume a constant and high concentration of RNAP, the concentrations of CI and YFP change due to their variable production and dilution rates. MAK uses the probability of binding from the TD model as the rate of production of YFP (Fig 4.2A).

We obtained the energy matrices for RNAP and CI from published works [Lagator *et al.*, 2020]. To fit the other parameters in the TD model, and hence to predict steady-state expression levels of P_L promoter mutants, we used an existing Lambda P_L random mutant library [Lagator *et al.*, 2020]. To fit the MAK parameters, and hence to model the dynamics of the system, we only used measurements of the wild-type P_L system and did not rely on any promoter mutants (Section 4.4.2).

To validate the performance of this model, we created 9 P_L promoter mutants, predicted to affect the binding of RNAP and CI in different ways: (a) not to significantly affect the binding of either; (b) primarily impair RNAP binding; (c) primarily impair CI binding; (d) impair the binding of both, RNAP and CI. We measured the temporal dynamics of these mutants when switching from "ON→OF" and "OFF→ON", and found that our combined model predicted their gene expression dynamics extremely well (Fig 4.2B).

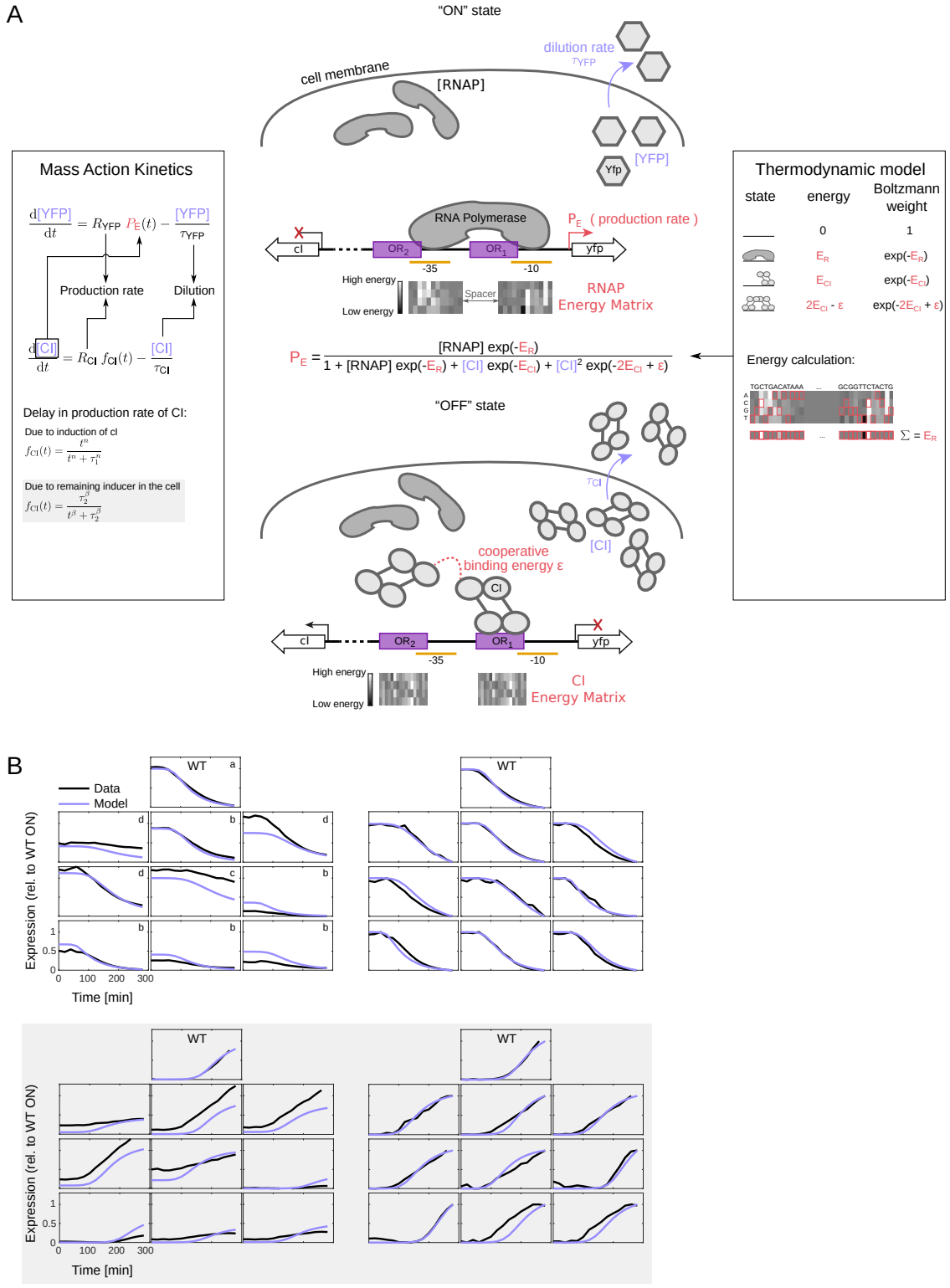

Figure 4.2: (Continued on the following page.)

Figure 4.2: **Combination of Thermodynamic and Mass action kinetics model gives accurate predictions of gene expression dynamics. (A)** Details of Mass Action Kinetics (left) and Thermodynamic model (right), with the representation of parameters used in the two models (middle). Mass Action Kinetics (MAK) model describes the change in CI and YFP concentration with a set of two ODEs; while the production of CI is delayed due to induction of cI (ON to OFF dynamics) or remaining inducer in the cell (OFF to ON dynamics), the production of YFP is determined via the Thermodynamic (TD) model. TD model description includes the system architecture (i.e., the possile binding states), their binding energies E and cooperativity ϵ between two CI dimers bound on O_{R1} and O_{R2}. The binding energies are determined using the energy matrix with binding energy being a linear sum of individual contributions from single base-pairs. The probability of RNAP bound, described by P_E, determines the rate of production of YFP. Each part of RNAP energy matrix consists of two parts, each 12bp long (see Section 4.4.1), with a spacer between them. **(B)** Comparison between experimental data and model prediction of gene expression dynamics for a wild-type and 9 promoter mutants. Mutations in each of the 9 mutants (a) have no significant impact, (b) impact RNAP binding, (c) impact CI binding, or (d) both. We mark the top right corner of each trajectories with (a-d) to show how mutations affect it. The left half shows expression dynamics with high agreement between data and model - Pearson correlation coefficient between data and prediction is $\rho_{ON\rightarrow OFF} = 0.92$ (top) and $\rho_{OFF\rightarrow ON} = 0.84$ (botom). Because some of these mutants had many mutations away from the wild-type (some containing as many as 10 mutations), it is not surprising that our predictions of steady-state expression levels are not ideal, as this is a known problem with the TD model [Vilar, 2010]. When correcting for these known errors of the TD model by setting the ON and OFF steady-state expression levels to 1 and 0, respectively, we find that our combined model predicts the dynamics of the system extremely well – $\rho_{ON\rightarrow OFF} = 0.98$ (top) and $\rho_{OFF\rightarrow ON} = 0.96$ (botom). The parameters for the model were obtained from independent calibration measurements, making this prediction fit-free. See Section 4.4.1.

4.2.3 Constraints in Genotype-Phenotype mapping

Starting with the wild-type Lambda P_L promoter, we used our model to exhaustively explore the effect of all possible single and double mutants on all six dynamical phenotypes (Fig 4.1). We were specifically interested in the constraints of different traits, asking whether mutations alter the phenotypes independently of each other or not. These constraints describe: (i) the size and shape of the accessible parameter space; (ii) the part of that space that is accessible by mutations; and (iii) correlations between pheontypes that detail how that space is covered by mutations. In other

words, these constraints limit the possible states that the system can adopt through mutation and define what phenotypes can be achieved through mutation, or, in other words, the distribution of phenotypic effects of mutations. To get a more complete picture of the constraints shaping GP mapping, we focused on double mutants (Fig 4.3), as they explore the two-dimensional phenotypic space more fully than the single mutants (Fig 4.9). While the constraints exist as a six-dimensional interaction between all measured phenotypes, we represent them as two-dimensional interactions between all possible pairs of phenotypes in order to better visualize them.

The GP map of the P_L promoter is heavily constrained, as double mutants explore only a portion of the possible landscape (Fig 4.3). This observation does not imply that the system is robust and that mutations cannot drastically alter one or more phenotypes. In fact, many double mutants have a large effect on the phenotypes (Fig 4.3). This finding goes against a common assumption of quantitative genetics – that small genetic changes (i.e. individual mutations) lead to small phenotypic changes [Milocco and Salazar-Ciudad, 2020]. While the observed constraints do not imply that the system is robust, they do set a limit to the possible phenotypic states that can be achieved. A more constrained system is less likely to lead to evolutionary innovations [Ciliberti *et al.*, 2007], as mutations result in a smaller set of possible phenotypic states, limiting the extent to which the system can explore the full, unconstrained phenotypic landscape. This is at least in part because the observed constraints point towards canalization in bacterial promoters [Wagner *et al.*, 1997], where the same value of a given phenotype can be achieved by many mutations that alter one or more other phenotypes. Because a more constrained system can assume a reduced number of possible phenotypic states, evolution is also more likely to be repeatable and to undergo the same pathways during the adaptive process [de Visser and Krug, 2014].

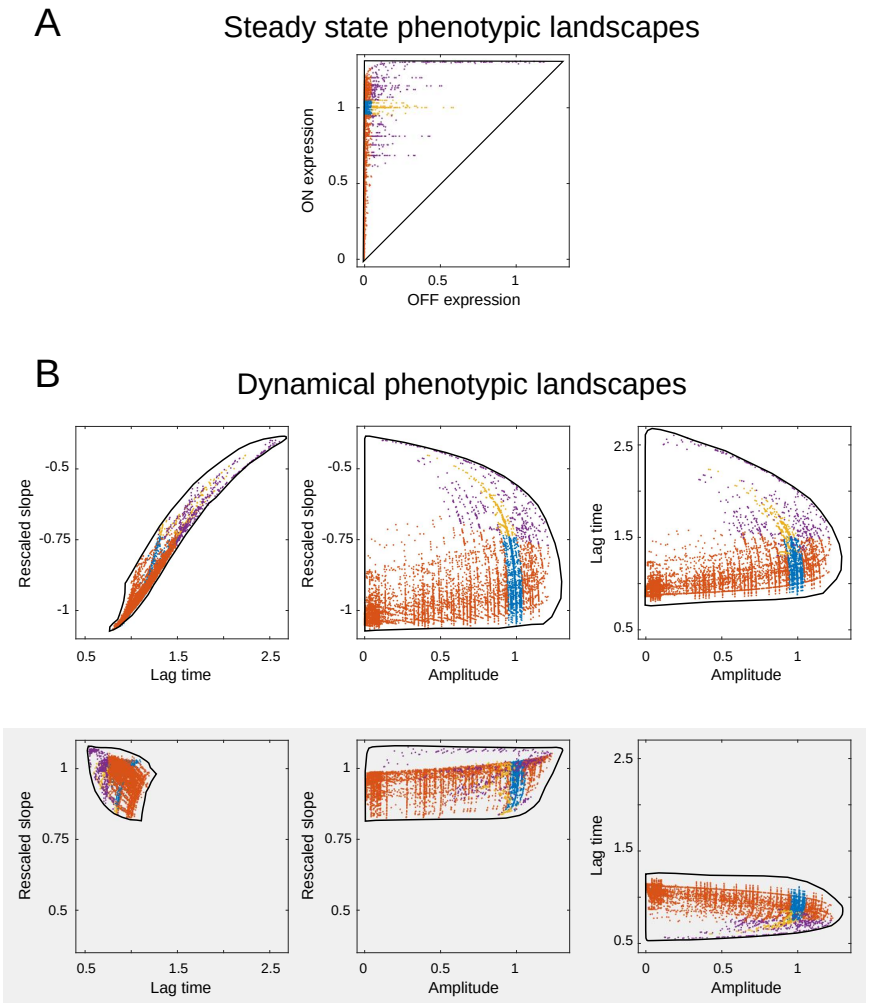

Figure 4.3: **Phenotypic landscapes are highly constraint. (A)** Phenotypic landscape of all double mutants, each represented by a dot, of steady-state phenotypes. **(B)** Phenotypic landscape of all phenotypes for ON to OFF dynamics (top) and OFF to ON dynamics (bottom, grayed). To keep the possible number of landscapes low, we use amplitude(=ON-OFF) as a proxy for both steady-state phenotypes (ON and OFF). As slope strongly depends on amplitude (twice the amplitude implies twice the slope), we use rescaled slope (i.e., slope/amplitude) as a phenotype. Black envelope represents the possible space that double mutants are theoretically able to explore, ignoring the constraints in architecture (e.g., overlap between binding sites), protein structure (energy matrix structure), or discrete space of genotypes. Different colors represent how steady-state phenotypes were affected: blue - WT-like mutants, red - change in ON expression >0.05, yellow - change in OFF expression >0.05, purple - change in both ON and OFF expressions > 0.05. All units are in wild-type units, with the exception of OFF expression – the expression (ON and OFF) is in the units of wild-type ON expression.

4.2.4 Mechanistic origins of constraints in dynamical phenotypes

The combined model that we used to describe the constraints in GP mapping of a bacterial promoter (Fig 4.3) also allowed us to understand the mechanistic origins of those constraints. Understanding not only what mutations do but also why is critical for developing a more predictive understanding of evolution, as it enables generalizing GP maps beyond a specific system being studied (in our case, the P_L promoter) to a range of other systems that share similar features (regulated bacterial promoters).

Looking at the combined model (Fig 4.2) identifies several key properties of the system that might impact the nature of GP mapping: (i) the factors that impact the concentrations of the relevant molecules in the system; (ii) the architecture of the promoter; and (iii) the factors that impact the binding energies of RNAP and CI. The factors that come from the MAK part of the model – (i) – are predominantly responsible for setting the limit to the phenotypes that can be achieved (black borders in Fig 4.3). The factors that influence the TD part of the model – (ii) and (iii) – primarily affect how freely mutations explore that space.

To summarize, the mass action kinetics of the system define the maximum values of phenotypes achievable through mutations, while the thermodynamics of binding define how easily those values are reached. This implies that, if all we are interested in is the maximum range of phenotypic values but not how easily those phenotypes can be realized by mutations, it is sufficient to represent the major aspects of the TD model (namely, the multi-variable energy matrices) through a small number of summary variables. We return to this point later in Section 4.2.6.

Molecule concentrations

The concentration of the CI repressor in the P_L system is affected by its production and dilution rates. The production rate is determined by presence of the inducer, while the dilution rate results from the combined effect of the cell division, transmembrane dilution, and protein degradation. In our model, the output of the system, YFP, has the same dilution rate as CI. We assume that RNAP concentration is always

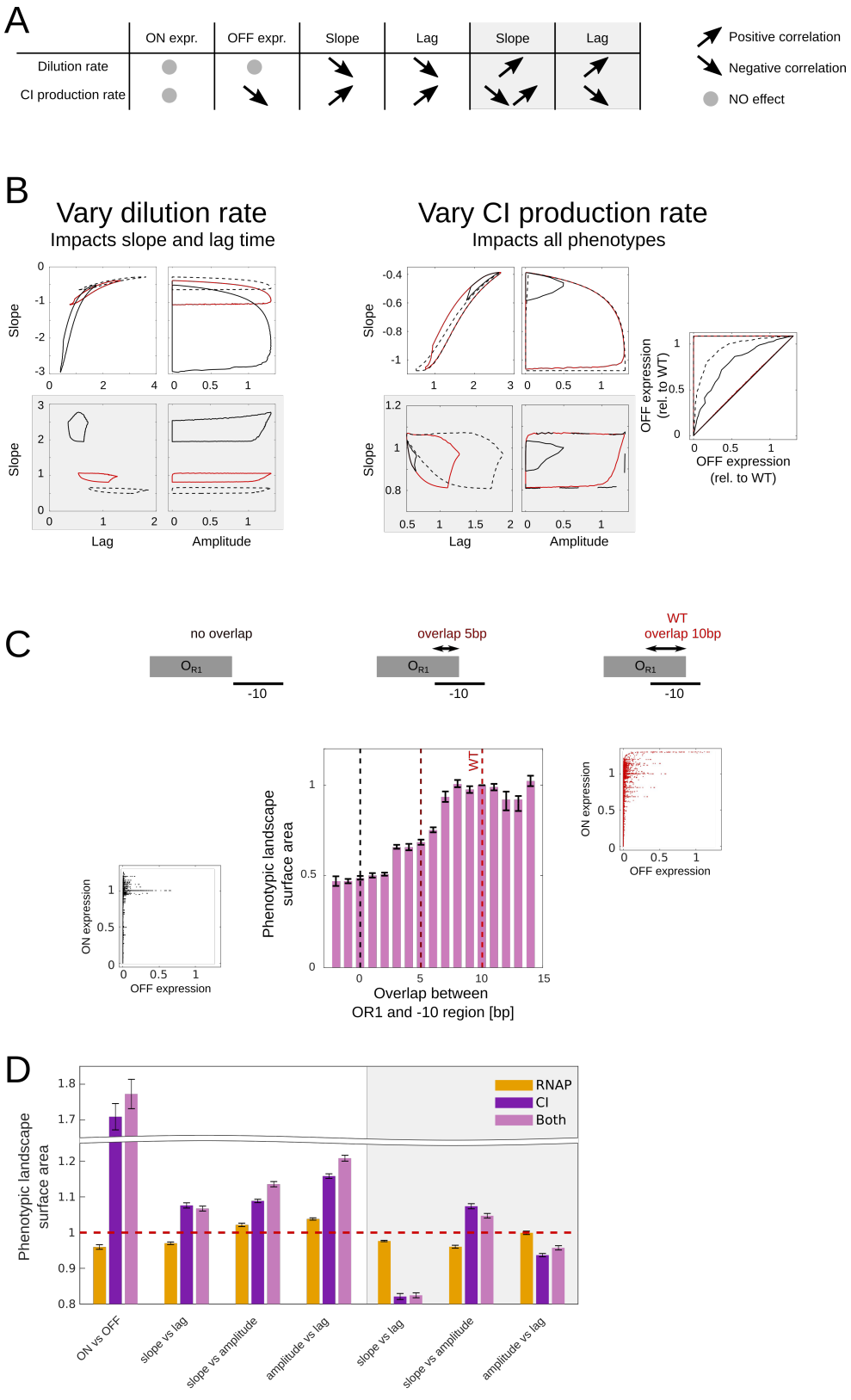

Figure 4.4: (Continued on the following page.)

Figure 4.4: **Mechanistic understanding of the phenotypic constraints. (A)** The effect of two important parameters in MAK model (dilution rate of CI and YFP, and CI production rate) on the individual phenotypes that double mutants around wild-type can explore. Both arrows are shown when an increase in a parameter leads to a significant increase in possible phenotype values in all directions (see Fig 4.11). **(B)** Similarly as in A, but showing how boundaries of phenotypic landscapes of double mutants change, showing the constraints between pairs of phenotypes more clearly. Full and dashed black envelope represent phenotypic landscape of increased and decreased parameter, respectively. For reference, red evelope shows the landscape with the original parameter value. Results for 'lag vs amplitude' look very similar to 'slope vs amplitude' - see Fig 4.12. As dilution time does not affect steady state dynamics, the ON vs OFF landscape is not affected and therefore not shown. Slope was rescaled with amplitude, i.e., slope/amplitude. All units are in wild-type units, with the exception of OFF expression – the expression (ON and OFF) is in the units of wild-type ON expression. **(C)** Overlap between -10 RNAP binding site and O_{R1} positively correlates with the size of phenotypic space of double mutants. The middle plot shows the sufrace area of ON vs OFF landscape as a function of number of overlapping base pairs. The surface area units are normalized to the wild-type (WT) overlap of 10bp. The energy matrix representation of RNAP also includes flanks of -10 binding site, thus wild-type having 10bp overlap with O_{R1} (see Fig 4.2A). The side plots shows comparison of ON vs OFF landscape for wild-type overlap of 10bp (right) and no overlap (left), demonstrating that decreasing one constraints (lower overlap) leads to increase of another constraints (lower space of possible phenotypes). **(D)** Surface area of all phenotypic landscapes with removed with-in correlations and structure of the energy matrices which represent protein structure. This is done by randomly shuffling the elements of the energy matrix, keeping the consensus sequence intact. Legend shows which energy matrix was shuffled. Units of surface area are normalized to the surface area of wild-type (non-shuffled) energy matrices - red dashed line. Error bars represent s.t.d. of 500 replicates. Note that y-axis does not start at zero. Grayed area shows results for dynamics OFF to ON. For details on surface area and overlap, see Section 4.4.1.

constant.

The CI production rate and the dilution rate impact most phenotypes individually and, in most cases, in a monotonic fashion (Fig 4.4A, Fig 4.11). When considering the constrains that emerge between pairs of phenotypes, the CI production rate and the dilution rate alter the limits of phenotypic values that can be achieved, but have a lesser impact on the specific nature of constraints (correlations) (Fig 4.4B). In other words, the concentrations of molecules in the system primarily affect the maximum range that each phenotype can achieve.

Promoter architecture

Each bacterial promoter has a specific architecture, determined by the relative position of RNAP and transcription factor binding sites in it. In the wild-type P_L promoter, there is one strong RNAP binding site consisting of the -10 and -35 feet, and two operators for the CI repressor, O_{R1} and O_{R2}. In the wild-type promoter, O_{R1} has a 10 base pair overlap with the -10 RNAP foot. This means that mutating those 10 positions in the promoter affects the binding of both, RNAP and CI simultaneously. In order to more clearly understand the role that promoter architecture plays in constraining the dynamical phenotypes, we considered a changing overlap of operator O_{R1} with the -10 foot of RNAP (see Section 4.4.3). We assumed that a given strength of CI binding is equally repressive, irrespective of the specific promoter architecture.

The critical property that changes as the overlap between the binding sites of two molecules changes is the number of positions that, when mutated, affect the binding of both instead of just one molecule. In other words, less overlap means more independent binding of each molecule in the system. We found that greater overlap decreased the constraints (Fig 4.4C), meaning that promoter architectures with more independent binding of RNAP and CI have a stronger correlation between phenotypes and hence could explore a smaller portion of the total phenotypic landscape surface area. This somewhat counter-intuitive finding stems from the fact that, when more than a single mutation emerges in the system (and here, we present all possible double mutant effects), greater overlap enables a higher possible number of phenotypic states to be assumed by the system. In other words, when there is no overlap, a point mutation can affect either the binding of CI or of RNAP, while with overlap it can affect the binding of one, the other, or, critically, both simultaneously.

Binding energies

The fundamental summary of a transcription factor's function is contained within its energy matrix, which describes the effect of every possible point mutation in the binding site on the energy of binding between the transcription factor and DNA.

Each energy matrix has its internal structure – between positions, some have a greater impact on binding than others; and within positions, some mutations alter the binding energy more than others. The energy matrix structure therefore implies a specific set of correlations between mutational effects..

We explored the extent to which the specific structure of RNAP and CI energy matrices (Fig 4.8) affected the constraints in dynamical phenotypes. To do this, we created 500 alternate energy matrices for both, RNAP and CI, in which we kept the wild-type sequence intact but shuffled randomly the specific entries in the matrix. Doing this altered the correlations between mutations that are inherent to the wild-type RNAP and CI energy matrices.

For most pairs of phenotypes, shuffled RNAP energy matrices decreased, while shuffled CI energy matrices increased the total surface area explored by mutations (Fig 4.4D). In other words, the wild-type RNAP imposes fewer constraints than one would predict based on randomized energy matrices, while the wild-type CI imposes greater constraints. This finding suggests that the internal structure of the energy matrices might have been selected for. RNAP is a molecule that requires flexibility in its binding, because it regulates the expression of >70% of all *E.coli* promoters [Salgado *et al.*, 2013]. Our results suggest that this functional requirement of RNAP is aided by the structure of its energy matrix, which can explore, and hence function, in a wider range of phenotypic states. For CI, which is supposed to bind only a few specific promoters, the energy matrix is more constrained than predicted. Hence, CI is less likely to lead to spurious binding, which can introduce fitness costs through binding to non-cognate promoters.

4.2.5 Evolution of regulated promoters

The mechanistic, predictive GP mapping for dynamical phenotypes provides the starting point for understanding how repressible promoters evolve. In population and quantitative genetics, evolution is typically described for a single, or, less frequently, for a pair of correlated phenotypes. Given the ability to accurately predict six phenotypes from a given genotype (Fig 4.1, Fig 4.2), we explore how evolution

proceeds when selection acts on six phenotypic dimensions.
Specifically, using the combined model we can describe several fundamental properties that affect how repressible promoters evolve, especially focusing on: (i) the distribution of phenotypic and fitness effects of the whole genotypic space and evolving promoters; (ii) the evolutionary trajectories and dynamics of populations that are evolving a repressible promoter; (iii) the outcomes of selection for repressible promoter function; and (iv) emerging simplified fitness landscapes that nevertheless correctly capture trait evolution.

Distribution of phenotypic and fitness effects of random sequences

Any random sequence can, in principle, act as a repressible promoter, if it binds RNAP and a repressor. We examined how the phenotypic effects of random sequences were distributed, in order to understand how likely a random sequence is to bind RNAP and CI. To explore the phenotypic effects of random sequences, we sampled $2 \cdot 10^9$ random 80 base-pair long sequences. We found that for all six phenotypes, functional sequences were very rare with most random sequences being non-functional (Fig 4.5A), as previously observed for individual promoters [Kinney *et al.*, 2010; Maerkl and Quake, 2007] and proteins [Maerkl and Quake, 2009; Jacquier *et al.*, 2013].
We also wanted to describe the effects of random sequence on the fitness of the organism. To describe the fitness of a given sequence, we adopted quadratic fitness landscape – one of commonly used models in evolutionary biology.
Using this model, we assigned a fitness value to each of the $2 \cdot 10^9$ random sequences in order to characterize the distribution of fitness effects (DFE) of the entire genotypic space. Most random sequences had low fitness, as most random sequences had neither any expression nor the ability to bind the repressor (Fig 4.5B). Functional repressible promoters (those with phenotypes at least somewhat similar to the wild-type P_L promoter) were extremely rare, occurring with probabilities of $10^{-5} - 10^{-7}$. Such a distribution of fitness effects were frequently observed [Jacquier *et al.*, 2013; Sanjuán *et al.*, 2004; Duveau *et al.*, 2017; Metzger *et al.*, 2016], although typically the observed distributions were smoother. This difference most likely arose from the fact

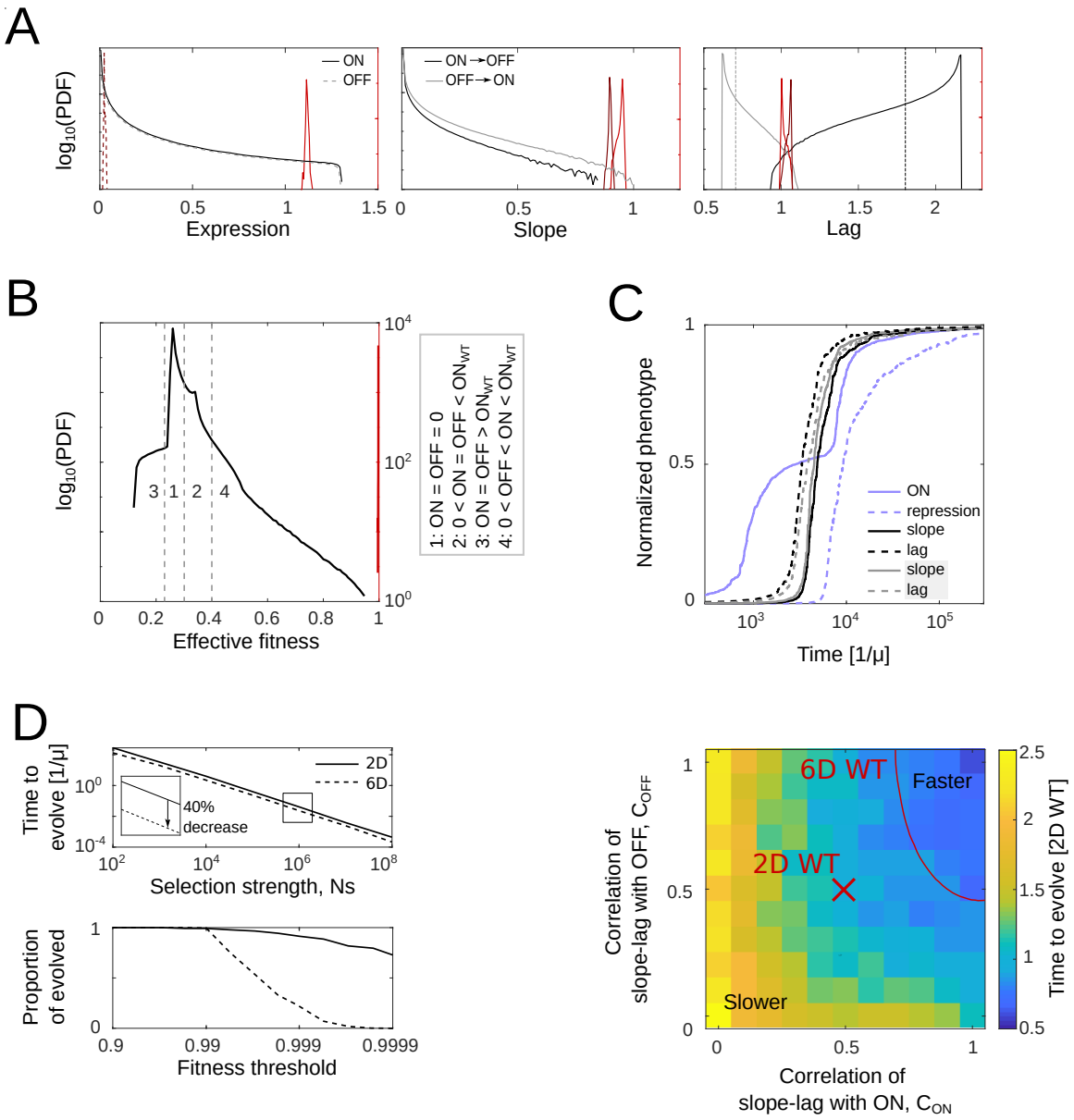

Figure 4.5: (Continued on the following page.)

that we link genotype to fitness through six phenotypes, while most measured DFEs either do not account for any phenotypic changes that alter fitness or, when they do, they only do so for a single phenotype [Orr, 2003; Eyre-Walker and Keightley, 2007; Bataillon and Bailey, 2014; Soskine and Tawfik, 2010].

Figure 4.5: **Higher dimensional phenotypes evolve faster. (A)** Probability density function (PDF) of phenotypes for $2 \cdot 10^9$ random sequence across the whole genotype space (black, gray, left y-axis) and distribution of 2000 evolved sequences (red, right y-axis). Black and dark red line represents phenotypes for ON to OFF dynamics, while gray and light red are phenotypes from OFF to ON dynamics. All units are in wild-type units, with the exception of OFF expression – the expression (ON and OFF) is in the units of wild-type ON expression. **(B)** Probability density function of effective fitness for random (black, left y-axis) and evolved (red, right y-axis) sequences, showing that selection strongly affects the distribution, leading to only very high fitness values. Solutions can be classified in 4 groups, depending on the functionality of RNAP and CI binding sites. Effective fitness is defined as $1 - \frac{1}{6}\sum_{i=1}^{6}(p_i - p_i^{\text{opt}})$, where p_i and p_i^{opt} are phenotype i and optimal value of phenotype i, respectively. Fitness is computed from phenotypes in (A). Note that the fitness function used in evolutionary model also depends on selection strength – Eq 4.17 and Section 4.4.4. **(C)** Time trajectories of phenotypes, showing the order of how phenotypes evolve. As OFF expression is selected for low values, we instead show repression which is the ratio of ON and OFF expression. Each curve represents a median of 2000 individual trajectories. For a better comparison on the order of how phenotypes evolve, we normalized phenotypes to start at 0 and end at 1. See Fig 4.13 for non-normalized results. **(D)** Top: Evolving 2D phenotypes is almost two-fold slower compared to 6D phenotypes. Time units are inverse mutation rate. Below: Proportion of evolved sequences as a function of fitness threshold which is a threshold above which sequence is considered evolved. With higher required fitness, increasing proportion of sequences get trapped in a local minimum, leading to lower proportion of evolved sequences. This effect is much more significant for 6D phenotypes. **(E)** Time to evolve, shown in color, where lag and slope were substitute with ON and OFF correlation – Eq 4.27 and Section 4.4.4. For $C_{\text{ON}} = C_{\text{OFF}} = 0.5$, the fitness function collapses to 2D phenotypes case, marked by red x. With higher correlations, evolution becomes faster. Red curve approximately marks the position where time to evolve equals that of 6D phenotypes. The time units are in 2D wild-type evolution time.

Evolutionary trajectories and dynamics

Theoretical models of evolution ordinarily consider a single, or, occasionally, two phenotypes that selection acts on. How evolution proceeds when multiple phenotypes are selected on remains poorly understood. The distributions of phenotypic and fitness effects of random mutations describe the potential starting points for repressible promoter evolution. As a miniscule portion of random sequences act as repressible promoters, selection must act on almost any random sequence in

order for a promoter to evolve. To simulate such evolution, we started with 2000 randomly selected sequences, and used a Strong Selection Weak Mutation (SSWM) model (adapted from [Tuğrul *et al.*, 2015]) with total population size N and selection strength of s to simulate evolution trajectories for each starting sequence. We selected for wild-type P_L function in terms of all six phenotypes, meaning that RNAP and CI binding sites had to emerge in the evolving promoter.

First, we were interested in whether the six phenotypes were fixed in the population in a specific order. While, unsurprisingly, ON expression always emerged first, there was also an order in which other phenotypes emerged in the population (Fig 4.5C). In other words, the evolutionary trajectories were not completely random, implying that selection for multiple phenotypes might be more predictable than expected. Furtermore, evolving some regulation on top of a weak constitutive promoter takes ten times longer then evolving the weak promoter itself (Fig 4.5C).

Intuitively, selecting for an additional phenotype ought to slow down evolution, because each phenotype needs to reach its own optimum. For example, resistance to a single antibiotic or pesticide generally evolves more rapidly than resistance to multiple ones [Durão *et al.*, 2018; Neve, 2007]. The combined model allowed us to compare the evolutionary trajectories and dynamics when selection acted on all six phenotypes (6D) or only on two phenotypes (ON and OFF expression – 2D). Starting from same random sequences, we compared how rapidly populations reached fitness comparable to that of the wild-type P_L.

Surprisingly, selecting on all six phenotypes led to more rapid rates of evolution, while being less precise (Fig 4.5D). Populations selected in 6D would more rapidly approach the optimum, but were less likely to reach the exact fitness of the wild-type P_L, compared to the populations selected only for ON and OFF expression. This effect becomes less significant with stronger selection (Fig 4.15), implying that selection in evolutionary steady state is unable to keep all traits at the maximum.

The observed evolutionary dynamics stem from the constraints that shape the GP landscape (Fig 4.3). The constrained nature of the GP landscape of the evolving promoters means that a mutation that alters one phenotype is likely to alter other phenotypes as well. As random sequences almost always have non-functional values

for all phenotypes (Fig 4.5A), any change, if it were to occur, was more likely to be positive. In fact, increasing the strength of the correlation between phenotypes increases the rates of evolution (Fig 4.5E) – explaining why evolution along a realistic constrained GP map is faster for six than for only two phenotypes. However, selection on six phenotypes is less likely to actually reach the precise optimum (Fig 4.5D bottom).

Distributions of mutational effects in evolving promoters

The distributions of phenotypic and fitness effects of random sequences (Fig 4.5A,B) provide an insight into the structure of the entire genotypic landscape, and, as such, inform about the potential starting points for de novo promoter evolution. Once a given sequence is under selection for promoter function, its evolution proceeds through individual mutations of that sequence. While many types of mutations can occur in nature, we here focus on one of the most common – single and double point mutations.

Distributions of phenotypic effects of point mutations (DME) summarize how point mutations might alter the given phenotype(s). While normally DME refers to a distribution of single point mutations, we also explored a DME for double mutants to increase the statistics (see Fig 4.9 for DMEs of single mutants). In contrast to the DME, distribution of fitness effects of point mutations (DFE) captures the fitness effect without explicitly accounting for how the mutation alters the underlying phenotypes. Numerous descriptions of DMEs and DFEs exist [Kemble *et al.*, 2019; Eyre-Walker and Keightley, 2007; Soskine and Tawfik, 2010], but their experimental determinations almost exclusively focus on DMEs and DFEs at a given point in evolutionary time. In other words, we lack any understanding of how DMEs and DFEs change as a given sequence evolves from non-functional to its optimum.

The detailed GP map combined with an evolutionary model allowed us to track how DMEs and DFEs of evolving sequences changed during the course of their evolution towards P_L-like function. The distribution of phenotypes of evolved promoters, show narrow landscapes around the reached evolutionary end points (Fig 4.5A). The range of phenotypic values that are reached after selection does not coincide

with phenotype values that are selected for, implying that wild-type P_L promoter differs from evolved promoters. Indeed, due to modeling difficulties (Section 4.4.1) our evolutionary model does not account for cooperativity between two CI dimers. The DFEs of evolving promoters are fairly smooth (Fig 4.6A, Fig 4.14). As the population moved towards its optimum, the frequency of deleterious mutations increased, although with a sharp decline for highly adapted promoters (Fig 4.6B). Furthermore, DFE characteristics depend only current fitness and not on the precise genotype, hinting that detailed description of genotype is not fully required for understanding DFEs (Fig 4.6A, Fig 4.6C top and middle). This puts these landscapes into the class that were studied before, yet it does not seem it is one of the standard thereotical landscapes [Kryazhimskiy *et al.*, 2009]. Meanwhile, evolving genotypes change their rate at which they travel through adaptive landscape (Fig 4.6C bottom). Adaptive landscape, shown on Fig 4.6D, contains long ridges of comparable fitness, suggesting abundant and potentially long neutral networks in which mutations might alter genotype but not fitness [Wagner, 2005], as commonly observed in larger gene regulator networks [Payne and Wagner, 2014].

DMEs and DFEs are not static properties but rather change, often dramatically, during the course of evolution. The changes in the shape and structure of the DFE can occur even after only one or two mutations, especially when the population is well adapted (Fig 4.6C, Fig 4.14). Therefore, modeling evolution must account for the changing nature of DMEs and especially DFEs, rather than assuming a fixed DFE throughout the course of the evolution.

Evolved promoters

From a theoretical perspective, predicting how a population traverses a given fitness landscape, or the dynamics of evolution, has received more attention than predicting the outcomes of evolution. This is, in large part, due to the lack of detailed GP maps, resulting in an advanced understanding of how selection operates but a relatively poor description of the genotypes that actually evolve.

The dynamical GP map that we developed allowed us to not only understand how repressible promoters evolve but also what genotypes were favored by selection.

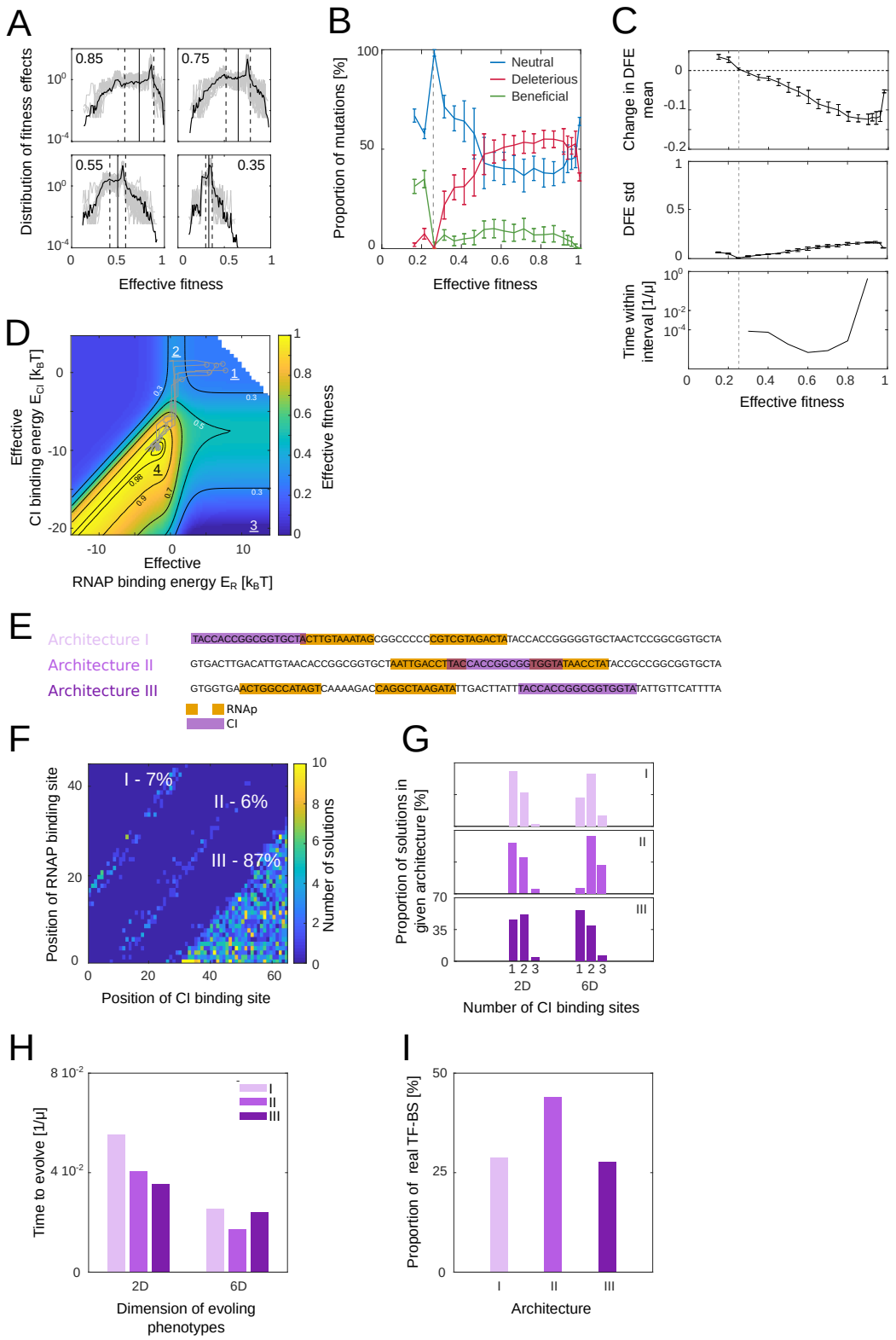

Figure 4.6: (Continued on the following page.)

Figure 4.6: **Selection gives rise to different architectures of binding sites. (A)** Distribution of fitness effects (DFE) for 30 different genotypes (gray) with the same effective fitness value (number writen in each plot), and their average (black curve). Vertical black and dashed lines represent mean +- std of the average distribution. **(B)** Proportion of neutral, deleterious, and beneficial mutations as a function of effective fitness. Error bars represent s.t.d. over 30 different genotypes with the same effective fitness. Deleterious and beneficial mutations are defined as mutations that decrease and increase effective fitness by at least 0.02, respectively. The vertical dashed line shows effective fitness 0.75 of non-functional sequences. **(C)** Top: the difference in mean of DFE and fitness of starting genotype as a function of fitness. Negative values represent that majority of mutations are deleterious. Effective fitness is defined in Fig 4.5B caption. Middle: Standard deviation of DFEs as a function of fitness. Error bars in top and middle plot represent standard deviation of mean and s.t.d. estimates over 30 replicates. Botom: Time spent within the fitness interval of ±0.1. The results are the median over 2000 evolving sequences. The vertical dashed line shows effective fitness 0.25 of non-functional sequences. **(D)** Representation of fitness landscape with the effective CI and RNAP binding energy (see Section 4.4.1). Contours show the lines of equal fitnes. Underlined numbers 1-4 show the classifications in Fig 4.5B. Gray lines represent 6 examples of evolutionary trajectory: first the RNAP binding site evolves (decreasing RNAP binding energy), reaching a ridge and starting to evolve CI binding site. White area denotes where the amplitude of expression is $< 10^{-15}$, reaching computer precision limit. The third parameters, describing binding configurations with CI bound upstream, does not have an effect on fitness landscape, it only changes the size of the white area (Section 4.4.1). **(E)** Example of three classes of evolved architectures with marked binding RNAP and CI binding sites. **(F)** Histogram of positions of strongest RNAP and CI binding sites. We mark the three architectures described in E with the percentages of each of them. **(G)** The distribution of number of significant CI binding sites between different architectures (top, middle, botom), together with 2D and 6D comparison. Significant binding site is defined as contributing at least 10% to total repression. Many of evolved sequences have more than one significant binding site. **(H)** Time to evolve varies not only between 2D and 6D, but also between different architectures. The results from D-H are from 2000 replicates with $N = 10^6$ and $s = 1$. **(I)** Using the data from RegulonDB, we obtained the position of binding relative to RNAP binding site of over 700 repressors. These were classified in one of the three architectures. A majority fall in architecture II which coincides with the fact that this architecture evolves fastest in 6D evolution.

Specifically, we were interested in what promoter architectures were more likely to emerge when random sequences evolved into repressible promoters. In addition to CI sterically excluding RNAP binding, we also considered any CI binding downstream of RNAP binding to have repressor function, and observed the number of

CI binding sites that evolved (see Section 4.4.1). We also observed the architecture of the promoters that emerged, defined as the relative position of the strongest (dominant) CI binding site relative to the RNAP binding site (Fig 4.6E).

The likelihood of emergence was not random for the three promoter architectures – the dominant CI binding site was more likely to appear downstream of the RNAP binding site (Fig 4.6F). The location of the strongest binding site (i.e. architecture) also impacted the total number of CI binding sites that needed to evolve in order to reach wild-type P_L levels of repression (Fig 4.6G). The likelihood of a given architecture emerging was related to its speed of evolution (Fig 4.6H), which was, at least in part, affected by the constraints associated with that architecture (Fig 4.6E). Furthermore, the RNAP binding site, which always evolves first (Fig 4.5C), introduced further constrains on the emergence of CI binding site(s). For example, when the dominant CI binding site evolved between the RNAP -10 and -35 binding site (architecture II), it often required additional CI binding sites to reach the fitness optimum (Fig 4.6G). This is because the dominant CI binding site in architecture II has direct overlap with RNAP binding sites, limiting the range of mutations that will increase CI binding without negatively affecting RNAP binding.

Selection acting only on two phenotypes (ON and OFF) predicts different evolutionary outcomes (promoter architectures and binding site numbers) to selection acting on all six phenotypes, with 6D selection resulting more frequently in multiple CI binding sites (Fig 4.6G). Furthermore, the predicted rates of evolution of the three architectures were also different between selection for two versus all six phenotypes (Fig 4.6H).

It remains completely unexplored whether selection in repressible promoters actually acts on dynamical (6D) or only on steady-state (2D) phenotypes. To indirectly examine this question, we collected the information about all known promoters in *E.coli* from RegulonDB [Salgado *et al.*, 2013]. Specifically, we classified all known repressible promoters into the three promoter architectures (shown in Fig 4.6E), using the information about the known position of repressor binding sites relative to RNAP binding sites. Interestingly, we found that the largest number of known promoters had a repressor binding site between the -10 and -35 RNAP sites (archi-

tecture II) (Fig 4.6I). Our model predicted this architecture to arise most rapidly when selection acts on six phenotypes, but not when it acts on only two. Therefore, while a multitude of factors likely contributed to architecture II being the most common in the *E.coli* genome, one of them might be that selection more frequently acts on dynamical rather than just steady-state phenotypes.

4.2.6 Generalizing beyond the studied system

In this work, we developed a comprehensive GP map capable of predicting how mutations in a promoter alter six dynamical gene expression phenotypes, and then used the model to understand how such promoters might evolve. To achieve this, we focused on the Lambda P_L promoter as a well understood model system in molecular biology and gene regulation [Ptashne, 2011; Lagator *et al.*, 2017a]. Now we ask how can we apply the lessons learned from Lambda P_L to other promoters and gene regulatory networks in general?

To address this question, we focused on the fundamental evolutionary property of any system under selection, the DFE, and explored its mechanistic origins. The most important component of our combined model for linking genotypic mutations to their effect on phenotype is the energy matrix (Fig 4.2A). Indeed, the constraints that define GP mapping (Fig 4.3) are in no small part attributable to the structure of the energy matrix (Fig 4.4D). And yet, while the energy matrix defines the specific nature of those constraints, the total range of phenotypes that could possibly be realized through mutations depends only on the range of values in the energy matrix but not on its internal structure (Fig 4.3A – black lines). In fact, the generalizable mechanisms (promoter architecture, concentrations of transcription factors, the total range of energies in the energy matrix and, importantly, the biophysical laws that are explicitly modeled) that determine how a transcription factor binds DNA set the limits for achievable phenotypes, while the specific energy matrix structure determines the correlations between phenotypes.

The question of how generalizable our findings are to other promoters becomes a question of whether predicting DFEs and evolutionary dynamics without an explicit

energy matrix is accurate. If the generalizable mechanisms of transcription factor-DNA binding indeed define the limits of what phenotypes can possibly be realized, then they ought to give sufficiently good predictions of DFEs and evolutionary dynamics without accounting for the internal structure of the energy matrix. The energy matrix is a multi-variable component of the model, connecting genotype to binding energies of RNAP and CI. Here, we summarized the two energy matrices and promoter architecture through three variables that capture the general range of energy values present in the energy matrices but not its internal structure (Fig 4.7A yellow rectangle, also Section 4.4.4). In this model, which we referred to as the 'geometric model on binding energies', a mutation alters one or more of the energy matrix summary variables (see Section 4.4.4). Importantly, mutational changes are represented as additive changes for energies but multiplicative for the rates in the model. Therefore, proper choice of effective variables is crucial and is informed by the underlying model. This model predicted DFEs that were comparable to those predicted by the combined model with the energy matrix (Fig 4.7B). This means that summarizing the energy matrices and promoter architecture with only 3 parameters resulted in consistent predictions of promoter evolution even though such a model did not account for the specific relationship between genotypic mutations and dynamical phenotypes. Therefore, the mechanisms that govern the GP mapping cannot account for all details of the underlying GP mapping nor are they sufficient to predict the phenotypic effects of specific mutations. However, they are sufficient to understand general trends of promoter evolution.

When modeling evolution, typically the fitness effects of mutations are drawn from an assumed distribution of phenotypes. That distribution is, most of the time, a theoretical assumption rather than built on experimental measurements of DMEs and DFEs. We implemented such an approach to model promoter evolution, where each mutation directly altered one or more phenotypes – we refer to this model the 'geometric model on phenotypes'. This model, which is in line with typical population genetics models does not account for any aspects of GP mapping (Fig 4.7A red rectangle, Fig 4.7B). The predictions of DFEs, and hence our understanding of evolution, are dramatically different if we do not account for the underlying GP

mapping or, at least, for the mechanisms that govern that mapping – questioning the common approaches used in theoretical evolution – Fig 4.7A.

In principle, extending our model (Fig 4.2A) to other regulated promoters, or even to more complex networks, is relatively straightforward. The fundamental aspects of the MAK ought to be true for any dynamical molecular system. Similarly, utilizing thermodynamics to predict steady-state gene expression levels is possible for any promoter or a network of any size [Bintu *et al.*, 2005a]. The major difficulty in accurately mapping GP in other networks comes from the fact that the predictive power of the TD model relies on having the relevant energy matrices [Kinney *et al.*, 2010; Lagator *et al.*, 2020], and obtaining energy matrices is labor- and time-intensive. And yet, our results suggest that for understanding evolution, using easy-to-derive summary of energy matrices might be sufficient, providing a key insight into how our model can be extended to other systems. This is why not only describing GP mapping, but also understanding its mechanistic origins, ought to form a crucial new direction in studying evolution – it provides a balance between detailed and exhaustive experiments and using generalizable assumptions that, at least in the case of promoters, often provide inaccurate understanding of evolution.

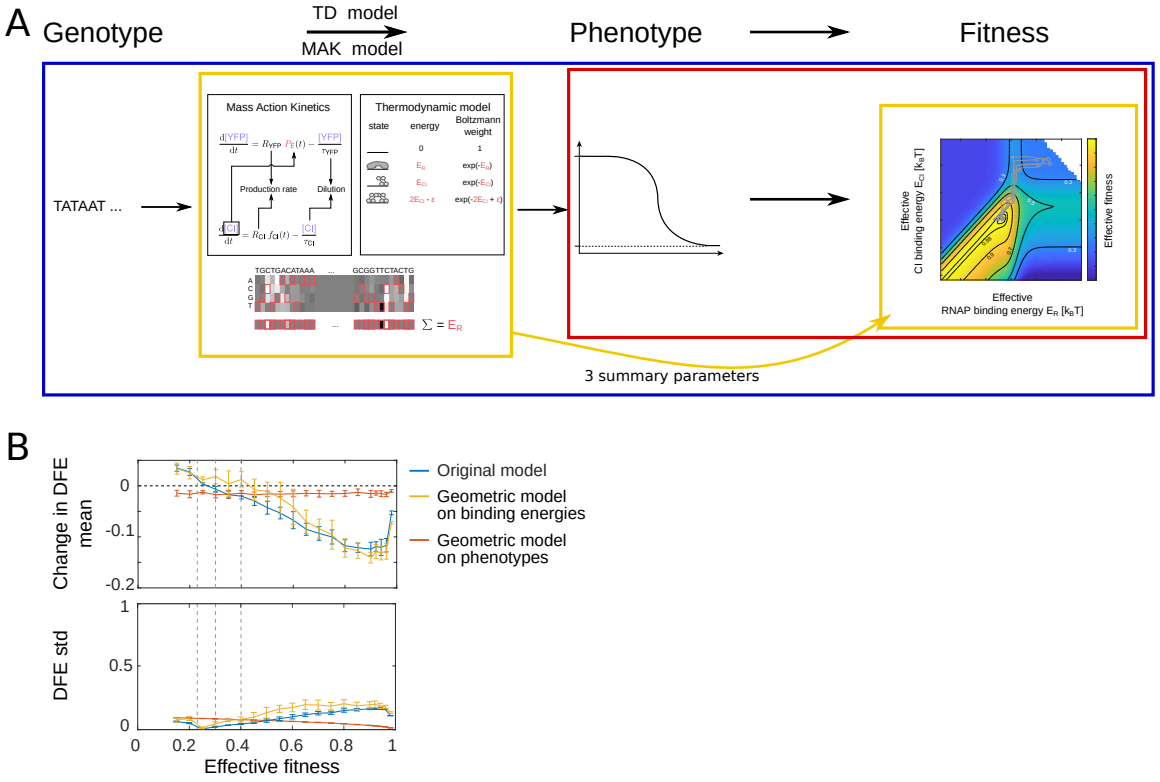

Figure 4.7: **The prediction of DFEs must account for the mechanisms that governs the GP mapping. (A)** The summary of the genotype-phenotype-fitness mapping, showing the major steps. Mechanistic model describing the relationship between genotype and phenotype can be used to produce 3 summary parameters - independent of genotype - which describe the fitness landscape and DFEs qualitatively accurate. **(B)** The comparison of DFE statistics for the original model (blue), including all the details of genotype-phenotype-fitness mapping, or two types of geometric model: one where mutations are represented as random effects on binding energies (yellow) represented by 3 summary parameters, or random effects on phenotypes (red). Geometric model with phenotypes shows independent change in DFE and decrease in DFE s.t.d with increasing fitness. This is in no agreement with real DFEs. Meanwhile, geometric model on binding energies qualitatively predicts the correct trends in DFE statistics, showing that understanding mechanisms governing GP mappings is needed to understand DFEs. See Section 4.4.1 for details.

4.3 Discussion

In this work, we developed a modeling approach that can accurately predict how mutations in a bacterial promoter alter the dynamics of gene expression. Doing so allowed us to, for the first time, examine how promoters evolve if selection acts on complex phenotypes determined by complex promoters, rather than only on simple

single environment expression levels. Dynamics of gene expression are a critical component of gene regulation, as the rate at which a gene is turned on or off can alter molecular decision making and organismal development. For example, in the system we studied, namely the Lambda P_L promoter that acts as a genetic switch between the phage lytic and lysogenic lifestyle, slowing down the expression of the repressor can result in a higher proportion of phages incorporating into the host genome [Ptashne, 2011]. On a broader genomic scale, it is plausible that selection acted on dynamical phenotypes, rather than only on steady-state expression levels, when shaping the architecture of *E.coli* promoters (Fig 4.6G,H). In spite of their importance, the role of gene expression dynamics in shaping the structure of gene regulatory networks remains poorly understood, as most studies focused on steady-state expression to describe how networks function and how they evolved [Payne and Wagner, 2014; Aguilar-Rodríguez *et al.*, 2017; Igler *et al.*, 2018; Taylor *et al.*, 2015; Babu and Teichmann, 2003]. The model we developed, which is extendable to most regulated bacterial promoters, can form the foundation for exploring how dynamics affect network structure and evolution in a more comprehensive manner.

Our model offers the prospect of improving the development of synthetic constructs in bacterial species. Synthetic biology often requires the development of gene regulatory cascades, when the expression of one component in the network triggers the expression of subsequent one [Trosset and Carbonell, 2013]. The optimization of promoters that constitute such synthetic networks can be critical for desired functioning of the construct, as changes in the levels or the dynamics of gene expression levels can alter construct performance [Singh, 2014]. Predicting the effects of mutations in silico, rather than having to experimentally create them in the lab, can speed up the process of developing a synthetic microorganisms for industrial purposes.

Describing, let alone predicting, how genotype maps onto phenotype has been a long-standing goal in evolutionary biology. In gene regulation, GP maps have been developed for entire gene regulatory networks [Payne and Wagner, 2014; Aguilar-Rodríguez *et al.*, 2018; Carter *et al.*, 2013] or for individual promoters [Otwinowski and Nemenman, 2013; Barnes *et al.*, 2019; Kinney *et al.*, 2010; Lagator *et al.*, 2017a].

However, all such maps consider only steady-state expression as the phenotype. More broadly, predictive GP networks have been developed for only a handful of biological systems – RNA folding and metabolic networks. Predictive RNA folding was the first biophysically-rooted GP map [Schuster *et al.*, 1994; Schuster, 2006], but the effects of altering RNA structure on fitness are difficult to understand. Models of metabolic networks, on the other hand, have a clear link between phenotypic changes and their effect on fitness, but can only account for large-scale mutations like deletions and knock outs [Yi and Dean, 2019; Segré *et al.*, 2000; Szathmáry, 1993]. Our combined model extends the ability to predict GP mapping to, plausibly, most bacterial promoters at the level of single point mutations, offering unprecedented detailed insights into the forces that shape GP mapping.

Developing a predictive GP map that accounts for mutational effects on six phenotypes allowed us to explore how biological systems evolve when selection acts on more than a single or a couple of phenotypes. Typical theoretical models of evolution focus on a single trait that is either controlled by a single or a large number of genes. Similarly, most experimental approaches that measured the effects of mutations either investigate how mutations directly alter fitness [Bataillon and Bailey, 2014; Keightley, 2000; Kassen and Bataillon, 2006] or describe how they affect a single phenotype of interest [Soskine and Tawfik, 2010; Lehner, 2013]. The detailed GP mapping enabled by our model identified a potential difficulty with the existing approaches to understanding mutational effects and evolution. Namely, the evolutionary outcomes of selection acting on multiple rather than a single phenotype can be drastically different (Fig 4.5, Fig 4.6).

Put together, our findings emphasize the need to improve our understanding of the mechanism that underpin biological function, and to understand the evolutionary consequences of those mechanisms [Yi and Dean, 2019]. Doing so allowed us not only to develop a predictive GP map for a repressible bacterial promoter, but also to understand how that promoter might evolve. Theoretical models of evolution commonly assume how mutations alter fitness without accounting for how mutations alter phenotypes that underpin fitness changes [Milocco and Salazar-Ciudad, 2020]. Doing so can result in misrepresentation of how a biological system evolved

(Fig 4.7B). In contrast, developing mechanistic models that link back directly to experimental observations offers promise of more accurate description and understanding of evolutionary processes.

4.4 Methods

4.4.1 Model

Thermodynamic model

The thermodynamic model is a well established model for gene regulation which provides a highly quantitative mapping from promoter sequences to gene expression levels that is compatible with biophysical measurements [Bintu *et al.*, 2005a; Bintu *et al.*, 2005b; Kinney *et al.*, 2010; Lagator *et al.*, 2020]. It assumes that we can use statistical mechanics to describe equilibrium probabilities of different molecules binding to the sequence of interest, and using these to describe the expression of the gene of interest.
The thermodynamic model requires us to know i) all the possible binding configurations, ii) binding energies (and interacting energies) of the binding configurations, and iii) available concentrations of the binding molecules.

Binding configurations. Binding configurations are specific to each system – in our system, the following binding states are possible (Fig 4.2A):

1. empty state, i.e., nothing is bound
2. RNAP bound to, e.g., P_{L} promoter,
3. CI dimer bound to, e.g., O_{R1} or O_{R2},
4. two CI dimers cooperatively bound, e.g., on O_{R1} and O_{R2}.

These are 4 major possible configurations – in each, different binding locations are possible. For example, RNAP can bind to its strongest binding site at -35 and -10, or at any other part of the sequence. Of course, binding to other random pieces of sequence is often very unlikely and will contribute very little to total binding. However, when there is no one clear strong binding site, binding to these weaker binding sites becomes important [Lagator *et al.*, 2020].
Technically, there are other possible configuration, such as RNAP and CI both binding at the same time to different binding sites (without steric hindrance). Another example would be 3 CI dimers simultaneously bound to the DNA. However, these

other configurations are extremely unlikely and contribute negligible amount to total binding. They would become significant (and important to include) only if strong RNAP and CI binding sites would not overlap. We include some of them in the evolutionary calculations – for details on that see the Section 4.4.4.

Binding energies. Each of the above mentioned configurations has an energy of binding that is obtained using an energy matrix. The energy matrix (EM) contains the information about how every possible point mutation in the DNA-binding site of a given molecule impacts its overall binding energy [Lagator *et al.*, 2020; Kinney *et al.*, 2010]. As such, each DNA-binding molecule has a unique EM associated with it, which can be thought of as a unique representation of that molecule's function, much like the amino acids sequence is a two-dimensional representation of that molecules 3D structure. Therefore, in our system we require two EMs, one for description of RNAP and one for CI – see Fig 4.8.
Therefore, EM can be represented by $4 \times L$ matrix (hence the name) whose elements give the energy contribution of the given nucleotide (rows) at given position (columns) to the total binding energy. The total binding energy is then the linear sum of individual energies, each contributed from individual nucleotides (Fig 4.2A). How do we obtain the energy of binding from the energy matrix? We align the binding site sequence with the EM, then for each position taking the EM element that corresponds to the correct nucleotide in the sequence. For an example see Fig 4.2A.

RNAP energy matrix includes also flanking regions. RNAP EM is described by two parts, one for each of -10 and -35 binding sites, with a spacer between them. As flanking regions of the common 'TATAAT' and 'TTGACA' sites also significantly contribute to binding [Lagator *et al.*, 2020], they are included in the energy matrix, making each part of RNAP energy matrix 12bp (not 6bp) long.

Flexiable spacer penalties As shown in [Lagator *et al.*, 2020], flexiable spacer has an impact on RNAP binding. In our model we use their published energy

penalties: 1.38, 0.54, 0, 0.17, and 0.94 for spacer variability between −2 and 2. These values are in the units of the largest element in the RNAP energy matrix – see Fig 4.8. Therefore, to obtain energy values in k_BT, these values must be multiplied with energy scale α, obtained in Section 4.4.2.

For default spacer with no energy penalty, energy of RNAP binding to a given position is E^R. However, for binding to the same binding site with spacer i, energy of binding should be modified to $E^R \rightarrow E^R + \delta_i$, where δ_i represents the energy penalty due to spacer i.

By having five possible spacer configurations, the total number of possible RNAP binding configurations increases by 5-fold.

Using the energy matrices shown in Fig 4.8, the default spacer with no energy penalty is 8bp.

Expression and probability of expressing state. One of the main assumptions of the thermodynamic model is that rate of expression – and thus steady state expression value – is proportional to the probability that expressing state occurs. What is an expressing state? That is each configuration of the system which leads to expression. In our case these are the configurations with bound RNAP.

Therefore, we can write the probability of finding the system in the state with RNAP bound as

$$P_E = \frac{\sum_i [\text{RNAP}] e^{-E_i^R}}{1 + \sum_i [\text{RNAP}] e^{-E_i^R} + \sum_i [\text{CI}_2] e^{-E_i^{\text{CI}}} + \sum_i [\text{CI}_2]^2 e^{-E_i^{\text{CI}} - E_{i+24}^{\text{CI}} + \epsilon}}. \tag{4.1}$$

The numerator in the above equation is the Boltzmann weight of the RNAP bound state, while the denominator represents the sum of Boltzmann weights of all possible configurations. E_i^R and E_i^{C} represent binding energies of RNAP and CI, respectively, to binding site i which represents different binding sites along the sequence. [RNAP] and [CI_2] represent the available RNAP and CI dimer concentration, respectively, and $\epsilon > 0$ the cooperativity energy between two CI dimers whose start of binding sites are 24bp apart.

For the wild-type sequence, there is only one significant RNAP binding site, and two CI binding sites (O_{R1} and O_{R2}) on which CI can cooperatively bind.

Furthermore, all energies must be in the units of $k_{\mathrm{B}}\mathrm{T}$.

The relation between CI monomer and dimer concentration. As binding to the CI binding site occurs by CI dimers (quantity required in thermodynamic model) and not monomers (quantity obtained from mass action kinetics model), we compute the relationship between the two. Let us denote the rate of two CI monomers forming a CI dimer as k_1 and the opposite dissociation rate as k_2. To a very good approximation, we can assume that the system is in chemical equilibrium, meaning that the processes of dimerization and dissociation occur faster than the changes in CI concentration. Therefore, we have the following chemical reaction: $2[\mathrm{CI}] \underset{\mathrm{k_2}}{\overset{\mathrm{k_1}}{\rightleftharpoons}} [\mathrm{CI_2}]$. Using the law of mass action in equilibrium, we can rewrite it as $k_1[\mathrm{CI}]^2 = k_2[\mathrm{CI_2}]$, and thus $[\mathrm{CI_2}] = \frac{k_1}{k_2}[\mathrm{CI}]^2$.
This means that we can rewrite Eq 4.1 as

$$P_E = \frac{\sum_i [\mathrm{RNAP}] e^{-E_i^R}}{1 + \sum_i [\mathrm{RNAP}] e^{-E_i^R} + \sum_i \omega_1 [\mathrm{CI}]^2 e^{-E_i^{\mathrm{CI}}} + \sum_i \omega_1 [\mathrm{CI}]^4 e^{-E_i^{\mathrm{CI}} - E_{i+24}^{\mathrm{CI}} + \epsilon}}, \tag{4.2}$$

where $\omega_1 = \frac{k_1}{k_2}$ contains the rates describing relation between CI monomers and dimers.

Reference points of energies. The quantities appearing in Boltzmann weights (Eq 4.2) are the binding energy of the state and available concentrations. As defined above, the binding energies are relative to the unbound state. However, the energy matrix produces only the change in binding energy, relative to some reference point. In our case, we assign this reference point to be binding to wild-type sequence to P_L and O_{R1} for RNAP and CI energy matrix, respectively.
Assuming that the binding energies in Eq 4.2 are only changes relative to these reference points, the binding energies of the reference points must be taken into account:

$$P_E = \frac{\sum_i g_1 [\mathrm{RNAP}] e^{-E_i^R}}{1 + \sum_i g_1 [\mathrm{RNAP}] e^{-E_i^R} + \sum_i \omega [\mathrm{CI}]^2 e^{-E_i^{\mathrm{CI}}} + \sum_i \omega [\mathrm{CI}]^4 e^{-E_i^{\mathrm{CI}} - E_{i+24}^{\mathrm{CI}} + \epsilon}}, \tag{4.3}$$

where $g_1 = e^{-E_{\mathrm{WT}}^R}$ represents the Boltzmann weight of RNAP binding to the wild-type sequence of P_L, and $\omega = \omega_1 e^{-E_{\mathrm{CI}}^{\mathrm{WT}}}$ the combination of dimer/monomer rates

(ω_1) and CI Boltzmann weights of binding to the wild-type sequence of O_{R1}.
As E^R and E^{CI} represent only the energies of mutational effects relative to the reference points, it means that their values equal zero for binding to said reference points.

Changes in CI concentration are much slower than the equilibration of the system. The thermodynamic model described above gives the prediction of expression where all quantities are assumed to be in equilibrium. However, in our system, the concentration of repressor varies, potentially violating this assumption. Yet, if the time scales on which CI concentration varies is much slower than the time scale on which the equilibrium is established, the assumption of equilibrium would still be satisfied. In our case, the time scale of varying CI concentration is hours, much longer than the typical processes in the cell.

Mass action kinetics model

The second model is the mass action kinetics (MAS) model which follows the concentration of repressor CI and fluorescence protein YFP. The concentration of CI is used to model the probability of RNAP being bound and rate of expression (Eq 4.3), while the YFP concentration is used as a proxy for gene expression.
We use two ODEs, one for each concentration. Both have two terms, one that describes the production of the molecule, and the second with processes that lower the concentration.

Concentration of CI. We model the total repressor concetration [CI] as

$$\frac{\mathrm{d}[\mathrm{CI}]}{\mathrm{d}t} = R_{\mathrm{CI}} f_{\mathrm{CI}}(t) - \frac{[\mathrm{CI}]}{\tau_{\mathrm{CI}}}, \tag{4.4}$$

where R_{CI} represents the production rate of CI, $f_{\mathrm{CI}}(t)$ the delay in production rate and takes values between 0 and 1, and τ_{CI} describes the effects of dilution and degradation.

Delay in production of CI for ON→OFF. When studying the dynamics of going from ON to OFF expression state, P_{TET} promoter is induced, leading to expres-

sion and production of CI. However, there exist a delay between the introduction of the inducer into the system and between the repressor being present and able to bind. This delay is due to i) diffusion of inducer to its cognate binding site, ii) transciption and folding of CI protein, and iii) diffusion of CI to its cognate binding site. We do not discuss the details of these three contributions in details but lump them into one delay, described by:

$$f_{\mathrm{CI}}(t) = \frac{t^n}{t^n + \tau_1^n}, \tag{4.5}$$

where τ_1 is the effective time scale of delay, and n the effective Hill coefficient (or sharpness) of delay. This makes sure that the production rate of CI for $t \ll \tau_1$ is zero, while for $t \gg \tau_1$ production rate converges towards R_{CI}.

Delay in production of CI for OFF→ON. Similarly as above, we now address the production of CI for the dynamics of OFF→ON, where inducer of CI production is removed from the system. However, there is a delay in stopping the production of CI and CI is being partially produced even when new environment (without inducer) occurs. This is due to the fact that in this new environment, some inducer is still left inside the cell and is not completely removed. This leads to production of CI also in the new environment where no inducer is present. We describe this process by:

$$f_{\mathrm{CI}}(t) = \frac{\tau_2^\beta}{t^\beta + \tau_2^\beta}, \tag{4.6}$$

where τ_2 is the effective time scale, and β the effective Hill coefficient (or sharpness). This makes sure that for $t \ll \tau_2$ the production of CI is still R_{CI}, then decreasing towards zero when $t \gg \tau_2$.

Concentration of YFP. Similarly as for repressor, we model concentration [YFP] as

$$\frac{\mathrm{d[YFP]}}{\mathrm{d}t} = R_{\mathrm{YFP}} P_E(t) - \frac{\mathrm{[YFP]}}{\tau_{\mathrm{YFP}}}, \tag{4.7}$$

where R_{YFP} represents the basal production rate of YFP, and $P_E(t)$ the probability of RNAP being bound (leading to expression) from Eq 4.2. P_E changes as a function

of time as the concentration of CI (which appears in P_E) also changes with time. P_E takes values between zero and one. Here we assume that the probability of RNAP being bound is linearly proportional to the rate of expression. τ_{YFP} describes the effects of dilution and degradation of the YFP protein.

4.4.2 Obtaining the parameters for the model

Each of the two models (MAK and TD) requires different set of parameters. TD model includes the following parameters (see Eq 4.3): prefactor in the RNAP bound state g_1[RNAP] (which we can treat as one parameter), scaling factors that determine the units of energy matrix elements α and ι for RNAP and CI energy matrices, respectively, prefactor in CI bound states ω, and cooperativity ϵ between two CI dimers bound at O_{R1} and O_{R2}.
Alternatively, MAK model parameters include (Eq 4.4, 4.7): YFP and CI production rates (R_{YFP} and R_{CI}, respectively), dilution and degradation times of YFP and CI (τ_{YFP} and τ_{CI}, respectively), and parameters β, n, τ_1, τ_2 describing delay in production of CI.
These model parameters were obtained from different independent data sets, described below.

Thermodynamic model parameters

The parameters for the thermodynamic model are the parameters that describe the steady state expression; in other words, the expression in ON and OFF state, without any temporal dynamics between the two. To obtain these parameters we use an existing Lambda P_R random mutant library [Lagator *et al.*, 2020]. This library includes over 25,000 unique mutants of the lambda P_R, each in two different environments – with and without CI repressor present, equivalent to ON and OFF environments. The distribution of mutational effects in each environment covers a wide range of expression value – from low to high expression in both environments. This means that the mutations have various effects, from impacting only RNAP or CI binding sites, to affecting both binding sites.

To fit the parameters of the TD model to this library, we separate the library into two disjoint sets; first part with 15,000 mutants which is used to fit the data (fitting set), and the remaining 10,000 mutants to test how good the fit is (evaluation set). Using the fitting set, we fit the steady state ON expression without present CI repressor to obtain g_1[RNAP] = 3.27 and RNAP energy matrix scaling $\alpha = 4.85 k_B T$. Next, we fit the TD model to the fitting set in the OFF environment, keeping the g_1[RNAP] and α fixed. There, we obtained cooperativity $\epsilon = 3.22 k_B T$, $\omega[\text{CI}]^2_{\text{steady state}} = 0.01$, and CI energy matrix scale $\iota = 3.00 k_B T$.

To test how good the fit is, we used the evaluation set from library data (on which the model was not fitted) and computed Pearson's correlation coefficient between predicted and measured data points. For ON expression (without repressor) we obtained $\rho_{\text{ON}} = 0.92$, and for OFF expression (with repressor) $\rho_{\text{OFF}} = 0.82$.

To fit these parameters, we minimized the sum of squared difference between model prediction and data.

Additionally, the two EMs were obtained from [Lagator *et al.*, 2020].

Mass action kinetics model parameters

The MAK model parameters are those that describe the temporal change in CI and YFP concentration. First, let us explore the steady state value of CI:

$$\frac{\mathrm{d}[\text{CI}]}{\mathrm{d}t} = R_{\text{CI}} f_{\text{CI}}(t) - \frac{[\text{CI}]}{\tau_{\text{CI}}} = 0 \tag{4.8}$$

$$[\text{CI}]_{\text{steady state}} = R_{\text{CI}} f_{\text{CI}}(t) \tau_{\text{CI}} = \begin{cases} 0, & \text{in ON environment} \\ R_{\text{CI}}\ \tau_{\text{CI}}, & \text{in OFF environment} \end{cases} \tag{4.9}$$

Alternatively, the steady state values of YFP equals:

$$\frac{\mathrm{d}[\text{YFP}]}{\mathrm{d}t} = R_{\text{YFP}} P_E(t) - \frac{[\text{YFP}]}{\tau_{\text{YFP}}} = 0 \tag{4.10}$$

$$[\text{YFP}]_{\text{steady state}} = R_{\text{YFP}} P_E \tau_{\text{YFP}} = \begin{cases} R_{\text{YFP}}\ P_E^{\text{ON}}\ \tau_{\text{YFP}}, \text{in ON environment} \\ R_{\text{YFP}}\ P_E^{\text{OFF}}\ \tau_{\text{YFP}}, \text{in OFF environment} \end{cases} \tag{4.11}$$

where P_E^{ON} and P_E^{OFF} represent P_E in ON and OFF expression, respectively. In P_E the appropriate value of $[\text{CI}]_{\text{steady state}}$ is used.

YFP production rate only determines units of YFP. We next show that YFP production rate, R_{YFP}, only determines the units of YFP concentration; in other words, we show that this production rate only scales YFP concentration. If we rewrite $[\text{YFP}] = [\text{yfp}] \cdot R_{\text{YFP}}$, and use this in Eq 4.7, we show that we obtain an ODE with rescaled YFP concentration (marked by [yfp]) but where the production rate doesn't appear in the ODE:

$$R_{\text{YFP}} \frac{\mathrm{d}[\text{yfp}]}{\mathrm{d}t} = R_{\text{YFP}} P_E(t) - R_{\text{YFP}} \frac{[\text{yfp}]}{\tau_{\text{YFP}}} \quad \rightarrow \quad \frac{\mathrm{d}[\text{yfp}]}{\mathrm{d}t} = P_E(t) - \frac{[\text{yfp}]}{\tau_{\text{YFP}}} \tag{4.12}$$

By demonstrating that YFP production rate isn't present in the ODE, we follows that it doesn't determine the dynamics of YFP.

CI production rate is determined from steady state OFF expression. Similarly as for YFP, CI production rate also determines only the units of CI concentration and not its dynamics. As the maximum effective steady state concentration of CI is already determined by the steady steady expression in the presence of CI, we set R_{CI} to be such that the constraint $\omega[\text{CI}]_{\text{steady state}} = 0.01$ is met. In practice, this means we can set $\omega = 0.01$ and $[\text{CI}]_{\text{steady state}} = 1$, following that $R_{\text{CI}} = 1/\tau_{\text{CI}}$.

Normalization of YFP to wild-type ON expression. Due to experimental reasons, YFP expression values in experimental data have arbitrary units. Therefore, we decided to use intuitive units of expression YFP and normalize all YFP results by wild-type ON expression. In other words, the wild-type ON expression is set to have value 1.

The steady state concentration of YFP of the wild-type sequence in ON environment is written as:

$$[\text{YFP}]^{\text{ON}}_{\text{WT}} = R_{\text{YFP}}\ \tau_{\text{YFP}}\ P_E^{\text{WT}}\ ([\text{CI}] = 0) \tag{4.13}$$

$$= R_{\text{YFP}}\ \tau_{\text{YFP}} \frac{\sum_i g_1 [\text{RNAP}] e^{-E^R_{i,\text{WT}}}}{1 + \sum_i g_1 [\text{RNAP}] e^{-E^R_{i,\text{WT}}}}. \tag{4.14}$$

Effectively, we set $R_{\text{YFP}} = 1$, and normalize all YFP results by $[\text{YFP}]^{\text{ON}}_{\text{WT}}$. Alternatively, one can think of it as constraining the YFP production rate R_{YFP} such that wild-type expression in ON environment equals to one.

Determining the dynamical MAK parameters. The remaining parameters that need to be determined are τ_{CI} and τ_{YFP}, which capture the dilution and degradation rate of CI and YFP, respectively, and the four parameters that describe delay in CI production (τ_1, τ_2, n, and β). We used two wild-type temporal expression curves, from ON to OFF and OFF to ON, to fit the above mentioned parameters. We obtained that $\tau_{\mathrm{CI}} = \tau_{\mathrm{YFP}} = 60$ min which corresponds to the dilution of both molecules due to cell growth. This agrees with the fact that degradation of both YFP and CI is much slower than the growth rate of 1h. Furthermore, we obtained $\tau_1 = 70$ min, $\tau_2 = 70$ min, $n = 2$, and $\beta = 5$ which are all within expected range of values.

Model agreement with the data.

To validate the performance of the model, we created 9 P_L promoter mutants, predicted to affect the binding of RNAP and CI in different ways: (a) not to significantly affect the binding of either; (b) primarily impair RNAP binding; (c) primarily impair CI binding; (d) impair the binding of both, RNAP and CI – see Fig 4.2B.
Importantly, all the parameters were obtained from either independent steady state data sets, or from the expression dynamics of the wild-type P_L system. This means that the prediction of these mutants is parameter free as no parameter was fit on this set of data.

To test the goodness of fit, we compute the Pearson correlation coefficient between all time points of the 9 predicted and measured temporal dynamics of gene expression. We obtain that $\rho_{\mathrm{ON}\to\mathrm{OFF}} = 0.90$ and $\rho_{\mathrm{OFF}\to\mathrm{ON}} = 0.90$, showing a very high agreement.

Next, we test the predictive power of TD and MAK models independently. For the TD model, we have already shown on the evalution sample of over 10,000 mutants that the predictive power is high (Pearson correlation coefficient of $\rho_{\mathrm{ON}\to\mathrm{OFF}} = 0.92$ and $\rho_{\mathrm{OFF}\to\mathrm{ON}} = 0.82$, see Section 4.4.2). Furthermore, to test only the goodness of fit of the MAK model, we wanted to remove the potential error in determining the steady state expressions. In other words, if the steady state values are wrong,

this will results in the wrong prediction of the temporal dynamics. Therefore, we normalized all the temporal dynamics curve of both model and experimental data in such a way that they all shared the same starting and ending point. This way, we compared if the model and data trajectories that now share the same steady state points, also share the trajectories – see Fig 2B. The agreement between model and data is very high with Pearson correlation coefficient $\rho_{\mathrm{ON}\to\mathrm{OFF}} = 0.98$ and $\rho_{\mathrm{OFF}\to\mathrm{ON}} = 0.96$.

This quantitatively confirms what is already seen on Fig 4.2B – that while TD gives good prediction and represents the state-of-the-art modeling, the MAK model gives almost perfect prediction of the dynamics with very little deviations from the experimental data.

4.4.3 Calculation of phenotypic landscapes

To compute the phenotypic landscapes, we use all double mutants of the wild-type sequence. The reason why we haven't used single mutants is that for the sequence of length $L = 67$bp, there are only 201 single mutants, most of them having little or no effect on the phenotypes. Alternately, the total number of all double mutants is $\approx 20,000$ which gives high enough sample to explore the properties of the phenotypic landscapes.
See Fig 4.9 for phenotypic landscapes of single mutants.

Phenotypic landscapes with continuous energies

To obtain the binding energies to a given binding site, we use the energy matrix on the binding site sequence. However, to disentangle the effects of the discreteness of sequence and binding site architecture, we explore the phenotypic landscapes where binding energies are not constraint by the sequence but can continuously take any value. This would give us the limits to the phenotypic space that the system can explore due to it's biophysical constraints, excluding the effects coming from the sequence (discrete energy and architecture of binding sites). Even though there are potentially many binding sites for RNAP and CI, we take into account only

bindings to the strongest binding sites: RNAP binding to it's strongest binding site, CI binding to O_{R1}, and CI binding to O_{R1}. In other words, we assume only these three binding sites exist which is a valid approximation as binding to other positions is much less likely. This gives us a three dimensional problem where three binding energies are independently and continuously varied.
Furthermore, to fairly compare the double mutant phenotypic landscapes of continuous energies with energies determined from the sequence, we limit the range of continuous binding energies. The range of the three binding energies is determined by the range that can be explored by double mutants in the original, sequence dependent system. For example, for continuous RNAP binding energy, we find (using the energy matrix) the highest and lowest energy of binding to P_L for double mutants. This range is is between $-5.32k_BT$ and $9.70k_BT$ (where wild-type binding energy is the reference point with energy zero). The range of binding energies for O_{R1} and O_{R2} is $(-0.75, 7.43)k_BT$ and $(-1.69, 7.43)k_BT$, respectively.

Varying overlap between O_{R1} and -10 region

To see the effect of overlap between RNAP and CI binding site, we varied the overlap between O_{R1} and -10 region of the wild-type sequence. Given that the representation of RNAP is in the energy matrix, we can adjust the RNAP energy matrix in a such a way that -10 region is moved, changing the overlap with O_{R1}. Our procedure was the following. If the -10 position was moved by h bp downstream, we increased the spacer between -35 and -10 by h. The spacer penalties were also corrected, allowing now for h larger spacer.
By moving the -10 position, we have effectively changed the size of energy matrix to $4 \times (L + h)$. To keep the binding to wild-type sequence unaffected by this change, we have to adjust the energy wild-type matrix elements. As per our definition, the energy matrix elements representing wild-type sequence have value zero. This means that in each energy matrix column, there is one element with zero energy contribution, representing wild-type nucleotide. The three remaining elements have non-zero value and represent mutational effects to other nucleotides. Therefore, each column in the new energy matrix is adjusted, such that the element representing WT

sequence is assigned zero value, while the remaining three elements in the column are given the three non-zero values. In other words, the wild-type nucleotide in the column is adjusted due to the movement of -10 region by h bp.
This ensures that expression of wild-type sequence would remain the same for both in presence and absence of CI.

Shuffling the elements of the energy matrices

We randomly shuffled the elements of an energy matrix without repetitions. This maintained the original distribution of elements in energy matrix but destroyed any internal structure of the energy matrix. To have a common reference point between different energy matrices, we fixed the elements in the energy matrix that represent the wild-type sequence. This way, the expression of the wild-type sequence was not affected by the shuffle.

Computing the surface area of phenotypic landscape.

A set of points in space doesn't have a volume (or surface). Therefore, to compute the surface area of a phenotypic landscape, we decided to assume that each mutant is represented with a square of edge length a, which is centered around the point of the mutant in phenotypic space. Assuming a square instead of a circle is computationally easier to implement, as surface area of a set of non-disjoint circles is non-trivial task. To compute the surface area of a set of mutants in the phenotypic space, we first represent the phenotypic space with a grid. The size of each tile in the grid must be much smaller than the size of mutant's square a. Next, for each mutant we mark which tiles of the grid are covered by this mutant's square. Doing this for all mutants, the surface area of mutants in the phenotypic space is proportional to the total number of marked tiles.
We measured all phenotypes in their wild-type value (with the exception of OFF expression which was in units of wild-type ON expression), meaning that the relevant scale of phenotypic values was 1. Therefore, our square size for each mutants was set of be $a = 0.01$, representing a small change in phenotypes which

could potentially be explored by intrinsic noise. However, the surface areas of various phenotypic landscapes do not qualitatively change with other values of a – see Fig 4.10.

4.4.4 Evolutionary model

For the evolution of promoter, we use the commonly used Strong-Selection-Weak-Mutation model with single point mutations being introduced into the system. The model assumes that that single mutations are rare, such that at any given time only a single mutation is competing to be fix.
As the average time scale is determined by the arrival of a new mutation, we use the inverse mutation rate as the unit of time.

What configurations lead to expression?

For the wild-type system, there already exist a strong RNAP binding sites with two CI binding sites overlapping it. However, when evolving a complex promoter de-novo, it is not entirely clear which configurations are productive and lead to expression. For example, it is clear that CI binding to O_{R1} stericaly excludes binding of RNAP and therefore represses expression. Similarly, CI binding further upstream of RNAP binding site, it is clear there is no interaction between CI and RNAP, thus allowing RNAP to bind, leading to expression. However, what if CI is bound downstream of RNAP binding site, far enough to allow RNAP to bind to its binding site? Does this configuration – of both RNAP and CI bound – lead to expression? In all results we have assumed that CI bound downstream of RNAP leads to repression.

Fitness function

To describe the fitness function closely around it's global maximum/peak, we can use Taylor expansion around it and write the quadratic term. However, as we would like fitness not to be negative, we instead describe fitness as

$$F \approx 1 - \frac{s}{d} \sum_{i=1}^{d} \left(\frac{p_i}{p_i^*} - p_i^{\mathrm{opt}} \right)^2 , \tag{4.15}$$

where s represents the selection coefficient, the sum goes over different phenotypes i, and p_i^* is the normalizing factor of phenotype p_i that determines the units. $p_i^* = p_i^{\mathrm{WT}}$ for all phenotypes, except for OFF expression which is measured in units of ON WT expression, i.e., $p_{\mathrm{ON}}^{\mathrm{WT}}$, as by construction. p_i^{opt} is the optimal value of p_i with highest fitness, and equals $p_i^{\mathrm{opt}} = 1$ for all phenotypes except for OFF expression which has value of $p_i^{\mathrm{opt}} = 0.0012$. In other words, we measure phenotypes in their wild-type units (with the exception of OFF expression), and would like to evolve them towards the optimal value which is their wild-type.
The sum inside the fitness function can be written as:

$$\sum_{i=1}^{d}\left(\frac{p_i}{p_i^*}-p_i^{\mathrm{opt}}\right)^2 = \overbrace{(\mathrm{ON}-1)^2+(\mathrm{OFF}-0.0012)^2}^{\mathrm{2D}}+ \left.\begin{array}{l} \left(\frac{\mathrm{slope}_{\mathrm{ON}\to\mathrm{OFF}}}{\mathrm{slope}_{\mathrm{ON}\to\mathrm{OFF}}^{\mathrm{WT}}}-1\right)^2+\left(\frac{\mathrm{slope}_{\mathrm{OFF}\to\mathrm{ON}}}{\mathrm{slope}_{\mathrm{OFF}\to\mathrm{ON}}^{\mathrm{WT}}}-1\right)^2+ \\ \left(\frac{\mathrm{lag}_{\mathrm{ON}\to\mathrm{OFF}}}{\mathrm{lag}_{\mathrm{ON}\to\mathrm{OFF}}^{\mathrm{WT}}}-1\right)^2+\left(\frac{\mathrm{lag}_{\mathrm{OFF}\to\mathrm{ON}}}{\mathrm{lag}_{\mathrm{OFF}\to\mathrm{ON}}^{\mathrm{WT}}}-1\right)^2, \end{array}\right\} 6D \qquad (4.16)$$

where ON and OFF are already, by construction, in the units of wild-type ON expression. If evolving only steady state expression, only the first two parts of the sum are taken (marked by brackets with 2D). Alternatively, evolving all 6 phenotypes requires all six contributions, as marked by brackets with 6D.
This makes sure that the optimal value of fitness is $F = 1$, and is $F < 1$ for any other non-optimal phenotypes.

However, the fitness function in Eq 4.15 is a quadratic form, approximating the peak only around the neighbourhood of the peak. To ensure that fitness function is limited between zero and one, we generalize the fitness function to

$$F = \exp\left[-s\frac{1}{d}\sum_{i=1}^{d}\left(\frac{p_i}{p_i^*}-p_i^{\mathrm{opt}}\right)^2\right], \qquad (4.17)$$

which can be approximated by a quadratic form for phenotypes close to the optimal value.

Fixation probability

The fixation probability is given by the Kimura fixation probability

$$p_{\text{fix}} = \frac{1 - e^{-2\Delta F}}{1 - e^{-2\Delta F N}} \tag{4.18}$$

where N is the population size and $\Delta F = \frac{F_{\text{new}}}{F_{\text{old}}} - 1$ the relative change in fitness between old and new allele. As typical bacterial population sizes are relatively large, the denominator will mostly have a small contribution.

Thermodynamic model and different configurations

To compute the probability of expressing state, we must take into account also those configurations that were highly unlikely before, when we were exploring the neighbourhood of wild-type sequence. This means that we need to extend the set of major configurations that we model: i) unbound state, ii) only RNAP bound, iii) only CI bound, iv) CI bound upstream of RNAP, and v) CI bound downstream of RNAP. As mentioned above, the productive states that lead to expression are ii) and iv). We can write the probability of states ii) or iv) occurring as

$$p_E = \frac{w_2 + w_4}{1 + w_2 + w_3 + w_4 + w_5}, \tag{4.19}$$

$$w_1 = 1 \tag{4.20}$$

$$w_2 \propto \sum_{i=1}^{M-L_R+1} [\text{RNAP}] e^{-E_i^R} \tag{4.21}$$

$$w_3 \propto \sum_{i=1}^{M-L_{\text{CI}}+1} [\text{CI}]^2 e^{-E_i^{\text{CI}}} \tag{4.22}$$

$$w_4 \propto \sum_{i=1}^{M-L_{\text{CI}}-L_R+1} [\text{CI}]^2 e^{-E_i^{\text{CI}}} \sum_{i=1}^{M-L_R+1} [\text{RNAP}] e^{-E_i^R} \tag{4.23}$$

$$w_5 \propto \sum_{i=1}^{M-L_R-L_{\text{CI}}+1} [\text{RNAP}] e^{-E_i^R} \sum_{i=1}^{M-L_{\text{CI}}+1} [\text{CI}]^2 e^{-E_i^{\text{CI}}}, \tag{4.24}$$

where w_{1-5} represent Boltzmann weights for states described by i) to v). M, L_R and L_{CI} are the total sequence length, RNAP binding site length, and CI binding site

length, respectively.
All other procedures don't change.

Computation and visualization of phenotypic and fitness landscape

If we re-write the Boltzmann weights as $w_3 = [\mathrm{CI}]^2\bar{w}_3$, $w_4 = [\mathrm{CI}]^2\bar{w}_4$, and $w_5 = [\mathrm{CI}]^2\bar{w}_5$, we can write the probability of expressing as:

$$p_E = \frac{w_2 + [\mathrm{CI}]^2\bar{w}_4}{1 + w_2 + [\mathrm{CI}]^2\,(\bar{w}_3 + \bar{w}_4 + \bar{w}_5)} \tag{4.25}$$

which can be represented as

$$p_E = \frac{K_1 + [\mathrm{CI}]^2 K_2}{1 + K_1 + [\mathrm{CI}]^2\,(K_2 + K_3)} \tag{4.26}$$

where $K_1 = w_1$ represents productive configurations with only RNAP bind, $K_2 = \bar{w}_4$ configuration with CI bound upstream of RNAP, both leading to expression. $K_3 = \bar{w}_3 + \bar{w}_5$ then represents unproductive configurations, not leading to expression – see Eq 4.19.
For a fixed value of (K_1, K_2, K_3) all phenotypes are exactly determined via TD and MAK models. Therefore, using these three parameters, we can characterize the whole fitness landscape as is shown in Fig 4.5G.

Correlating expression with slope and lag

To test how evolution of 2D phenotypes compares with evolution with additional correlations between phenotypes, we substituted lag and slope with an effective correlation to ON and OFF expression. This way we could directly test what effect do correlations have on evolution speed.
Fitness function was modified to:

$$F = \exp\left(-\frac{s}{6}\left[(1 + 4C_{\mathrm{ON}})(\mathrm{ON} - 1)^2 + (1 + 4C_{\mathrm{OFF}})(\mathrm{OFF} - 0.0012)^2\right]\right), \tag{4.27}$$

where $C_{\mathrm{ON}} \in (0, 1)$ and $C_{\mathrm{OFF}} \in (0, 1)$ represent average correlation of expression with other 4 phenotypes (lag and slope in both ON to OFF, and OFF to ON dynamics). For values $C_{\mathrm{ON}} = C_{\mathrm{OFF}} = 0.5$, fitness function takes the form for 2D evolution:

$$F_{\mathrm{2D}} = \exp\left(-\frac{s}{2}\left[(\mathrm{ON} - 1)^2 + (\mathrm{OFF} - 0.0012)^2\right]\right). \tag{4.28}$$

Geometric model

To test how genotype affects the distribution of fitness effects (DFE), we used the evolutionary geometric model to simulate a set of mutants, thus forming a new DFE. We tested two different models: where each mutation had an effect either on phenotypes or Boltzmann weights.

Geometric model on phenotypes. In the geometric model where each mutation was represented by a random change in phenotypes, we varied all phenotypes in the following way. Each mutation had a fixed effect size in the phenotypic space, meaning that the mutant had a fixed Euclidean distance in the phenotypic space from the initial point. This means that vector of changes in phenotypes was described by $\vec{dp} = \sum_i dp_i$, where dp_i is a change in phenotype i and $|\vec{dp}|$ is fixed. Therefore, new value of phenotype i is $p_{\text{initial}} + dp_i$, where p_{initial} is the initial value of this phenotype.
To obtain $\vec{dp}$, we randomly drew numbers in the range of $(-1, 1)$ for all phenotypes, and at the end normalizing the vector to the desired amplitude. We used $|\vec{dp}| = 0.3$.

Geometric model on binding energies. As we can represent our phenotypic landscape with 3 effective Boltzmann weights (Section 4.4.4), we also modeled each mutation as an effective change in the sizes of Boltzmann weights. Similarly as for phenotypes, mutations are represented by a vector $\vec{dp} = \sum_i dr_i$ of fixed size in the Boltzmann weight space. However, as Boltzmann weights take values that vary many orders of magnitude, we decided that the each mutation will have a relative effect on each Boltzmann weight: $K_i^{\text{mutant}} = K_i(1 + dr_i)$, where K_i represents Boltzmann weight i. We constrain $|\vec{dp}|$ to be fixed: $|dp| = 1$.
Relative change in Boltzmann weights can be represented by an additive change in effective binding energies of the 3 effective configurations of the system. We can write each of the three Boltzmann weights as $K_i = e^{-E_i}$, $i \in \{1, 2, 3\}$, with E_i representing the effective binding energy of configuration i. Therefore, as each mutation leads to a relative change in Boltzmann weights K_i, we can write $K_i^{\text{mutant}} = K_i(1 + dr_i)$ and $-E_i^{\text{mutant}} = \log K_i^{\text{mutant}} = \log\left(K_i(1 + dr_i)\right) = \log K_i + \log dr_i = -E_i + dr_i$, if $dr_i \ll 1$.

Computing DFEs for the geometric model

To compute the DFEs of the two above mentioned model, we use the original model (with genotypes) and randomly drew sequences (to avoid any bias), until we found a genotypes with the desired fitness value.
Now, having this Boltzmann weights and phenotypes of this genotype, we compute 10^4 single mutants around it using Geometric model on either phenotypes or effective Boltzmann weights. For each mutant, we randomly drew a vector of fixed size $\vec{dr}$ which represents the change in either phenotypes or Boltzmann weights.
To compare DFE with the original model with a genotype, we have to compute double mutants. We do this by applying the procedure described above twice; first to obtain single mutants, and the second time to get double mutants. In other words, first a random vector $\vec{dr}_1$ is applied on the original phenotypes/Boltzmann weights to get a single mutation, and then a second random vector $\vec{dr}_2$ is applied on the new phenotypes/Boltzmann weights of the single mutation. This gives us a double mutant.
To compute the means and standard deviations of these DFEs, we used 30 sequence with different genotypes for each given fitness value.

Obtaining the architecture of *E.coli* promoters.

Using the data from RegulonDB, we obtained sequences of 2000 promoters and consensus of $\approx$ 3700 TF-BS from *E.coli*. To find the cognate binding sites for each TF, we matched consensus sequences of each TF-BS to different promoters, connecting promoters with TF whole consensus perfectly matches. To find -35 and -10 region of each promoter, we used RNAP energy matrix – the position with highest likelihood for RNAP to bind was considered as RNAP binding site. Furthermore, to be consistent with out experimental system, we used only the data where TF was a repressor, leading to 700 positions of TF-BS relative to RNAP binding site. These were then classified in one of the three architectures. A majority fall in architecture II which coincides with the fact that this architecture evolves fastest in 6D evolution.

4.4.5 Experimental system and measurements

We used a synthetic system based on the Lambda phage switch, in which we decoupled the *cis*- (promoter) and *trans*- (transcription factor) regulatory elements, as previously described in Lagator et al. [Mato *et al.*, 2017]. We removed *cI* and substituted *cro* with *venus-yfp* [Nagai *et al.*, 2002] under control of P_L promoter, followed by a T1 terminator sequence. The O_{R3} site was removed in order to remove the P_{RM} promoter. Separated by a terminator sequence and 500 random base pairs, we placed *cI* under the control of P_{TET}, an inducible promoter regulated by TetR [Lutz and Bujard, 1997], followed by a TL17 terminator sequence. In this way, concentration of CI transcription factor in the cell was under external control, achieved by addition of the inducer anhydrotetracycline (aTc). The entire cassette was inserted into a low-copy number plasmid backbone pZS* carrying a kanamycin resistance gene [Lutz and Bujard, 1997].

We measured the ON→OFF dynamics of gene expression in the wild-type P_L system in the following manner. Six replicates were grown overnight in M9 media, supplemented with 0.1% casamino acids, 0.2% glucose, and 50μg/ml kanamycin. The absence of the inducer aTc indicates that these cells were grown in the ON state overnight. Overnight cultures were diluted 100x, grown for 2h under the same conditions, and then diluted again at 100x. At this point, each replicate population was diluted into two conditions: same as the overnight growth (in this case, ON state); different state to the overnight, in this case achieved by adding 10ng/ml aTc. Fluorescence of growing replicate populations was measured every 15 minutes in Bio-Tek Synergy H1 platereader. The measured fluorescence was always corrected for the autofluorescence of the media. Populations were always grown at 37°C. To measure OFF→ON dynamics, we used the same protocol, but have grown overnight cells in the presence of 10ng/ml aTc. These wild-type P_L measurements served as the basis to derive model parameters.

Sequence
GATAAATATTTATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTGATACTGAGCACATCAGCAG
GATAAATATTTATATCTGGCGGTGTTGACATAAATACCACTGGCGGTGATACTGAGCACATCAGCAG
GATAAATATTTATCTCTGGCGGTGTTGACATAAATACCACTGACGCTGATACTGAGCACATCAGCAG
GATAAATATTTATCTCTGGCGGTGTTGACATAAATACCACTGACGCTGATACTGAGCACATCAGCAG
GATAAATATTTATCTCTGGCGATGTTGACATAAATACCACTGGCGGTCATACTGAGCACATCAGCAG
GATAAATATTTACCTCTGGCGGTGTTGACCTAAATACCACTGGCGGTGATACTGAGCACATCAGCAG
GATAAATATTTATCTCTGGCGGTGTTGTCATAAACACCACTGGCGGTGATACTGAGCACATCAGCAG
GATAAATATTTATCTCTGGCGGTGTTGACCTAAATACCACTGGCGGTGATACTGAGCACATCAGCAG
GATAAATATTTTTCTCTGGCGGTGTTGACATGAATACGACTGGCGGTGATTCTGAGCACATCAGCAG
GATAAATATTTATCTCTGGCGGTGTTGACATAAATACCGCTGGCCGTGACACTGAGCACATCAGCAG

Table 4.1: **Oligo sequences.** Top sequence is the wild-type and all mutations are shown in red.

4.4.6 Experimental validation of model predictions

To validate model prediction, we created 9 P_L promoter mutants. To select the mutations in these mutants, we wanted to impair either: (i) RNAP binding with minimal disruption to CI binding; (ii) CI binding, with minimal disruption to RNAP binding; (iii) the binding of both molecules – see Table 4.1. We ordered oligonucleotides containing the desired mutants from Sigma Aldrich, and cloned them into the wild-type P_L system by restriction/digestion. We verified each cloned mutant by Sanger sequencing. We measured the ON→OFF and OFF→ON dynamics of six replicates of each mutant in the same manner as described for the wild-type P_L system.

4.5 Supporting Information

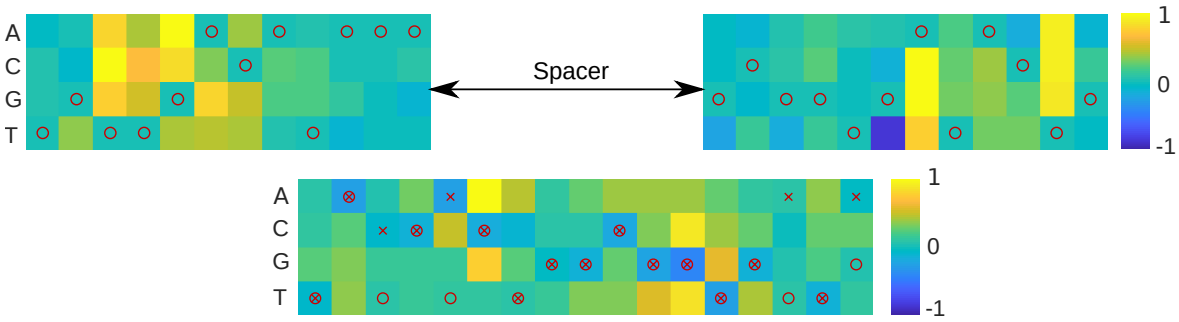

Figure 4.8: **Energy matrices of RNAP (top) and CI (bottom).** As it was shown that flanking region of −35 (left) and −10 (right) RNAP binding site significantly influences the prediction of binding, we also include them in the RNAP energy matrix. We mark the wild-type sequence of the strongest binding site with red 'o' (P_L RNAP binding sites, and O_{R2} CI binding site) and with red 'x' (O_{R1} CI binding site). The unit scale is normalized to be between −1 and 1 which is determined by the largest element in each matrix. The transforming factors to real energy units k_BT for both RNAP and CI EM, α and ι, are determined in section 4.4.2.

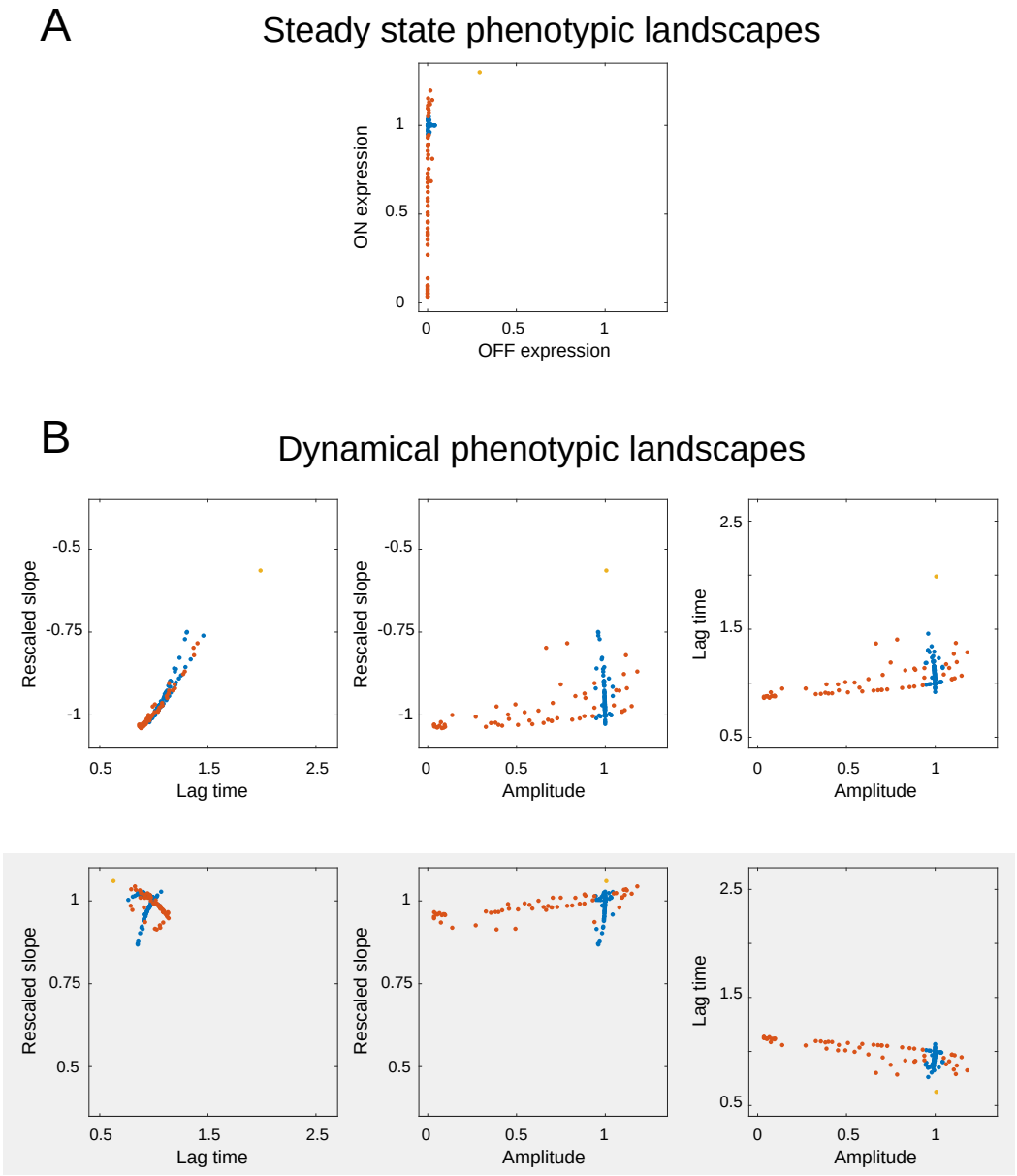

Figure 4.9: **Phenotypic landscapes of single mutants.** Showing the same phenotypic landscape as in Fig 4.3, with the difference that all *single* mutations are show. As seen, single mutations do not explore a large portion of phenotypic space compared to double mutants in Fig 4.3. **A** Phenotypic landscape of all double mutants, each represented by a dot, of steady-state phenotypes. **B** Phenotypic landscape of all phenotypes for ON to OFF dynamics (top) and OFF to ON dynamics (bottom, grayed). To keep the possible number of landscapes low, we use amplitude(=ON-OFF) as a proxy for both steady-state phenotypes (ON and OFF). As slope strongly depends on amplitude (twice the amplitude implies twice the slope), we use rescaled slope(=slope/amplitude), not slope, as a phenotype. Color scheme is the same as in Fig 4.3. All units are in wild-type units, with the exception of OFF expression - the expression (ON and OFF) is in the units of wild-type ON expression.

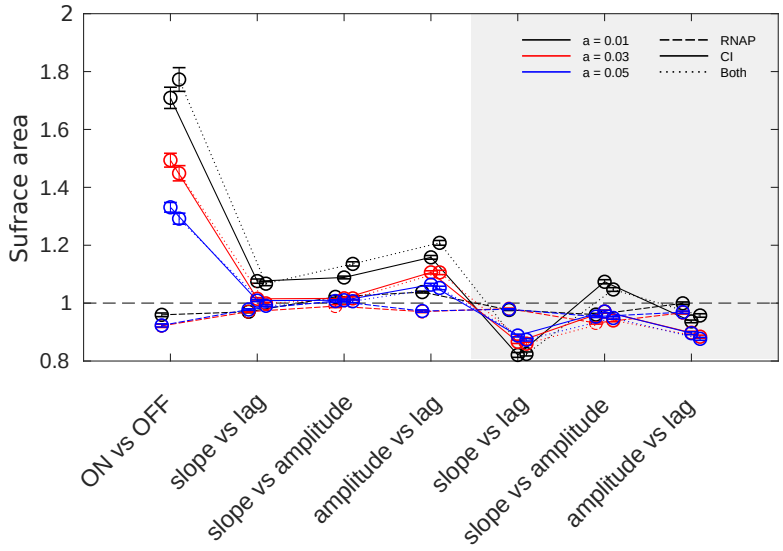

Figure 4.10: **Surface area of phenotypic landscape is independent of the details how it is calculated.** Comparison of surface area of phenotypic landscapes for different 'area of effect' a for each mutant, showing that qualitative results do not change with a – Section 4.4.3. For each value of a (each color), results are qualitatively the same as in Fig 4.4D. Units of surface area are normalized to the surface area of wild-type (un-shuffled) energy matrices - gray dashed line. Error bars represent s.t.d. of 500 replicates. Note that y-axis does not start at zero. Grayed area shows results for dynamics OFF to ON.

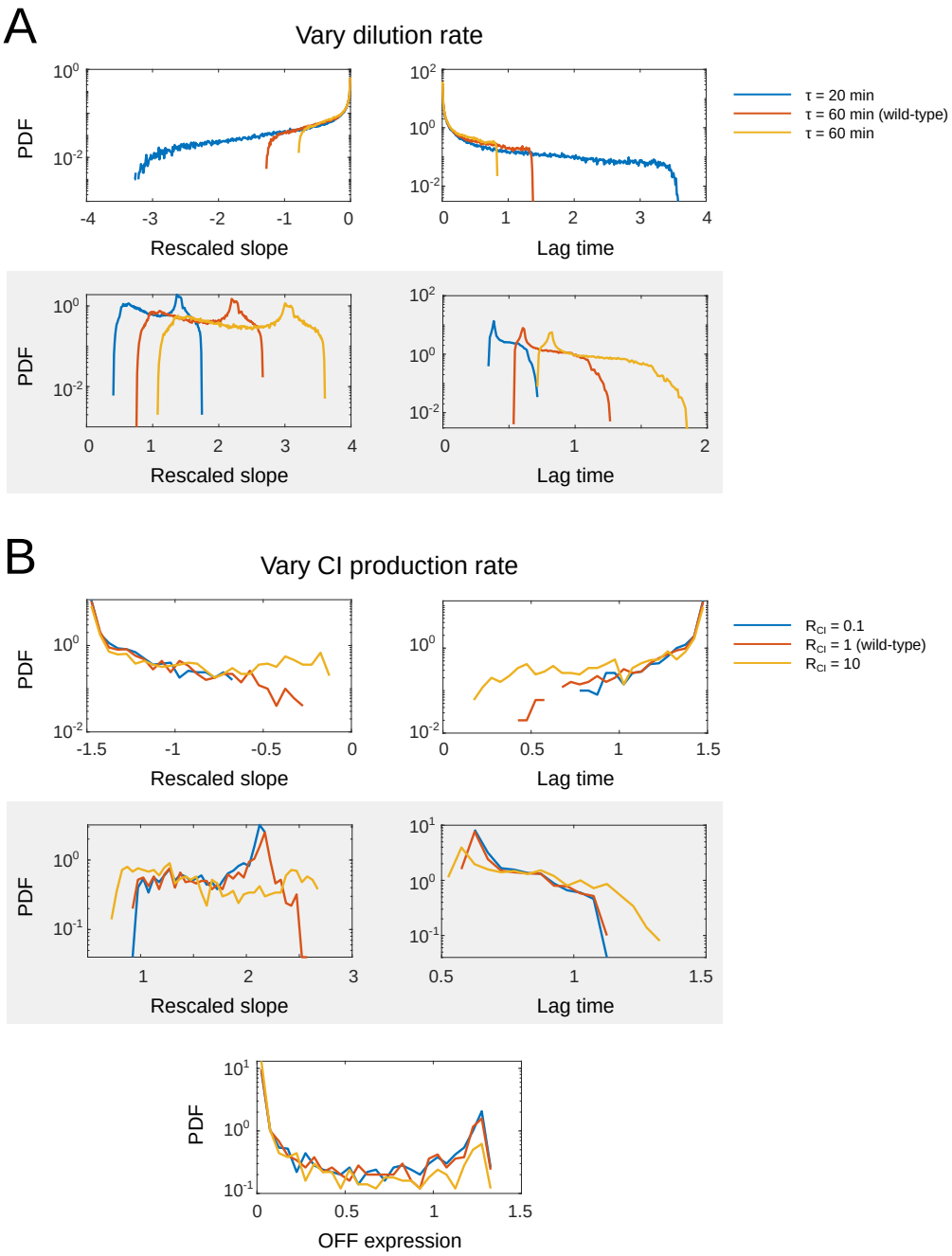

Figure 4.11: **Probability density function (PDF) of phenotypic values that all double mutants are able to explore** for various values of dilution rate **(A)**, and for CI production rates **(B)**. The summary of this figure is shown in Fig 4.4A. As dilution rate doesn't affect steady state expression, the effect of varying dilution rate on them is not shown. Similarly, as CI production rate affects only OFF expression, PDF of ON expression is not shown.

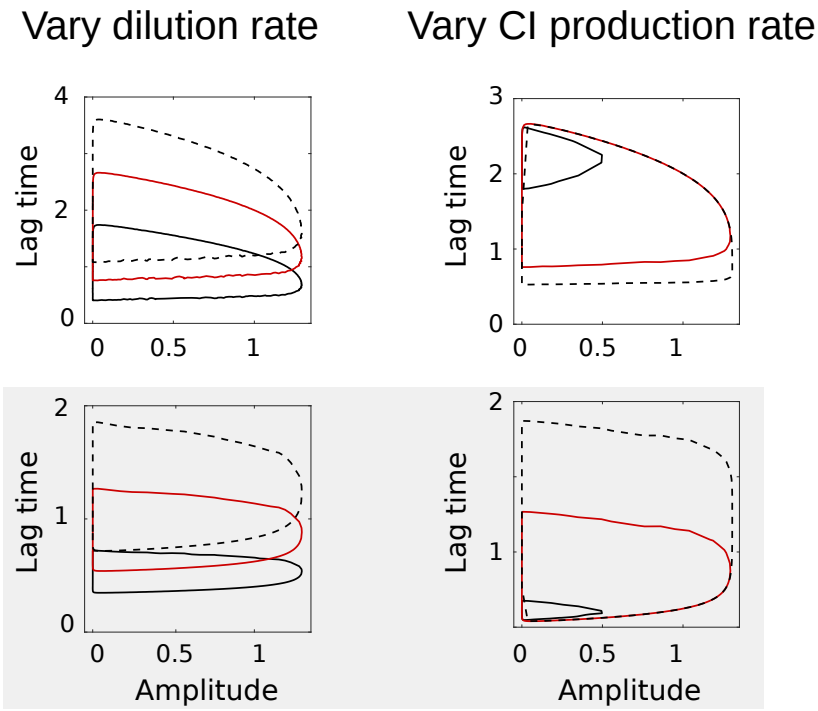

Figure 4.12: **The effect of duplication and CI production rates on lag and rescaled slope does not qualitatively differ.** Boundaries of phenotypic landscapes of double mutants, showing the constraints between pairs of phenotypes more clearly for varying dilution rate or CI production rate. As lag time strongly correlates with rescaled slope, these results are already well described by landscapes of 'rescaled slope vs amplitude' on Fig 4.4B. Full and dashed black envelope represent phenotypic landscape of increased and decreased parameter, respectively. For reference, red envelope shows the landscape with the original parameter value.

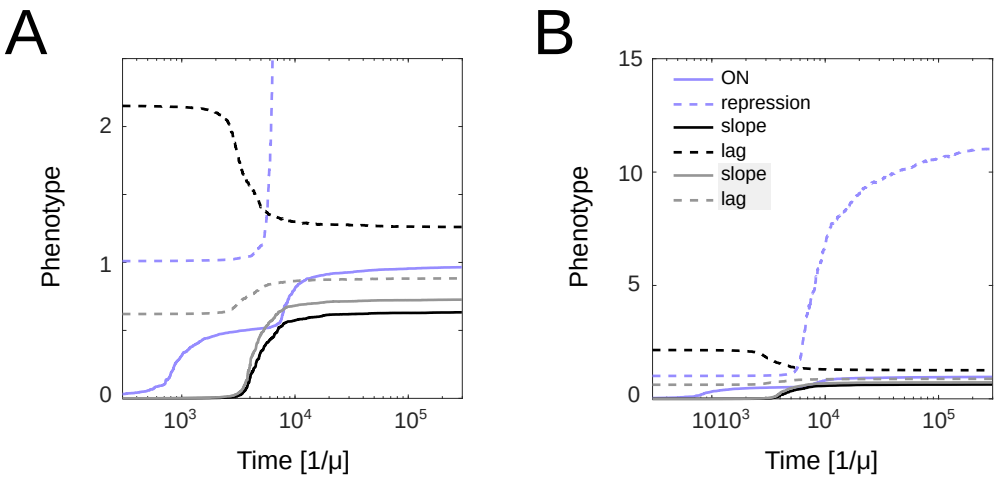

Figure 4.13: **Non-normalized time trajectories of all phenotypes.** To better represent evolving repressor binding site, we plot repression instead of OFF expression. While **(A)** focuses on majority of phenotypes, **(B)** zooms out, showing how repression changes relative to other phenotypes. Resulting trajectories are a median of 2000 replicates. Units of phenotypes are in their wild-type units, with the exception of repression which is the ratio of ON and OFF expression, both in the units of wild-type ON expression. Time units are inverse mutation rate μ. Normalized trajectories are shown on Fig 4.3C.

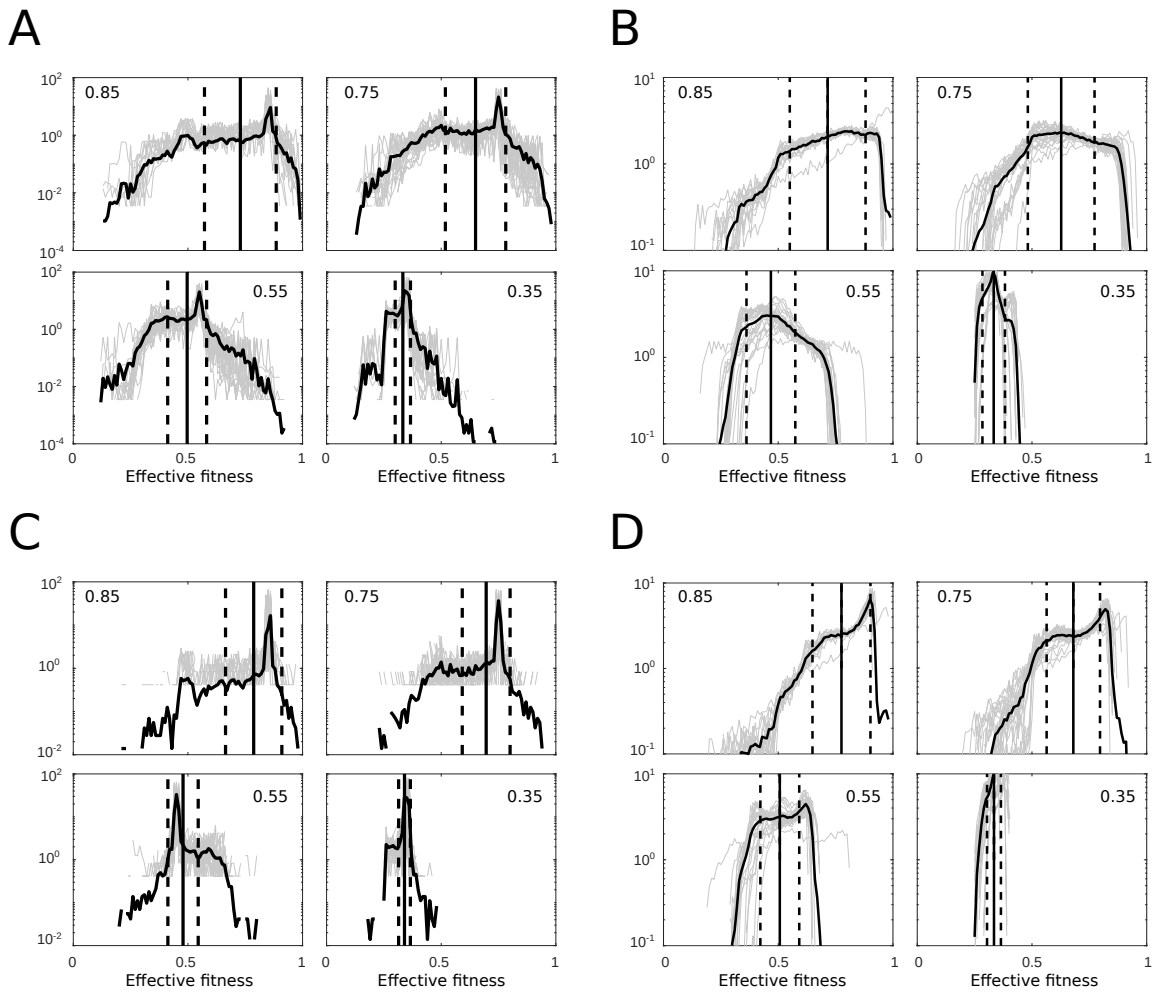

Figure 4.14: **DFE for original model and geometric model on binding energies of single and double mutants.** Distribution of fitness effects for full genotype-phenotype-fitness model (Fig 4.7) for double **(A)** and single **(C)** mutations, and for geometric model on binding energies (Fig 4.7) for double **(B)** and single **(D)** mutations. The value inside each plot represents the fitness value around which DFE was calculated. (A) is identical to the plot on Fig 4.6A. Black line represents a mean over 30 replicate DFEs (shown in gray). Vertical black and dashed lines represent mean and s.t.d. of the mean DFE.

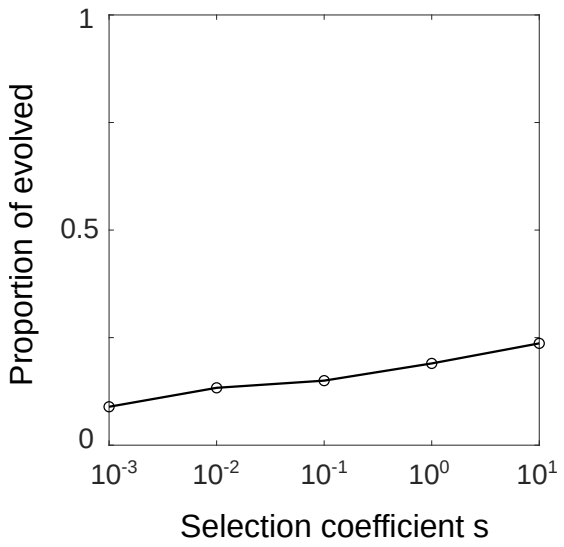

Figure 4.15: **Proportion of evolved sequences increases with selection coefficient, s.** These results were computed for 6D phenotypes, a fitness threshold (above which sequence is considered evolved) of 0.999, for population size $N = 10^6$, and a sample of 300 sequences.

5 Gene amplification as a form of population-level gene expression regulation

Organisms cope with change by employing transcriptional regulators. However, when faced with rare environments, the evolution of transcriptional regulators and their promoters may be too slow. We ask whether the intrinsic instability of gene duplication and amplification provides a generic alternative to canonical gene regulation. By real-time monitoring of gene copy number mutations in *E. coli*, we show that gene duplications and amplifications enable adaptation to fluctuating environments by rapidly generating copy number, and hence expression level, polymorphism. This 'amplification-mediated gene expression tuning' occurs on timescales similar to canonical gene regulation and can deal with rapid environmental changes. Mathematical modeling shows that amplifications also tune gene expression in stochastic environments where transcription factor-based schemes are hard to evolve or maintain. The fleeting nature of gene amplifications gives rise to a generic population-level mechanism that relies on genetic heterogeneity to rapidly tune expression of any gene, without leaving any genomic signature.

Published as Tomanek I*, **Grah R***, Lagator M, Andersson AMC, Bollback JP, Tkačik G, Guet CC. Gene amplification as a form of population-level gene expression regulation. *Nature Ecology & Evolution*. 4(4):612- 625, 2020.

* These authors contributed equally.

Contributions: Tomanek I has done all experimental measurements. Grah R has constructed the model and has done model analysis. Andersson AMC has done single cell analysis. Tomanek I and Grah R have done data analysis and interpretation.

Some changes have been made to the text in order to integrate it into this thesis.

5.1 Introduction

Natural environments change periodically or stochastically with frequent or very rare fluctuations and life crucially depends on the ability to respond to such changes. Gene regulatory networks have evolved into an elaborate mechanism for such adjustments as populations were repeatedly required to cope with specific environmental changes [Moxon *et al.*, 1994; Savageau, 1974; Gerland and Hwa, 2009]. Gene regulation requires many dedicated components – transcription factors and promoter sequences on the DNA – for information processing to occur. However, due to low single base-pair mutation rates, complex promoters cannot easily evolve on ecological time scales [Tuğrul *et al.*, 2015; Berg *et al.*, 2004].

Gene copy number mutations might provide a fundamentally different adaptation strategy, which neither depends on existing regulation nor requires regulation to evolve. Gene duplications arise by homologous or illegitimate recombination between sister-chromosomes. Depending on the genomic locus, duplication rates (k_{dup}) can vary between 10^{-6} and 10^{-2} per cell per generation in bacteria [Anderson and Roth, 1981; Reams *et al.*, 2010; Pettersson *et al.*, 2009; Sun *et al.*, 2012]. This means that a typical bacterial population will contain at any given time a large fraction of cells with a duplication somewhere on the chromosome [Sun *et al.*, 2012; Roth *et al.*, 1996]. Due to the long stretches of homology, duplications are highly unstable: at rates (k_{rec}) between 10^{-3} and 10^{-1} per cell per generation [Reams *et al.*, 2010; Pettersson *et al.*, 2009] *recA*-dependent unequal crossover of the repeated sequence leads to deletion of the second copy – restoring the ancestral state – or to further amplification (Fig 5.1a). If a gene is under selection for increased expression [Nicoloff *et al.*, 2019; Bass and Field, 2011; Albertson, 2006], the process of gene duplication and amplification (GDA) can dramatically increase organismal fitness by increasing gene copy numbers. Due to their high rates of formation, amplifications provide fast adaptation and facilitate the evolution of functional innovation [Andersson and Hughes, 2009]. In contrast, their high rate of loss makes amplifications transient and difficult to study [Andersson and Hughes, 2009]. Surprisingly, until recently it has not been appreciated how this high loss rate impacts the distribution of

copy numbers and associated expression levels in the population, a phenomenon causing antibiotic heteroresistance [Nicoloff *et al.*, 2019; Hjort *et al.*, 2016]. Moreover, amplifications have been studied only under constant selection for increased expression [Näsvall *et al.*, 2012; Elde *et al.*, 2012], while natural environments are rarely ever constant. While a large body of work suggests that phenotypic heterogeneity serves as an adaptation to fluctuating environments [Kussell and Laibler, 2005; Veening *et al.*, 2008], it is not known how the genetic heterogeneity resulting from copy number mutations impacts survival in fluctuating environments.
Here, we ask whether the intrinsic genetic instability of gene amplifications allows bacterial populations to tune gene expression in the absence of evolved regulatory systems. To test this idea experimentally we devised a system of fluctuating environmental selection, which selects for the regulation of a model gene. In this fluctuating environment, we track, in real time, copy number mutations in populations as well as single cells of *Escherichia coli*. Using this system, we test the ability of GDA to effectively tune gene expression levels on ecological timescales, when environmental perturbations occur at rates far too fast for transcriptional gene regulation to emerge de novo.

5.2 Results

5.2.1 Amplification-mediated gene expression tuning (AMGET) occurs in fluctuating environments

To test whether GDA can act as a form of gene regulation at the population level, we experimentally introduced environmental fluctuations, such that a given level of expression of a model gene is advantageous in one, but detrimental in another environment. As the model gene, we used the dual selection marker *galK*, encoding galactokinase. Expression of *galK* is necessary for growth on galactose, but deleterious in the presence of its chemical analogue, 2-deoxy-galactose (DOG) [Barkan *et al.*, 2011]. Using *galK* with an arabinose-inducible promoter, we mapped the relationship between galK expression level and growth in (i) galactose, which selects for high galK

expression levels and which we refer to as the 'high expression environment'; and in (ii) DOG, which selects for low *galK* expression and which we refer to as the 'low expression environment' (Fig 5.1b). In order to establish a strong selective tradeoff between high and low expression, we used 0.1% galactose for the high expression environment and 0.0001% DOG for the low expression environment in all experiments.

We then constructed a reporter gene cassette to monitor expression and copy number changes of *galK* (Fig 5.1c) based on a previously described construct [Steinrueck and Guet, 2017]. In this construct, *galK* is not expressed from a promoter but harbors p_0, a randomized 188 bp nucleotide sequence matching the average GC content ofÂă*E. coli* instead [Steinrueck and Guet, 2017]. This allowed for the selection of increased expression of *galK*. The reporter cassette harbors two fluorophores that allowed us to distinguish the two principal ways of increasing *galK* expression in evolving populations: promoter mutations and copy number mutations (Fig 5.1c). The promoterless *galK* gene is transcriptionally fused to a yellow fluorescence protein (*yfp*) gene, which reports on *galK* expression. Directly downstream, but separated by a strong terminator sequence, an independently transcribed cyan fluorescence protein (*cfp*) gene provides an estimate of the copy number of the whole cassette (Fig 5.6a). We inserted this cassette into the bacterial chromosome, close to the origin of replication (*oriC*) – a location with an intermediate tendency for GDA [Steinrueck and Guet, 2017]. However, our results also hold for a second locus, which is flanked by two identical insertion sequence (IS) elements and has a much higher tendency for GDA [Steinrueck and Guet, 2017] (Fig 5.9).

The ancestral strain carrying the promoterless *galK* construct does not visibly grow in the high expression environment. After one week of cultivation at 37°C, mutants with increased *galK* expression appeared (Fig 5.6b). We randomly selected one evolved clone with increased CFP fluorescence ('the amplified strain') and analyzed it in detail (see Section 5.4) to confirm its amplification. This amplified strain was then used for further experiments in alternating environments (Fig 5.2a-c).

In all three alternating regimes, which change on a daily timescale, mean CFP levels of 60 replicate populations of the amplified strain tracked the environments for the

full duration of the experiments. The adaptive change in *galK* copy number (Fig 5.2b) occurred within the imposed ecological timescale, rapidly enough to maintain population growth given the daily dilution bottleneck under all three alternating selection regimes (Fig 5.8a). We confirmed the observed changes in copy number using whole genome sequencing (Fig 5.7b). To understand these population-level observations, we monitored changes in expression of *galK* and *cfp* at the single cell level for two consecutive environmental switches (Fig 5.2c). Expression of *galK-yfp* (Fig 5.8b) was tightly correlated with the observed changes in gene copy number (Fig 5.8c), indicating that gene expression was effectively tuned by GDA. We refer to this phenomenon as amplification-mediated gene expression tuning (AMGET).

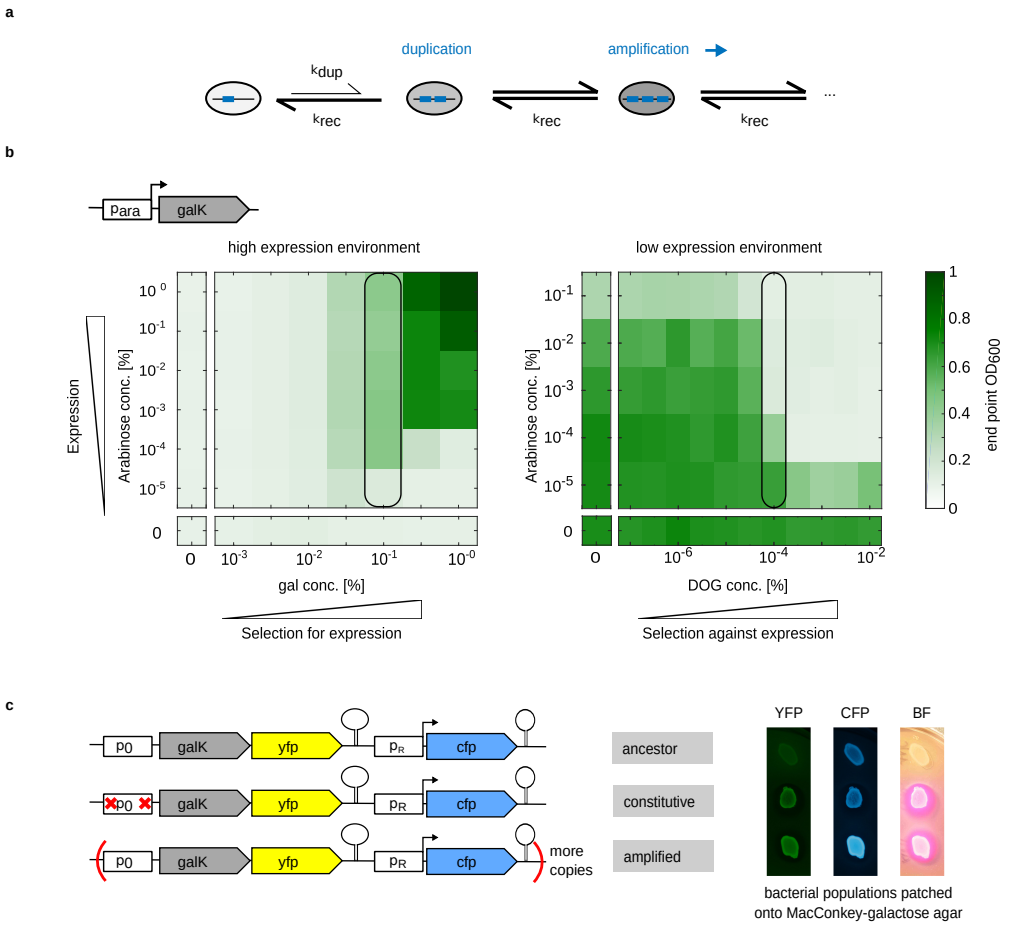

Figure 5.1: (Continued on the following page.)

Figure 5.1: **An experimental system for monitoring gene copy number under fluctuating selection in real time. (a)**, Gene duplication and amplification (GDA). Genomic loci duplicate at rate (k_{dup}) 10^{-6} – $10^{-}2$ per cell per generation. The two gene copies oriented in tandem provide long stretches of identical sequence allowing for homologous recombination at rate (k_{rec}) 10^{-4} – 10^{-1} per cell per generation with recA-dependent unequal crossover leading to further duplication (amplification) or deletion. Grey shading of cells symbolizes the amount of gene product made: increases in copy number result in increased gene expression. **(b)**, Schematic of chromosomal cassette used. Expression of the selection marker, galK, is driven by an arabinose-inducible promoter (para). Growth (as measured by end point OD_{600}) in a 2D gradient of arabinose with galactose (high expression environment) or DOG (low expression environment), respectively. Boxes mark concentrations of 0.1% galactose and 0.0001% DOG, which result in a strong selective tradeoff between high and low expression and were used for further experiments. **(c)**, Schematic showing *galK* reporter cassette (p_0 = random sequence/'non-promoter', pR = strong constitutive lambda promoter, terminator sequences downstream of yfp and cfp, respectively) and genetic changes of strains evolved in the high expression environment with resulting phenotypes on MacConkey galactose agar. Both evolved strains show increased *galK-yfp* expression over the ancestral strain (YFP) and the ability to grow on galactose (BF = bright field image, white versus pink colonies). The amplified strain shows increased CFP fluorescence (CFP) over the ancestral and the constitutive strain, indicating a gene copy number increase.

5.2.2 AMGET depends on selection acting on a gene copy number polymorphism

The rapid population dynamics observed during environmental switches (Fig 5.2c) might simply be explained by selection acting on gene copy numbers with different fitness (Fig 5.2d; Section 5.5.1). We therefore hypothesized that AMGET occurs because of the intrinsic genetic instability of gene amplifications, which continuously and rapidly generate copy number polymorphisms that selection can act on. Re-streaking a single bacterial colony of the amplified strain resulted in colonies with different CFP levels, sometimes with sectors of different CFP expression levels within individual colonies (Fig 5.3a), demonstrating the intrinsic genetic instability of the amplification. Importantly, this genetic instability is dependent on homologous recombination, as a Δ*recA* derivative of the amplified strain failed to show a decrease in CFP fluorescence (and thus copy number) in response to increasing concentrations

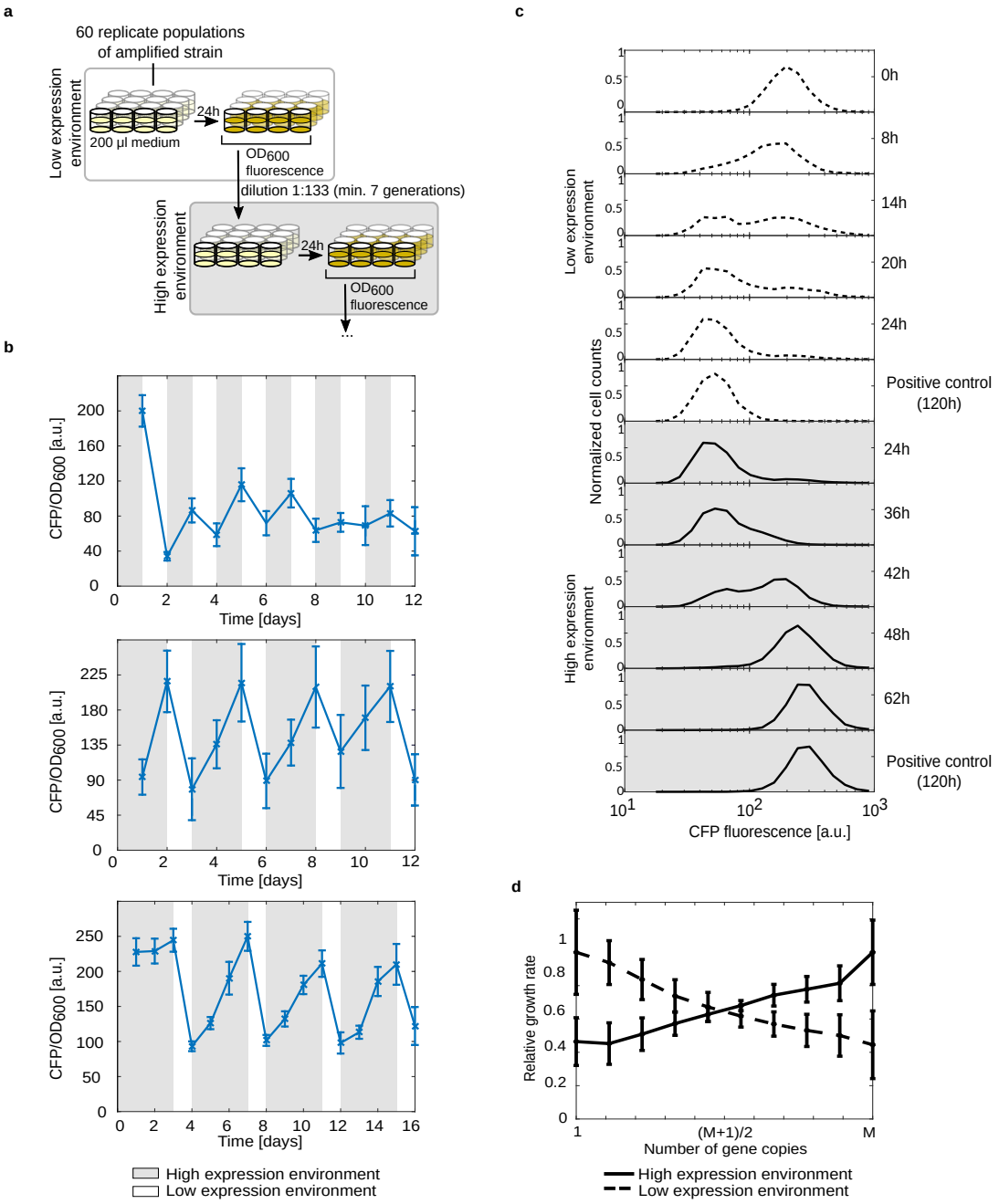

Figure 5.2: (Continued on the following page.)

of DOG (Fig 5.8d). Similarly, Δ*recA* populations were not able to track fluctuating environments as their *recA* wild-type counterparts did (Fig 5.8e).

To determine the rate at which copy number polymorphisms are generated in an amplified population, we followed individual bacteria over ~40 generations in a mother-machine microfluidic device [Bergmiller *et al.*, 2017; Wang *et al.*, 2010]

Figure 5.2: **Amplification-mediated gene expression tuning (AMGET) occurs in fluctuating environments. (a)**, Experimental design of alternating selection in 96-well plate batch cultures, with a daily dilution of 1 : 133. A minimal duration of 24h per environmental condition (no shading = low expression environment, grey shading = high expression environment) allows measuring OD_{600} and fluorescence in populations that have reached stationary phase after dividing at least seven times after their last dilution. **(b)**, Alternating selection of 1 day - 1 day, 2 days - 1 day and 3 days - 1 day in high and low expression environment, respectively. Normalized CFP fluorescence as proxy for gene copy number of 60, 48 and 60 populations of the amplified strain. Error bars represent standard deviation (SD) over all populations. **(c)**, Flow cytometry histograms (one of six replicates from two independent experiments; see d. for an overview of the full dataset) following the adaptation of an amplified bacterial population to low and high expression environments. Positive controls represent populations grown in respective environment for 5 days. **(d)**, Fitness as a function of copy number in the two environments. Growth rates relative to those of maximally adapted populations (positive controls in c) as a proxy for fitness were calculated from the population's shift in CFP fluorescence over time (see Section 5.4). M denotes the maximum copy number, which we estimate to be approximately 10 (see bulk measurements of M in Fig 5.6a and Fig 5.7a, and single cell-based measurements in Fig 5.10b). Note that results do not depend on the precise value of M). Error bars represent the standard deviation of six replicates from two independent experiments.

and monitored their CFP levels. Mutations in copy number were clearly visible as changes in CFP fluorescence of the mother cell. In approximately 35% of cases, these changes were accompanied by a reciprocal fold-change of fluorescence in the daughter cell (Fig 5.3b) as expected from unequal crossover [Reams and Roth, 2015]. In order to quantify the combined rate of copy number gain and loss events by homologous recombination, we analyzed the fluorescence time trace of 1089 mother cells. 55% of traces exhibit constant levels of CFP fluorescence (Fig 5.3c – panel 1) indicating stable inheritance of copy number. In about 7% of traces, the constant level of CFP is interrupted by a sudden decrease or increase (Fig 5.3c – panel 2-3). The corresponding fold-changes of fluorescence are consistent with gains or losses of entire copies of *cfp*. We estimated the lower bound for the average number of copy number mutations, k_{rec}, to be 2.710^{-3} per cell per generation, by automatically selecting only clear step-wise transitions in fluorescence, which are indicative of single copy-number mutation events (Section 5.4, Fig 5.10). Interestingly, 34% of all

traces (Fig 5.10c) exhibit more complex behaviors (Fig 5.3c – panel 4) and cannot be explained in terms of single step transitions.
Complex traces are expected to contain more than one duplication or deletion event even under the expectation that copy number variations are independent events (Fig 5.10d). In addition, it is conceivable that copy number mutations are not independent, i.e., an increased probability exists for a second mutation after the first copy number increase occurred. However, we cannot exclude the possibility that most of the complex traces are due to expression noise of one or both fluorophores, especially since CFP expression noise increases with copy number. Moreover, microfluidics experiments showed transient growth defects visible as filamentation. Given that the amplification includes the origin of replication (*oriC*), complex traces might in part result from replication issues. Transiently stalled replication forks could result in an overproduction of CFP relative to mCherry, which is located at phage attachment site *attP21*, almost opposite on the *E.coli* chromosome. Thus using only single clear step-wise transitions provides a very conservative lower bound for the rate of copy number mutations.

5.2.3 AMGET requires continual generation of gene copy number polymorphisms

Because the mechanism behind AMGET is selection acting on copy number polymorphism, we asked whether it differs from selection acting on single nucleotide polymorphisms (SNPs). To do so, we artificially created a polymorphic population comprised of an equal ratio of two strains – the ancestral strain with no detectable *galK-yfp* expression and a strain with two SNPs in p_0 (Fig 5.1c) resulting in constitutive expression of *galK* (Fig 5.4a). Importantly, this 'co-culture' contained standing variation in galK expression, but because it is not due to amplification, variation is not replenished at high rates. While the 'co-culture' population tracked short-term environmental fluctuations in a manner similar to the amplified population (Fig 5.4b), the long-term dynamics of the two populations were crucially different. Despite being grown from a single cell, the amplified population was able to respond to

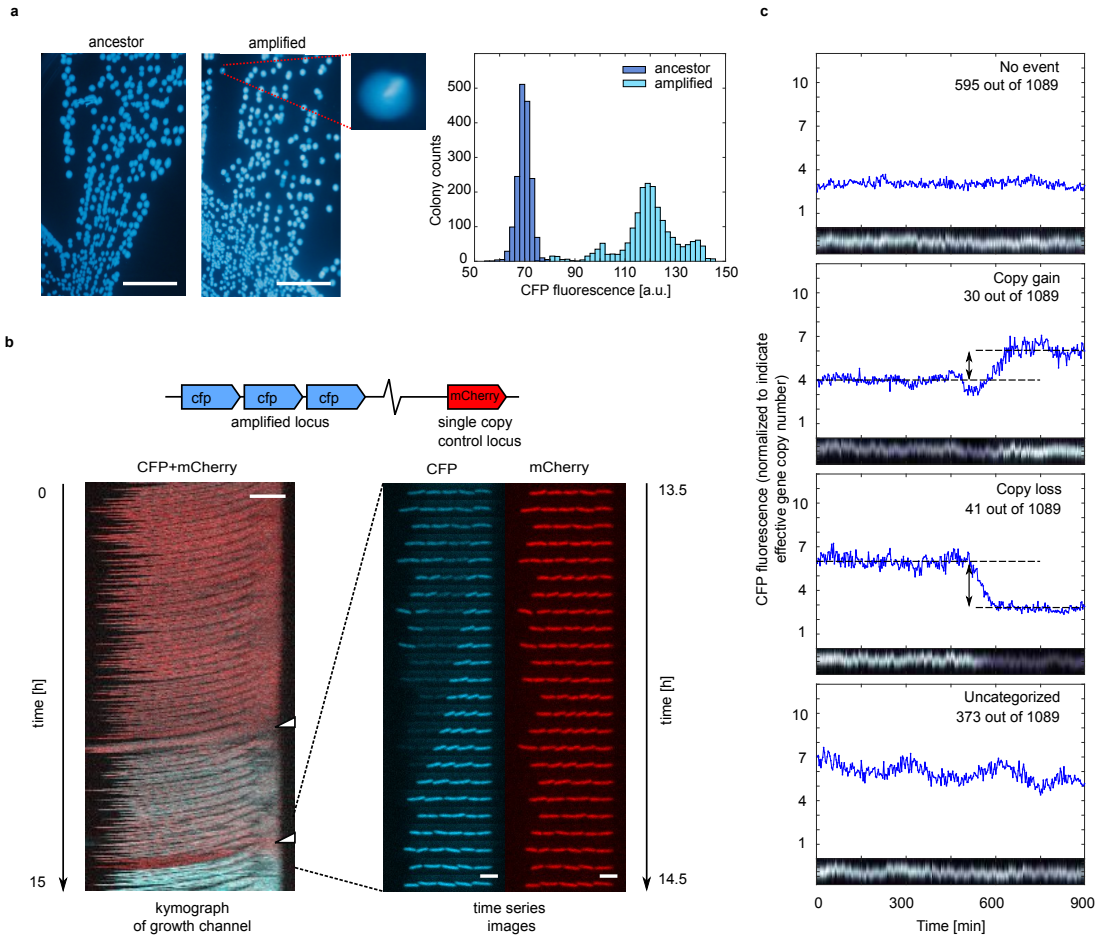

Figure 5.3: (Continued on the following page.)

environmental change rapidly after being maintained in a constant high expression environment for increasingly longer periods (Fig 5.4c). The 'co-culture' population, in stark contrast, progressively lost the ability to respond to sudden environmental change (Fig 5.4d). While standing variation in the 'co-culture' provided some ability for a population to adapt in the short run, it is only replenished at the rate of point mutations. Hence, this variation – as well as the ability to adapt – is depleted by prolonged selection as the genotype with higher fitness goes to fixation in the population.

Figure 5.3: **High-frequency deletion/duplication events in the amplified locus create gene copy number polymorphism in populations. (a)**, Re-streaks of a single bacterial colony on nonselective agar. Ancestral strain bearing a single copy of cfp (left), amplified strain (middle) colonies display sectors of different CFP fluorescence (inset). Scale bars, 10 mm. Histogram of single-colony mean CFP intensities obtained by resuspending and diluting five ancestral and amplified colonies, respectively (right). **(b)**, The amplified strain carrying a single copy of mCherry in a control locus (top) was grown in a microfluidics device to allow tracking of cell lineages in the absence of selection. Overlay of kymographs of CFP and mCherry fluorescence for one microfluidics growth channel (left). Two recombination events are visible as pronounced changes in CFP relative to mCherry fluorescence (white arrows). Time series images of CFP and mCherry fluorescence (right) of the same channel during the second amplification event. An increase in CFP fluorescence of the mother cell (rightmost position in the growth channel) occurs concomitantly with reciprocal loss of CFP fluorescence in its first daughter cell. As mother and daughter cell divide again, their altered level of CFP fluorescence is inherited by their respective daughter cells. mCherry fluorescence of the control locus stays constant during the recombination event. Scale bars, 5μm. **(c)**, Examples of single-cell time traces (kymographs and CFP fluorescence sampled from the mother cell) for four representative behaviors: constant expression, stepwise increase and decrease in expression, and complex expression changes. Frequencies of each behavior across 1089 channels from three independent experiments are shown in figure panels.

5.2.4 AMGET is a general and robust mechanism

The experimental results have qualitatively shown that both, gene copy number polymorphism and selection acting on it, are necessary for AMGET to occur. Using population genetics theory, we developed a generic mathematical model to quantitatively predict the observed experimentally observed population dynamics (Fig 5.2b). The model describes how gene copy number changes over time in a population under selection. Each copy number is treated as a distinct state, and these states differ with respect to growth rates in each of the two environments. Duplication and amplification events are the only source of transition between states. Importantly, all model parameters (the strength of selection and the rate at which the copy-number polymorphism is introduced as shown in Fig 5.1a) are obtained from independent measurements (Table 5.2). Thus, without specifically fitting any parameters, the generic model fully captured the experimentally observed dynamics

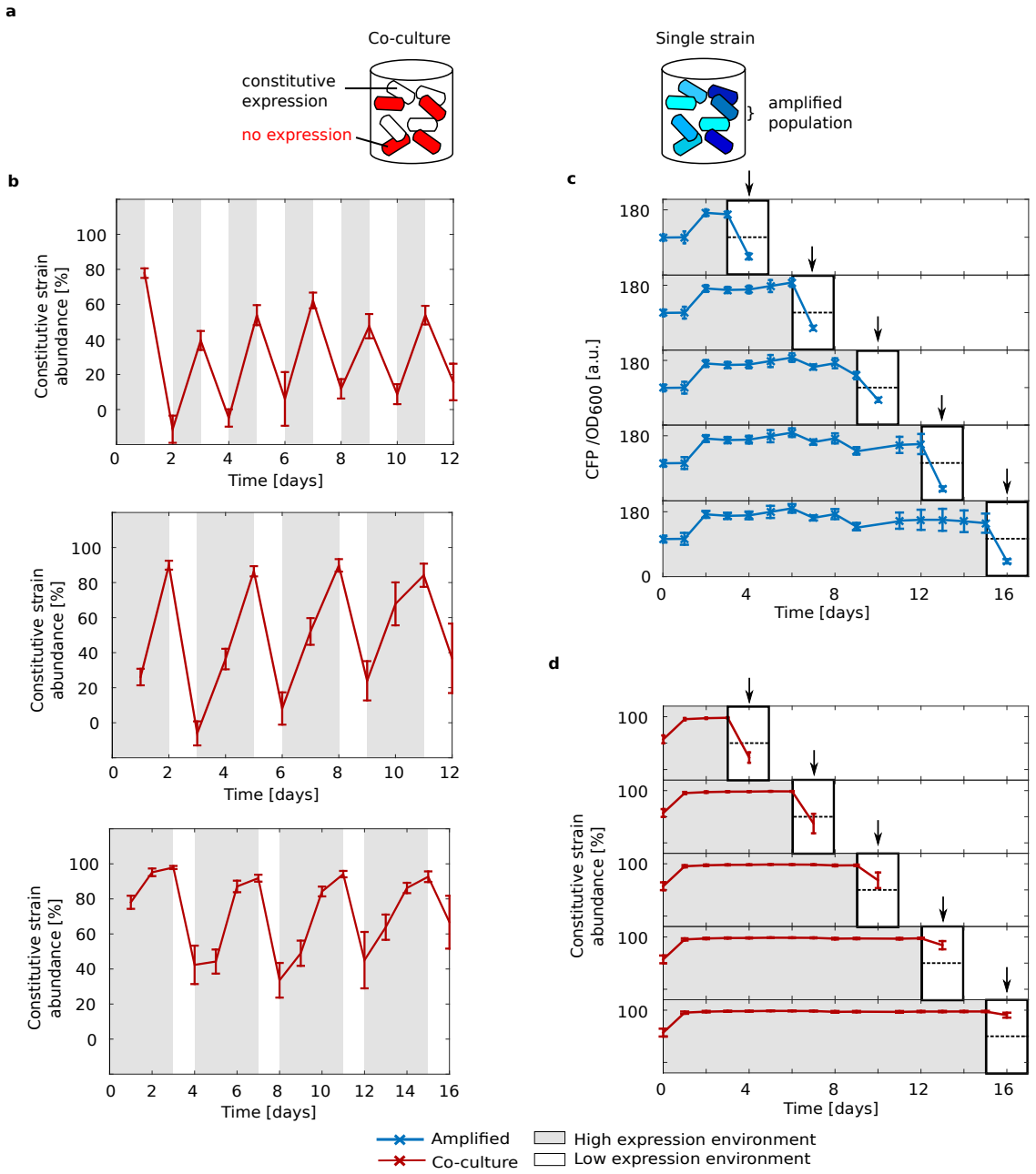

Figure 5.4: (Continued on the following page.)

of AMGET (Fig 5.5a, Fig 5.11a). The good fit between model and experimental data meant that we could use the model to expand the understanding of the basic conditions under which AMGET can act as an efficient de facto mechanism of population-level gene regulation.

Qualitatively, the model revealed that for a population to respond to environmental change at all, two conditions must be met: (i) constant introduction of gene copy

Figure 5.4: **AMGET requires continual generation of gene copy number polymorphisms. (a)**, Schematic of a co-culture composed of the ancestral strain without galK expression and a strain with two SNPs in p_0 (Fig 5.6C) resulting in high galK expression (left). Fluorescently labeling the ancestor allows monitoring relative strain abundance (Section 5.4). A population consisting of a single amplified strain (right) contains cells with different galK copy numbers and, accordingly, expression levels. **(b)**, Alternating selection following the scheme 1 day - 1 day, 2 days - 1 day and 3 days - 1 day in high and low expression environment, respectively. Constitutive strain abundance of 18 co-culture populations tracks environments, with the non-expressing strain being abundant in the low expression environment and the constitutive strain being abundant in the high expression environment. Error bars represent the SD of 18 replicates. **(c-d)**, To estimate a populationÂťs ability to respond to a change in the environment, periods of increasing length spent in the high expression environment are followed by one day in the low expression environment. **c**, Copy number of amplified populations as measured by CFP fluorescence is adjusted to the low expression environment (black arrows) even after prolonged growth in the high expression environment. **(d)**, In contrast, response of the co-culture to the low expression environment after prolonged growth in the high expression environment decreases with time spent in the high expression environment. The mean response on day 16 (1.11 for co-culture, 4 for amplified) differs significantly ($p < 10^{-3}$, two-sided t-test) between populations of co-culture (d) and amplified (c) (see Section 5.4). Error bars represent the SD of 36 replicates.

number variation (i.e. non-zero duplication/recombination rate), and (ii) selection acting on it. If either of these are not present, the population is not able to maintain any long-term response to environmental change.

In order to more quantitatively examine the environmental conditions under which a population can respond to environmental change through AMGET, we defined the response R as the maximum fold change in gene expression before and after an environmental change.

We used the model to expand the range of environmental durations beyond those tested in experiment. In periodic environments, we find a sharp, switch-like transition from no response to full response for environments that switch typically on a day or longer timescale (Fig 5.5b). In stochastically fluctuating environments, the transition is more gradual (Fig 5.5c), yet no less effective. Furthermore, AMGET maintains its efficiency to tune gene expression in bacterial populations over order-of-magnitude variations in the duplication and recombination rates, as well as for any fitness cost

of expression (Fig 5.12).

5.2.5 AMGET tunes gene expression levels when transcription factor-based schemes are hard to evolve or maintain

Canonical gene regulation is unlikely to evolve or be maintained when a population is exposed to an almost constant environment that is sporadically interrupted by a rare environmental perturbation [Gerland and Hwa, 2009]. We tested if AMGET might provide a generic mechanism of regulating expression under such conditions, by asking how long a population that is fully adapted to one environment needs for responding to a step-like environmental change (Fig 5.5b top and side part of heat map; Fig 5.11b). Our model results showed very rapid responses to step-like environmental changes on the order of one to six days, for all biologically relevant parameter values of amplification and duplication rates, as well as fitness cost of expression (Fig 5.5d; Fig 5.11c-e). AMGET is also a viable mechanism for practically any population size, especially for typical bacterial ones, although its efficiency drops for small populations (Fig 5.11f). Therefore, AMGET efficiently tunes gene expression levels across a wide range of environments where transcription factor-mediated regulation would take prohibitively long to evolve [Tuğrul *et al.*, 2015; Berg *et al.*, 2004].

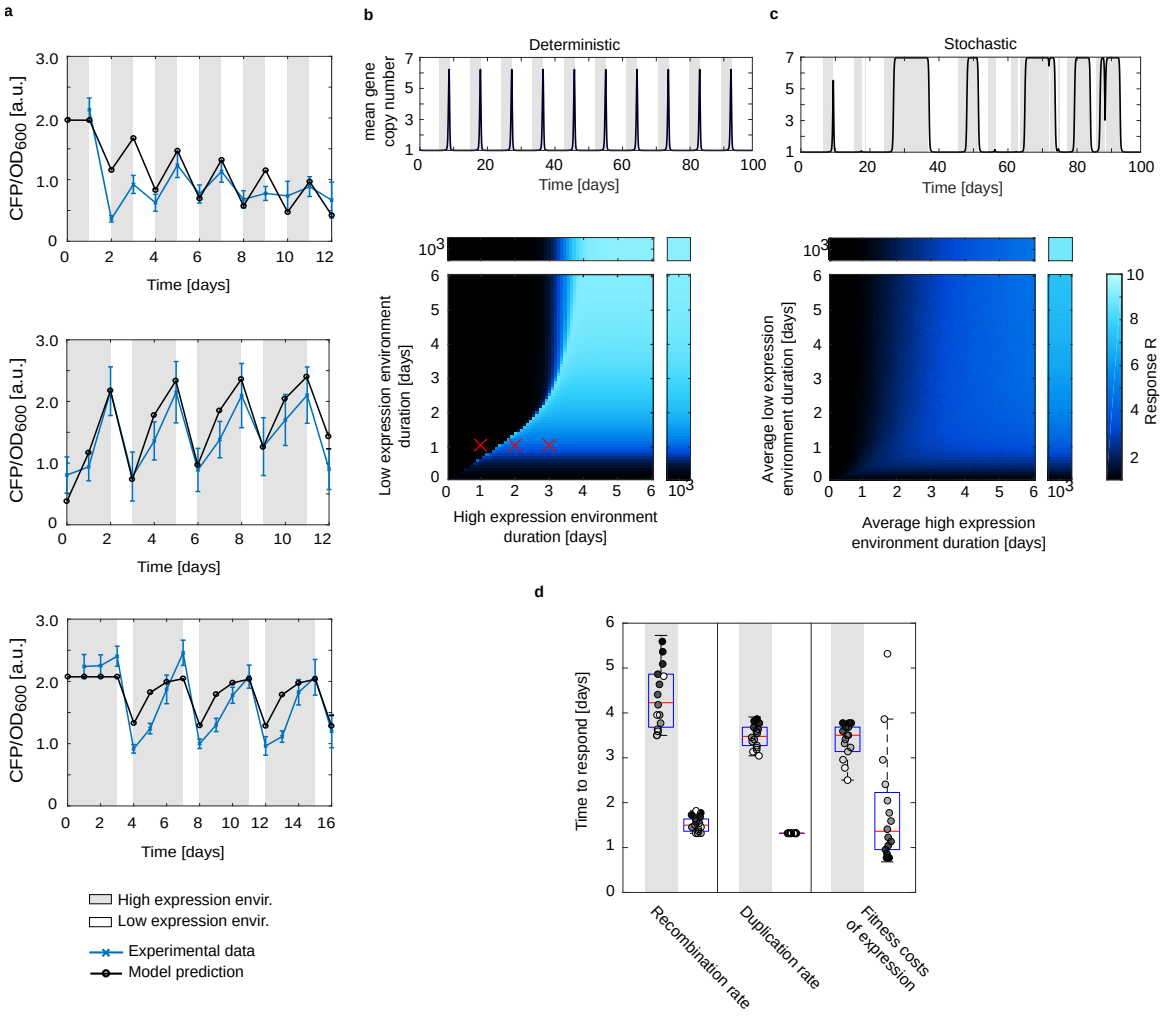

Figure 5.5: (Continued on the following page.)

Figure 5.5: **AMGET is a robust strategy for population level gene expression tuning across a range of environments. (a)**, Comparison of model predictions (with all parameters derived from independent calibration experiments; see Section 5.4) and experimental data for three different environmental durations. Pearson correlation between data and model: 0.72 (top), 0.92 (middle), 0.87 (bottom). See Fig 5.11a for parameter sensitivity. Error bars represent standard deviation (SD) over of 60, 48 and 60 bacterial populations, respectively. **(b-c)**, Top: example of gene expression time trace for deterministic (b) and stochastic (c) environment durations. Bottom: response R (maximum expression fold change before and after the environmental change), shown in color, as a function of the two environment durations. Red crosses in b mark environments shown in a. The gradual increase in response in c occurs because of averaging across responses, which are deterministic for each individual environmental transition (c top). **(d)**, Variation of response time when uniformly sampling sets of parameters (black circles) in the range of $10^{-4} - 5 \cdot 10^{-2}$, $10^{-5} - 10^{-3}$, and $0.1 - 1$ for recombination rate, duplication rate, and fitness costs of expression, respectively (Fig 5.11c-e). The plot shows the median (red line) with the 25th and 75th percentile (blue box). In all plots, when not varied, we use recombination and duplication rates $k^0_{\text{rec}} = 1.34 \cdot 10^{-2}$ and $k_{\text{dup}} = 10^{-4}$, respectively. All rates have units of cell^{-1} generation^{-1}. In our setup, one-day timescale is equivalent to between 10 and 23 generations (lower and upper bound, respectively; the bounds are estimated from the minimum and maximum growth rate of the least and best adapted copy number types, Table 5.2, Fig 5.2d).

5.3 Discussion

Biology often relies on messy solutions, be it due to physical limitations or because evolution proceeds by opportunistic tinkering [Tawfik, 2010; Jacob, 1977]. For organisms living in constantly fluctuating environments even the crudest form of gene regulation [Troein *et al.*, 2007] or gene expression heterogeneity [Wolf *et al.*, 2015] increases fitness compared to not having any regulation at all. Here, we showed that the intrinsic instability of gene amplifications, rapidly tunes gene expression levels when gene regulation is required but no other molecular regulatory mechanism is in place.

Despite resembling canonical gene regulation when observing populations as a whole (Fig 5.2b), AMGET does not allow all single cells to change their gene expression concurrently. Instead, only a fraction of the population grows after the environment changes. Thus, AMGET may effectively work by allowing bacterial populations to 'hedge their bets' for expression levels that could be required in a future environment. Unlike traditional descriptions of bet-hedging, where genetically identical individuals show variability in their phenotypic states [Veening *et al.*, 2008], AMGET populations differ in their genotype due to the intrinsic instability of gene amplifications, thus passing on the adaptive state with high probability. Moreover, bet-hedging is typically characterized by switching between a small number of alternative phenotypic states [Veening *et al.*, 2008], while in an amplified locus, expression can adopt a graded response due to a wide range of copy numbers.

Because AMGET enables rapid dynamics and at the same time graded responses, it can be thought of as a form of primitive gene expression regulation at the population level [Anderson and Roth, 1977]. Mechanistically, AMGET bears no resemblance to canonical gene regulation, which employs sensory machinery to alter gene expression in the course of just a single generation. Yet, despite the mechanistic difference, AMGET operates on the time scales of days and thus closer to those of canonical gene regulation, compared to the process of transcriptional rewiring by point mutations, which occur several orders of magnitude less frequently.

AMGET may be one of several ways by which populations can make use of variation

in expression levels to rapidly adapt to environmental change. While point mutations occur at lower rates, regulatory rewiring can be surprisingly fast [Taylor *et al.*, 2015], especially when there is pre-existing variation in the precise architecture of regulatory networks. Moreover, noise propagation within gene regulatory networks can create an abundance of different expression levels, which are – in principle – tunable by selection [Wolf *et al.*, 2015]. However, as the results of our co-culture experiment (Fig 5.4) show, pre-existing variation can be easily depleted from a population if under strong selection. While it was previously shown that variation can be maintained in the form of multiple plasmid copies [Rodriguez-Beltran *et al.*, 2018], our results highlight that multiple copies of a genomic region actively regenerate heterogeneity due to the high recombination rate. Due to this property, AMGET provides a means of tuning expression to rare environmental fluctuations, where canonical gene regulation cannot evolve or be maintained [Gerland and Hwa, 2009]. AMGET is fast in bacteria because their generation times are short and their population sizes are usually large. However, our model results show that AMGET is in principle applicable to any other organism, but would take much longer time in relatively small populations (Fig 5.11f). A compelling example for the 'up-regulation' of a gene on relatively short evolutionary time-scales is that of the salivary amylase in humans, where variation in AMY1 copy number correlates with dietary starch content of human populations [Perry *et al.*, 2007].

Because any genomic region can be potentially amplified, AMGET can act on essentially any bacterial gene, providing regulation when the promoter is lacking altogether or when the existing promoter is not adequately regulated [Gil *et al.*, 2006; Latorre *et al.*, 2005]. For instance, horizontally transferred genes tend to be poorly regulated, as their integration into endogenous gene regulatory networks can take millions of years [Lercher and Pál, 2008; Pál *et al.*, 2005]. At the same time, they are enriched in mobile genetic elements [Dobrindt *et al.*, 2004; Juhas *et al.*, 2009], providing repetitive sequences for duplication by homologous recombination [Andersson and Hughes, 2009; Pettersson *et al.*, 2005]. Indeed, genes with a recent history of horizontal transfer are often amplified [Gusev *et al.*, 2014; Hooper and Berg, 2003; Eme *et al.*, 2017].

Similarly, gene amplifications can confer resistance to antibiotics and pesticides, but they are often accompanied by a fitness cost in the absence of the compound [Nguyen *et al.*, 1989]. In fact, heteroresistance caused by copy number polymorphisms is much more prevalent than previously thought and can lead to antibiotic treatment failure [Nicoloff *et al.*, 2019]. Repeated use of antibiotics or pesticides can therefore create alternating selection regimes [Gladman *et al.*, 2015], where AMGET might play an important, yet previously overlooked, role in bacterial adaptation.

In spite of their ubiquity, GDA has been underappreciated [Andersson and Hughes, 2009; Elliott *et al.*, 2013]. In principle, fixed amplifications can easily be detected in next generation sequence data by an increase in coverage and mismatches corresponding to the duplication junctions (Fig 5.7, Section 5.4). However, duplications revert to the single copy state at high rate without leaving any traces in the genome (Fig 5.7a). This implies that populations have to be kept under selection prior to sequencing, a condition that may not typically be met, especially not for environmental isolates [Eydallin *et al.*, 2014]. However, despite this challenge, there are many reports of cases where amplified genes have been detected in the sequences of environmental strains and were found associated with adaptation to environmental conditions [Gil *et al.*, 2006; Gusev *et al.*, 2014; Greenblum *et al.*, 2015].

The notion that GDA "might be thought of as a rather crude regulatory mechanism" [Anderson and Roth, 1977] is more than 40 years old. However, so far almost all experimental work has focused on the benefits of amplification in constant, stable environments, thereby selecting for increased expression only [Näsvall *et al.*, 2012; Dhar *et al.*, 2014]. Here, we demonstrated how flexible GDA is in rapidly altering gene expression levels of populations in response to a wide range of environmental fluctuations. AMGET is thus a critical, and a critically underappreciated, mechanism of bacterial survival.

	Regulation	**Amplification**	**Adaptation (rewiring via point mutations)**	**Bet-hedging strategies**
mechanism	hard-wired response of individual cells	mutation	mutation	phenotypic differences between genetically identical cells
rate ON	1	10^{-6} – 10^{-2} cell^{-1} gen.$^{-1}$ [Anderson and Roth, 1981; Reams *et al.*, 2010; Pettersson *et al.*, 2009; Sun *et al.*, 2012]	10^{-9} bp^{-1} cell^{-1} gen.$^{-1}$ [Drake, 1991; Elez *et al.*, 2010]	$> 10^{-5}$ variants per total cells [Bayliss, 2009]
rate OFF	1	$10^{-}3$ – 10^{-1} cell^{-1} gen.$^{-1}$ [Reams *et al.*, 2010; Pettersson *et al.*, 2009]	10^{-9} bp^{-1} cell^{-1} gen.$^{-1}$ [Drake, 1991; Elez *et al.*, 2010]	$> 10^{-5}$ variants per total cells [Bayliss, 2009]
active sensing machinery required	yes	no	no	no
can substitute for regulation on ecological time scales	-	yes	no	yes
expression state genetically heritable	no	yes	yes	no
tuning (allows graded expression)	typically not	yes	yes, but very long timescales	typically not
high reversibility (rate OFF > rate ON)	yes	yes	no	yes
suitable for rare stresses	no	yes	probably not, due to slow reversibility	depends on cost and rate

Table 5.1: **Comparison of regulation, amplification, adaptation and bet-hedging strategies.**

5.4 Methods

5.4.1 Bacterial strain background construction

Except when noted otherwise, all changes to the E.coli chromosome were introduced by pSIM6-mediated recombineering [Datta *et al.*, 2006]. All recombinants were selected on either 25μg/ml kanamycin or 10μg/ml chloramphenicol, to ensure single-copy integration. All resistance markers introduced by recombineering were flipped by transforming plasmid pCP20 and streaking transformants on LB at the non-permissive temperature of 37°C [Datsenko and Wanner, 2000]. We used strain MG1655 for all experiments, except for testing galactose and DOG concentrations (Fig 5.1c). For that purpose, we placed *galK* under control of the *pBAD* promoter and used strain BW27784, which allows relatively linear induction of the *pBAD* promoter over a 1000 fold range of arabinose concentration [Khlebnikov *et al.*, 2001]. In both strain backgrounds the genes *galK*, *mglBAC*, and *galP* were altered in order to allow galactose- and DOG-selection.
Endogenous galK was deleted by P1-transduction of *galK*::kan from the Keio-collection [Baba *et al.*, 2006]. The *mglBAC* operon was deleted to avoid selective import of galactose but not DOG [Nagelkerke and Postma, 1978]. To express *galP* for DOG to be imported in the absence of galactose, its endogenous promoter was replaced by constitutive promoter J23100 [Zhou *et al.*, 2017]. For this, the fragment texttttBBa_K292001 (available at the Registry of Biological Parts, http://parts.igem.org/Part:BBa_K292001) was cloned into pKD13 [Datsenko and Wanner, 2000] yielding plasmid pMS1 with FRT-kan-FRT upstream of J23100. The cassette FRT-kan-FRT-J23100 was used for recombineering.

5.4.2 Assembly of the chromosomal gene cassettes

The chromosomal reporter gene cassette used for experimental evolution (p_0-RBS-galK-RBS-yfp-p_R-cfp; Fig 5.1c) was assembled on plasmid pMS6* using standard cloning techniques. Plasmid pMS6* is based on plasmid pMS7, which contains the 'evo-cassette' (p_0-RBS-tetA-yfp-p_R-cfp) [Steinrueck and Guet, 2017]. To obtain pMS6*

we replaced the translational fusion of tetA-yfp on pMS7 with *galK* from MG1655 in a transcriptional fusion with *yfp venus*, originally derived from pZA21-yfp [Lutz and Bujard, 1997]. In addition, XmaI and XhoI restriction sites were added directly upstream and downstream of p_0 by two consecutive inverse PCRs.
The chromosomal gene cassette for testing galactose and DOG concentrations (*pBAD-galK*, Fig 5.1b) was assembled on plasmid pIT07, which was obtained by cloning *galK-yfp* as well as a chloramphenicol resistance flanked by FRT sites from pMS6* into pBAD24 [Guzman *et al.*, 1995]. Gene cassettes were integrated into chromosomal loci 1 and 2 (corresponding to locus D and E in Ref [Steinrueck and Guet, 2017]) by recombineering [Datta *et al.*, 2006] and checked by PCR with flanking primers and sequencing of the full-length construct.

5.4.3 Strain modification for microfluidics

The amplification of locus 1 was moved from the evolved strain IT028-EE1-D8 to the ancestral background (IT028) by P1 transduction to isolate it from the effect of other potential mutations in the evolved background, including a sticky phenotype, which clogged the microfluidic devices. In order to obtain a single copy control locus p_R-mCherry from our lab collection was introduced into the phage 21 attachment site (*attP21*) by P1-transduction [Bergmiller *et al.*, 2017].

5.4.4 *RecA* deletion in amplified strain locus 1 (Fig 5.8d,e)

RecA was deleted in the amplified strain by replacing it with the kanamycin cassette from pKD13 [Datsenko and Wanner, 2000]. In order to maintain the amplified state, recombinants were selected on M9 0.1% galactose medium supplemented with 25μg/ml kanamycin and verified by sequencing.

5.4.5 Culture conditions

All experiments were conducted in M9 medium supplemented with 2 mM $MgSO_4$, 0.1 mM $CaCl_2$ and different carbon sources (all Sigma-Aldrich, St. Louis, Missouri). For evolution experiments 0.1% galactose (high expression environment) or 1% glycerol

combined with 0.0001% 2-deoxy-d-galactose (DOG) (low expression environment), respectively, were added as carbon sources. For microfluidics experiments M9 medium was supplemented with 0.2% glucose and 1% casein hydrolysate and 0.01% Tween20 (Sigma-Aldrich, St. Louis, Missouri) was added as surfactant prior to filtering the medium (0.22 μm).
All bacterial cultures were grown at 37°C. Growth and fluorescence measurements in liquid cultures were performed in clear flat-bottom 96-well plates using a Biotek H1 platereader (Biotek, Vinooski, Vermont).

5.4.6 Mapping the relationship between galK expression level and growth growth

For the 2D gradients of arabinose and galactose or DOG (Fig 5.1b), respectively, an overnight culture of the test-cassette strain was diluted 1:200 into 96-well plates containing 200 μl of M9 supplemented with carbon sources, DOG and the inducer arabinose, as indicated in Fig 5.1b. Cultures were grown in the platereader with continuous orbital shaking.

5.4.7 Evolution experiments

For all evolution experiments (1. experimental evolution of the amplified strains in the high expression environment and 2. alternating selection experiments), cultures were grown in 200μl liquid medium in 96-well plates and shaken in a Titramax plateshaker (Heidolph, Schwabach, Germany, 750 rpm). Populations were transferred to fresh plates using a VP407 pinner (V&P SCIENTIFIC, INC., San Diego, California) resulting in a dilution of ~1:133.

1. Evolution of the amplified strains in the high expression environment

To obtain the amplified strains of locus 1 and 2, respectively, an overnight culture inoculated from a single colony of the ancestral strain carrying the reporter gene cassette in the respective loci (IT028; Fig 5.6b-c) or 2 (IT030; Fig 5.9b) was started in LB-medium. Cells were pelleted, washed twice and diluted 1:100 into M9 0.1%

galactose (locus 1) or M9 0.1% galactose supplemented with 0.1% casamino acids (locus 2). For locus 1, the timing of each dilution into fresh medium (~1:133) was chosen such as to maximize the number of rescued populations and to minimize the amount of time spent in stationary phase for grown populations. The transfers happened at days 10, 13, 15, 17, 18 and 19 (Fig 5.6c). The first signs of growth were detected in several wells only after approximately one week of cultivation in minimal galactose medium (Fig 5.6b). The evolving populations were monitored by spotting them onto MacConkey galactose agar in 128 x 86mm omnitray plates prior to transfer. For locus 2, the evolving populations were transferred daily (~1:133, corresponding to seven generations) and spotted on to LB plates supplemented with 0.5% charcoal (Fig 5.9b) to improve fluorescence quantification. Colony fluorescence of all experiments was recorded using a custom-made macroscope set-up (https://openwetware.org/wiki/Macroscope) [Chait *et al.*, 2010]. For the isolation of clones, evolved populations were streaked twice for purification on LB agar and grown in M9 galactose medium prior to freezing. For both locus 1 and 2, respectively, all further experiments were started from the original freezer stock of the amplified strain. This was done for two practical reasons: i) to save the time needed for duplications (and higher order amplifications) to evolve (one week in M9 galactose medium used for locus 1 and one day in M9 medium supplemented with casaminoacids used for locus 2), and more importantly, ii) to allow interpretation and reproducibility of the fluorescence data of the alternating selection experiments. As the reporter gene cassette allows selecting for increased *galK* expression but not for amplification itself, it is necessary to screen mutants with increased *galK* expression for increased CFP fluroescence. During amplification the initial duplication step is rate-limiting and break-points differ between evolving populations. We therefore limited ourselves to two amplified strains (locus 1 and 2), which we analyzed in detail. Amplified populations were thus started from single colonies, which were grown non-selectively on LB (Lennox) agar by streaking the original freezer stock. Due to the high rate of recombination, any given streak of the original amplified freezer stock contains colonies with a single copy of *galK* (Fig 5.3a, right panel). In order to pick only amplified colonies, we examined CFP fluorescence using the

macroscope.

We characterized evolved amplified strains by Sanger sequencing of the p_0 region, amplification junctions and the *rho* gene, which was found mutated in a previous study using the same locus [Steinrueck and Guet, 2017]. For the strain amplified in locus 1 (IT028-EE1-D8), increased *galK* expression is achieved by increased *galK* copy number as evident from increased CFP fluorescence (Fig 5.1c), as well as through a missense mutation in the termination factor *rho* (S265>A), allowing for baseline-expression via transcriptional read-through from the upstream *rsmG* into *galK* [Steinrueck and Guet, 2017]. The amplified region spans 16 kb from *atpB* at the left replicore over the origin of replication to *rbsD* into the right replicore.
For the strain amplified in locus 2 (IT030-EE11-D4), *galK* expression comes solely from the increase in copy number (no mutations in p_0 were detected). In this case, inverse PCR and sequencing confirmed that two identical IS elements (*IS1B* and *IS1C*) form the junction of the amplified segment [Steinrueck and Guet, 2017]. Whole genome sequencing of both amplified strains confirmed amplification junctions and the *rho* mutation detected with PCR and Sanger sequencing and revealed two additional single nucleotide changes in the amplified strain locus 1 (*coaA*, pos. 4174770, C>T, resulting in R>H; *wcaF*, pos. 2128737, C>A, resulting in G>V).

2. Alternating selection experiments

For the experiments in Fig 5.2b, a pre-culture of the amplified strain (IT028-EE1-D8) was grown in M9 0.1% galactose overnight, which was then inoculated 1:200 into the medium as indicated. For the experiment alternating two days in high and one day in low expression environment (Fig 5.2b – middle panel), populations were first subjected to a scheme of daily alternating selection for six days prior to switching to the 2-1 scheme.

For the co-culture experiments (Fig 5.4), a pre-culture of the amplified strain (IT028-EE1-D8) was grown in M9 0.1% galactose overnight. In parallel, the ancestral strain carrying a single silent copy of *galK* in locus 1 (IT028) and a strain constitutively

expressing *galK* in locus 1 (IT028-H5r), were grown overnight in M9 1% glycerol and mixed in a 1:1 ratio. We labeled the ancestral strain by transduction of *attP21*::p_R-mCherry (IT034). The constitutive strain was obtained by oligo-recombineering two point mutations into p_0 of the ancestral strain and selecting recombinants on M9 0.1% galactose agar. These two point mutations (-29 A>T and -37 G>T) have initially evolved in parallel to the amplified strain and result in a similar level of *galK* expression (Fig 5.1c). To quantify the relative abundance of the two strains in the co-culture, we calculated the expression ratio of the two strains, using an exchange rate between CFP and mCherry units from the ancestral strain expressing both fluorophores (IT034).

5.4.8 Whole genome sequencing

We isolated gDNA from overnight cultures of single clones of i) the ancestral strains ii) the amplified strains after initial selection in the high expression environment (galactose) as well as iii) the amplified strains after overnight selection in the low expression environment (DOG), for Locus 1 and Locus 2, respectively. In all cases overnight cultures were inoculated from colonies grown non-selectively on LB agar. For the overnight culture M9 1% glycerol was used for the ancestral and DOG-selected clones, while M9 0.1% galactose was used for the galactose-selected clones. A whole genome library was prepared and sequenced by Microsynth AG (Balgach, Switzerland) on an Illumina Next.Seq (with a mean read length of 75 bp). Fastq files were assembled to the MG1655 genome (Genbank accession number U00096.3) using the Geneious alignment algorithm with default options of the software Geneious Prime version 2019.2.1. SNPs were analyzed using the variant finding tool of Genious.

5.4.9 Flow Cytometry

Three colonies of the amplified strain and the constitutive control strain, respectively, were inoculated into culture tubes with 2ml M9 0.1% galactose (high expression environment) and grown for three days with transfers every 24h. This population

was inoculated into M9 + 1% glycerol + 0.0001% DOG (low expression environment). OD_{600} was monitored to assure continuous exponential growth by regular dilutions. Samples for flow cytometry were frozen at the indicated time points (Fig 5.2c). After 24h in the low expression environment, the populations were transferred back to the high expression environment with dilution and sampling occurring in the same manner. In parallel, the positive controls were grown for five days in both selection environments, respectively, with transfers occurring every 24h. Fluorescence was measured using a BD FACSCanto^{TM} II system (BD Biosciences, San Jose, CA) equipped with FACSDiva software. Fluorescence from the Pacific Blue channel (CFP) was collected through a 450/50nm band-pass filter using a 405nm laser. Fluorescence of the FITC channel (YFP) was collected through a 510/50 band-pass filter using a 488nm laser. The bacterial population was gated on the FSC and SSC signal resulting in approximately 6000 events analyzed per sample, out of 10,000 recorded events.

5.4.10 Microfluidics experiments

For the microfluidics experiments, a single colony of the amplified strain was picked and grown overnight in nonselective LB (Lennox) medium.
Microfluidics devices were prepared as described previously [Bergmiller *et al.*, 2017]. Briefly, devices had dimensions 23μm×1.3μm×1.3μm (l,w,h) for the growth channels with 5μm spacing along a trench for growth medium. Devices were fabricated by curing degassed polydimethylsiloxane (Sylgard 184, 1:10 catalyst:resin) inside epoxy replicate master molds produced from primary wafer-molded devices. Microscopy was performed on an inverted Nikon Ti-Eclipse microscope and with a previously described set-up [Bergmiller *et al.*, 2017]. Per experiment, multiple positions of a single mother machine were imaged using a 60× 1.4 NA oil immersion objective lens. To image constitutive mCherry, the green LED (549±15nm) was used at a light intensity of 670μW and an exposure time of 170-200ms. To image CFP, the cyan LED (475±28nm) at a light intensity of 270μW and an exposure time of 90-100ms was used.

Analysis of microfluidics data

The mother machine allowed tracing of mother cells for ~38 divisions, thereby following the fate of arising copy number mutations in the absence of selection. In three experiments, we analyzed 336, 369 and 384 mother cell lineages, respectively, equaling a total of approximately 40.000 cell divisions (with a division time of 23.6 ± 1.5 min as determined by counting septation lines in growth channel kymographs). Microfluidics data analysis was based on mother cell time traces (Fig 5.4c). To this end, we used Fiji/ImageJ to create kymographs, by laying a line through the middle of mother cells perpendicular to the growth channel using the built-in Multi-Kymograph tool with a pixel width of 9. Kymographs of CFP and mCherry were then analyzed using MATLAB.Âă

Determining what data to include

To minimize the influence of three unknown factors (maturation rate and bleaching of the two fluorophores, and the degree of bleedthrough between channels on the microfluidic chip), we were restrictive with the colonies we included.

1. We excluded all fluorescence changes that occurred when the cells were dying. Only colonies (mother cell lineages) that continuously grew until the end of the experiment were included. Specifically, the last 10 frames of mean mCherry fluorescence of mother cells needed to exceed the background threshold (68%, 76%, 82% of total colonies included, respectively, for the three experiments).

2. Some colonies exhibited a large variation in growth rate, due to temporary slowdown and/or filamentation. In the kymographs this was seen as a large variance in the constitutive mCherry channel. We excluded colonies with a variance >1.5 times the mCherry experiment-wide variance (thus including 96%, 96%, 96% of total colonies included for the three experiments, respectively).

3. In some cases there was significant bleedthrough between adjacent colonies. To avoid double counting transitions, the colony that was less bright was

removed from the data set if two adjacent colonies had a correlation of 0.6 or higher (99%, 98%, 98% of total colonies included, respectively, for the three experiments).

For the identified colonies the maximum fluorescence value per time point was extracted for both, mCherry and CFP channels. These were plotted against each other and a rectangular area, bounded by a manually selected max and min for each channel was chosen such as to include all but extreme outliers (Fig 5.10a). Accordingly, 99% of data points were included in all three experiments.

Normalization

To correct for slow temporal drift in the signal of CFP and mCherry, a time average over all colonies was taken and a 7th degree polynomial fitted. All time points were divided by the corresponding polynomial estimates.
Furthermore, mCherry fluorescence was flat-field corrected based on the expectation that mCherry is roughly constant across all colonies.ÂăTo do so, a line was fitted to the coordinate to get an estimate of the background of each location. The data was divided by the corresponding estimated value.

Probability density function

For the probability density function (PDF) in Fig 5.10b we normalized for differential growth rate by dividing the CFP fluorescence by the constitutively expressed mCherry fluorescence. To reduce noise, a median filter (MATLAB medfilt1) was applied to the ratio of CFP and mCherry over 20 data points. To get an estimate of the PDF of the CFP/mCherry single cell fluorescence, we used a kernel density estimation (KDE) (MATLAB function ksdensity). To estimate a proxy for copy numbers, we found points where the first and second derivative of the PDF is zero. These points were set as initial conditions for a pairwise fitting of peak mean and variance. All but the first and the last peak had two estimates for mean and variance. For the mean, the average of the two was taken and for the variance the smaller one was chosen. To assign boundaries for states, the estimated variance was halved.

For plotting, the height of each peak was set to match the peak height. No weight was fitted. The mean inter-peak distance for each PDF was used as a proxy of copy numbers for plotting in Fig 5.4c.

Estimation of nS2R2 for classification of single cell traces

We have classified the single cell traces using a normalized R squared, the proportion of variance explained, which we call nS2R2. In this adjustment, each element in both the residual and the total sum of squares is normalized by the predicted value:

$$\text{nS2R2} = 1 - S_{\text{res}}^{\text{norm}} / S_{\text{total}}^{\text{norm}} \tag{5.1}$$

where

$$S_{\text{res}}^{\text{norm}} = \sum_i (y_i - f_i)^2 / f_i^2, \tag{5.2}$$

$$S_{\text{res}}^{\text{norm}} = \sum_i (y_i - y_0)^2 / f_i^2, \tag{5.3}$$

where y_i, f_i, and y_0 represent measurements, fitted/predicted values, and mean of the measurements, respectively. This normalization takes into account that the intrinsic noise increases with expression and thus penalizes it less. Next, the algorithm fits one constant to the start and one constant to the end value of the CFP/mCherry trace, and reports this estimation parameter (nS2R2) based on which it classifies traces as shown in the pie charts of Fig 5.10c. Clear transitions exhibit an nS2R2 score of>0.5 and were verified by eye analyzing microfluidics movies in detail. The algorithm classifies no-events (”flat lines”) if the nS2R2 score lies between 0 and 0.5. Traces, which cannot be classified unambiguously neither as clear transition nor as a clear no-event, i.e. with nS2R2 below 0, are classified as ”complex traces”. This occurs if the start and end of CFP/mCherry trace values are similar but vary significantly in between.

5.4.11 Quantitative PCR

For qPCR, DNA was isolated using Wizard Genomic DNA purification kit (Promega, Madison, Wisconsin) from 50 μl of frozen samples from different time points

(1,4,9,10,11, gal 10, single copy control, DOG 8, DOG 10) of one flow cytometry experiment grown for 4-5 generations in LB. To quantify fluorescence, the same cultures were patched onto LB agar supplemented with 0.5% charcoal and imaged using the macroscope.

We performed qPCR using Promega qPCR 2x Mastermix (Promega, Madison, Wisconsin) and a C1000 instrument (Bio-Rad, Hercules, California). To quantify the copy number of samples of an evolving population, we designed one primer within *cfp* (target) and used one primer within *rbsB* as a close reference, which lies outside the amplified region. We compared the ratios of the target and the reference loci to the ratio of the same two loci in the single copy control. Using dilution series of one of the gDNA extracts as template, we calculated the efficiency of primer pairs to be 89.01% and 92.57%, for *cfp* and *rbsB*, respectively. We quantified the copy number ofÂă*cfp* in each sample employing the Pfaffl method, which takes amplification efficiency into account [Pfaffl, 2001]58. qPCR was done in three technical replicates.

5.4.12 Measurement of colony fluorescence (Fig 5.6c, Fig 5.9b, Fig 5.3a)

Colonies were grown without selection and imaged using the macroscope set up. To obtain mean colony CFP fluorescence intensity, a region of interest was determined using the ImageJ plugin 'Analyze Particles' (settings: 200px-infinity, 0.5-1.0 roundness) to identify colonies on 16-bit images with threshold adjusted according to the default value. The region of interest including all colonies was then used to measure intensity.

5.4.13 Mathematical model

A simple mathematical model recapitulates the change in *galK* copy number of the amplified population (Fig 5.5a). Importantly, the parameters for the model were estimated purely from calibration measurements (growth rates, fitness in the two environments with respect to copy number (flow cytometry experiments), number of generations spent in each environment, and recombination rate, k_{rec}) and the

literature (k_{dup}, [Andersson and Hughes, 2009]). Their values are listed in Table 5.2. No parameter was fit to reproduce the measurements in Fig 5.5a.

The model describes the time evolution of a population where cells with different gene copy numbers are represented by distinct states. The duplication and amplification events are the only source of transition between states. The time evolution proceeds iteratively, with discrete times representing synchronous cell divisions in the population.

The size of subpopulation N_j of cells with gene copy number j at time $t+1$ equals:

$$N_j(t+1) = \underbrace{\overbrace{(1-k_{\text{rec}}s_j)N_j(t)}^{\text{daughter 1}} + \overbrace{(1-k_{\text{rec}}-k_{\text{dup}}\delta_{j,1})s_jN_j(t)}^{\text{daughter 2}}}_{\text{no duplication or amplification event}} + \underbrace{\sum_{k=2}^{M} k_{\text{rec}}P_{kj}s_kN_k(t)}_{\text{amplification event}} + \underbrace{k_{\text{dup}}s_1N_1(t)\delta_{j,2}}_{\text{duplication event}} \tag{5.4}$$

where s_j is the relative growth rate of the subpopulation with j gene copies in the given environment (taken from Fig 5.2d), δ_{jk} a Kronecker delta which equals 1 if $j = k$ and 0 otherwise. The equation for single and double gene copy numbers ($j = 1$ or $j = 2$, respectively) has an additional term to reflect duplication events. As we assume that the rate of recombination per copy is constant, the overall recombination is proportional to the number of gene copies k; $k_{\text{rec}} = kk_{\text{rec}}^0$ (ref 8). P_{kj} represents the transition probabilities given an amplification event and is computed in the following way: assuming a homologous recombination between sister chromosomes occurs somewhere in the gene, we computed all possible combinations of how genes can be recombined to form different number of gene copies between the two daughter cells. P_{kj} then represents the probability that, given a recombination event, a daughter cell obtains j gene copies with its mother having k of them before the event. For example, starting with three gene copies, there is 22% probability to obtain four gene copies, or 22% probability to have one copy in the daughter (Fig 5.11h). We have observed in microfluidics experiments that most (65%) copy number changes happen only in the mother cell while the daughter cell remains unchanged. Therefore, we do not model recombination as a reciprocal event.

Based on platereader bulk experiments, observations indicated an upper limit for the copy number a cell can have. Thus, in our model, a cell can have up to M gene copies; if that number is exceeded, the cell stops dividing. This upper limit for gene copy number was confirmed in microfluidics and qPCR experiments, indicating to be between 6 and 12. Our single cell analysis showed that $M = 10$ is a good estimate (Fig 5.10b, according to number of states in the probability density function, see Analysis of the microfluidics data). However, the results of the mathematical model do not depend on the precise value within the measured range, as all results remain qualitatively the same for any value in the range of 6 and 12. Fig 5.11g shows that relative growth rates, obtained from flow cytometry experiments, are independent of M.

5.4.14 Measurements of model parameters (Table 5.2)

T1 & T2, generations per day in 96 **well plates**

In order to model the alternating selection experiment (Fig 5.5a), we needed to know the maximal growth rate of the amplified strain (IT028-EE1-D8) in the high and low expression environments, respectively. Because the exact details of cultivation (such as culture volume, shaking speed and temperature fluctuations) strongly affected growth rate, we were unable to measure growth curves while keeping cultures under the conditions of the original experiment. Hence, we estimated growth rate indirectly without perturbing the experiment, by determining the maximal number of generations possible in 24h (number of generations = 24[h] · growth rate[1/h]/log(2)) from a dilution series experiment. Populations pre-adapted to the respective environment were grown to carrying capacity of the respective medium and diluted by a factor of approximately 2^n (with n ranging between 7 and 28). We sought the maximal dilution that could still be compensated by growth (by requiring after 24h of growth the OD_{600} to reach the OD_{600} of the stationary phase). All dilutions of equal to or less than 1:222 and 1:223 were able to reach stationary phase in the high and low expression environment, respectively, yielding model parameters $T_1 = 22$ and $T_2 = 23$ for the maximal possible number of

generations.

T10 & T20, generations per day in culture tubes

Parameters T10 and T20 were necessary for obtaining the fitness landscape in Fig 5.2d (and the resulting relative growth rates s_j). T10 and T20 generations per day, measured under the exact conditions of the flow cytometry experiment (Fig 5.2c), namely exponential growth in culture tubes with 2ml volume of M9 0.1% galactose or M9 1% glycerol + 0.0001% DOG, respectively. We measured OD_{600} with a WPA Biowave spectrophotometer (Biochrom, UK).

Determining fitness landscape and relative growth rates s_j

The relative growth rates for each genotype (copy number state) in the high and low expression environments, respectively, were computed from flow cytometry time series experiments assuming exponential growth with no duplication/amplification event ($k_{\text{dup}} = 0$, $k_{\text{rec}} = 0$). This is a valid approximation as long as the two rates are small enough, such that the population structure consists of all copy number types, i.e., that each subpopulation is much larger than the additional cells created by a single amplification or duplication event. The flow cytometry measurements of the distribution of CFP expression at different times were split in M equal-width bins. The lowest and highest bins were set according to the equilibrium fluorescence distribution in DOG and galactose, respectively. For the lowest bin, we took the values of fluorescence < 85, while for the high bin we took the mode fluorescence values of the measured distributions, corresponding to > 160 for the first, and > 245 for the second set of flow cytometry experiments. Each bin represents a given gene copy number. The distributions between different times were then compared using iterative exponential growth model:

$$N_j(t_2) = (1 + s_j)^{(t_2 - t_1)/t_{1/2}} N_j(t_1), \tag{5.5}$$

where N_j is the population size with j gene copy number, $t_{1/2}$ is the doubling time, t_1 and t_2 are two measurement times, and s_j represents the relative growth of cells with j gene copies. The population distributions for all time points were

obtained from the flow cytometry data given the binning described above. Using this model, we obtained growth rates s_j for each pair of consecutive distributions at times t_i and $t_i + 1$ in the following way: given population distribution at time i, we predicted the new distribution given Eq. 5.5. We found such s_j values that minimize the Euclidean difference between the predicted and observed population distribution at time $i + 1$. We repeated this for all pairs of consecutive distributions (at times t_i and $t_i + 1$) and different replicates to obtain a set of solutions for s_j. Using this approach, we acquired only relative growth rates, which still allowed constants to be added to the growth rates. To tackle this, we added such constants to each growth rates in order to i) minimize the χ^2 of the differences between each growth rate solution and the mean of all solutions, which optimally removes the replicate-to-replicate variability (error bars in Fig 5.2D) on the inferred relative growth rates but does not affect their mean value; and ii) force the average growth rate of the adapted state to be 1 (i.e., for $j = 1$ in low expression environment and $j = M$ is high expression environment, $s_j = 1$) by adding a term to the χ^2 error function of the form (adapted state expression $- 1)^2$. Fixing s to be 1 in a reference environment is a convention that mathematically will not affect any subsequent results.

The absolute maximal growth rates in the two environments were measured in populations grown in high and low expression environments for 120h, respectively. Thus, they represent the growth rates of populations with the highest and lowest possible copy number (Fig 5.2c, positive controls). The estimated fitness values for both high expression environment (s_j^{HEE}) and low expression environment (s_j^{LEE}) can be found in Table 5.2.

Estimation of recombination rate $k_{\mathbf{rec}}$ from microfluidics data

We obtained a conservative estimate for the lower bound for the average number of copy number mutations from single step transitions in the pie charts (Fig 5.10c). Out of 72 mother cell time traces classified as clear transition events, we verified 67 by detailed analysis of microscopy images. We accordingly calculated the lower bound for the mutation rate as 67 events/1089 lineages/22.7 generations yielding

$k_{\text{rec}} = 2.7 \cdot 10^{-3}(\pm 0.74 \cdot 10^{-3})$ per cell per generation.

To estimate the mean recombination rate to be used in the model, two corrections have to be made: i) because our model assumes that the recombination rate is proportional to the number of gene copies 8, we had to take into account that cells with higher initial gene copy number are more likely to undergo a recombination event; and ii) as our experimental setup only allowed us to see if there has been a change in gene copy numbers or not, we had to take into account that there are some recombination events that do not change the gene copy number.

To account for i), we first computed the probability distribution that a given number of independent recombination events occur (Fig 5.10d): given the assumed independence of recombination events, the probability of observing a certain number of recombination events for a given cellular trace is approximately Poisson distributed, with the parameter being the expected number of events per microfluidic experiment duration (i.e., the effective recombination rate times the number of generations). The total number of observed generations was: 37.7, 36.3, and 41.3 for the three microfluidics experiments, respectively. Our approach is an approximation, namely it assumes a constant effective recombination rate for each trace throughout the experiment, which can be violated if more than one recombination event occurs. For example, the first recombination event can change the gene copy number, which in turn changes the probability of subsequent recombination events happening. While it is in principle possible to take this into account, it substantially complicates the inference of the recombination rate from data and makes it strongly model dependent.

As per our model assumption, the effective recombination rate is equal to the initial number of gene copies times the basal recombination rate. Therefore, we used all single cell traces to estimate a starting gene copy distribution. To do this, we averaged the normalized fluorescence (as a proxy for the starting effective gene copy number, see Fig 5.3c) over the time points 20 through 50. Next, we computed a Poisson probability distribution of obtaining k events ($k = 0, 1, \dots$) in the time of the experiment for each individual trace, with the basal recombination rate multiplied with the starting gene copy number (Fig 5.10d). For example, if a single cell trace

started with 4 gene copies, the expected number of events per experiment would be 4 times the basal recombination rate times the number of generations. Next, we averaged over all computed Poisson probability distributions, obtained from all single cell traces. This effectively means obtaining a total probability distribution for seeing 0, 1, or more recombination events over all recorded single-cell traces, taking into account point i).

Next, we consider point ii), taking into account the effect of recombination events that do not change the gene copy. We know from the P_{kj} matrix that the probability of keeping the gene copy numbers is the reciprocal of the initial gene copy number. Therefore, we took into account all events that would be seen as zero or single events (but are not) and adjusted the probability distributions. For this, we defined two probability distributions: the distribution of observed events, p_{observed}, which we are trying to find; and the distribution of âĂIJactualâĂİ number of events, p_{actual}, which we computed as described above. For example, in the observed distribution that is compared with experimental data, we classified as single events all double events where one of the recombination events leaves the copy number unchanged, all triple events where two events keep the copy numbers unchanged, etc. Therefore, the probability of observed events also includes the actual probability from states with $k > 0$ in which recombination did not change the copy number:

$$p_{\mathrm{observed}}(k=0) = p_{\mathrm{actual}}(k=0) + \sum_j p_{\mathrm{actual}}(j)/\epsilon_0^j, \quad \text{for all } j > 0, \tag{5.6}$$

with $p(j)$ being the probability of having j recombination events, and ϵ_0 being the initial gene copy number in the given single cell trace (estimated from experimental single cell traces). The $(1/\epsilon_0)^j$ represents the probability of having j consecutive recombination events, all of which leave the gene copy number unchanged. Analogously, the observed probability for a single event ($k = 1$) to occur is:

$$p_{\mathrm{observed}}(k=1) = p_{\mathrm{actual}}(k=1) + \sum_j (j-1) p_{\mathrm{actual}}(j)/\epsilon_0^{j-1}, \quad \text{for all } j > 1. \tag{5.7}$$

The prefactor $(j-1)$ comes from the number of different possibilities of having events that keep the gene copy number unchanged. For example, having 3 recombination

events, there are 3 different ways of having two events that keep the gene copy number unchanged while one event changes it.

After taking both corrections into account, we obtain a probability distribution of observing k recombination events (Fig 5.10d). The estimate of the basal recombination rate, k^0_{rec}, is based on the proportion of traces classified by our algorithm as no mutation events. We looked for such a recombination rate that best matched the number of no-events in the probability distribution (Fig 5.10c-d). We obtained $k^0_{\text{rec}} = 0.01434$ per cell per generation, which is approximately 5× larger than the conservative lower bound.

5.4.15 Model comparison with experimental data

For comparison of the model with the experimental data (Fig 5.5a), we simulated the full experimental protocol (for parameter values, see Table 5.2):

1. We exposed a single copy, ancestral population to a week of high expression environment, driving the population structure close to equilibrium. This mimicked the evolution of the amplified strain in the high expression environment such that both experimental and simulated population started with the same degree of copy number polymorphism.

2. The population spent one day in the low environment (for details on procedure in each day, see below).

3. For the experiment shown in Fig 5.5a top panel, the population was additionally exposed to three daily oscillations between high and low expression environment.

4. The population was exposed to the environments indicated in Fig 5.5a.

5. For every experiment, bacterial culture was diluted by a factor of $D = 133$ every day, thus limiting growth. This growth limitation was enforced by multiplying all growth rates by $g(c) = (1 - \min(c/133, 0))^0.01$, with c being the number of

cells, relative to the number of cells after each dilution. The exponent 0.01 was chosen such that $g(c)$ was smooth but nearly a step function.

6. To compare the units of experimental and simulated data, we obtained a common reference point. We took this to be the expression value after one week in the high expression environment, when the population has already equilibrated. We aligned these two points to have the same expression value. This value varies between different experiments.

The simulation of one day consisted of (for parameter values see Table 5.2):

1. Given the recombination rate and number of states M, we computed the transition matrix P_{kj} (see Eq 5.4) in the following way: given k copy numbers, the probability of going from k to $j < k$ copy numbers equals j/k^2, while probability for k to $j >= k$ equals $(2k - j)/k^2$ [Pettersson *et al.*, 2009]. Furthermore, we assumed that no transitions that increase copy numbers beyond M are allowed. We implemented this by setting all probabilities that go over M gene copies to zero.

2. Next, to update the current population structure following Eq 5.4, we used the current population structure, N_j, selection on the states (growth rates) in the given environment, s_j (Fig 5.2d), transition matrix, P_{kj} (probability of having j copies given k copies), the duplication and recombination rate (k_{dup} and k_{rec}, respectively), and the dilution factor D. First, we computed the total population growth since the last dilution, i.e., the ratio of population size of current time point and the size after last dilution. Second, we computed $g(c)$ (taking into account the saturation of the population) and multiplied it with each of the selection values s_j in Eq 5.4. Then, we used these new values to compute N_j at the new time point.

3. We repeated the step 2 for 23 or 22 times for low or high expression environment, respectively. These numbers represent the number of cell divisions per day and were determined experimentally. Steps $2 - 3$ represent time evolution of the population over the period of one day.

4. We diluted the population by a factor of $D = 133$.

5. We repeated the steps $2 - 4$ according to the environment the population is exposed at on the new day (selection different between the two environments). With this step, we simulate different days, diluting after each (step 4).

6. For each time point, we computed expression as the average gene copy number: $E = \sum_j w_j$, where w_j is the the proportion of cells with j gene copies and sum goes over all gene copy numbers.

7. At the end, we returned the population distribution and expression at each time point.

For simulation of the stochastic environmental durations, we followed the same procedure as for the deterministic ones, except that the environment durations here were randomly drawn from an exponential distribution.

5.4.16 Finite size population model

To compute the response times for a finite size population (Fig 5.11f), we used the Wright-Fisher model where the population size is kept constant. The procedure was:

1. Given all parameters of the system and using the infinite size population model (Eq 5.4), we obtained the equilibrium distribution of the population in the starting environment. We computed the equilibrium distribution of copy numbers in the infinite population size limit by computing the eigenvector corresponding to the largest eigenvalue of the transition matrix (obtained from r.h.s. of Eq 5.4), and obtained the starting finite population as a multinomial draw of N individuals from this equilibrium distribution.

2. After the environmental transition, we updated the distribution after each division. The new distribution was computed using the Eq 5.4.

3. We computed the new population, as a multinomial draw of N individuals, randomly drawn from the new population distribution.

4. After each division, we computed the expression of the population.

5. We repeated steps $3-5$ until response $R = M/2$ has been reached. The number of generations until this point represents the time to response. We define response as the ratio of mean copy numbers before and after the environmental switch.

Fig 5.11f shows the response time as the average over 100 replicate simulations of the algorithm above.

5.4.17 Quantification and Statistical Analysis

Statistical details of individual experiments, including number of replicate experiments, mean values, and standard deviations, are described in the figure legends and indicated in the figures. For the t-test in Fig 5.4c-d we computed the response as the fold change between mean expression of days $1-15$ in the high expression environment and mean expression in the low expression environment on day 16 for amplified populations (Fig 5.4c). For the co-culture populations (Fig 5.4d), we analogously computed the response as fold change between mean constitutive strain abundance of days $1-15$ in the high expression environment and mean constitutive strain abundance in the low expression environment on day 16. We used a two-sided t-test (Matlab function ttest2) to compute the p-value ($2.6 \cdot 10^{-68}$) for the difference in mean response between amplified (Fig 5.4c) and co-culture populations (Fig 5.4d). For measuring the linear dependence between the experimental data and model prediction in Fig 5.5a, we computed the Pearson correlation coefficient using the inbuilt Matlab function corrcoef.

5.5 Supporting Information

5.5.1 Supplementary Note. An upper limit for copy number exists in locus 1.

CFP levels of the amplified strain stabilize in populations following prolonged exposure to the high expression environment (Fig 5.7c, positive control), indicating that there is a cost to increasing copy number above a certain point. Indeed, microfluidics experiments revealed that increasing copy number beyond the maximum attainable level of CFP fluorescence generally led to cell death, which lead to the exclusion of all such lineages from our analysis (Section 5.4). Both, microfluidics experiments (Fig 5.10b) and qPCR (Fig 5.6a), consistently estimate a maximum copy number between six and ten. This upper limit to copy number might be due to the fact that the origin of replication lies within the amplified segment (see Section 5.4) and could thus be specific to the strain we are using. This is corroborated by the fact that the copy number of the strain amplified in locus 2 is estimated to be 39 according to read-depth (Fig 5.7b). If there is a strict limit to copy number in locus 2, it is much higher than in locus 1.

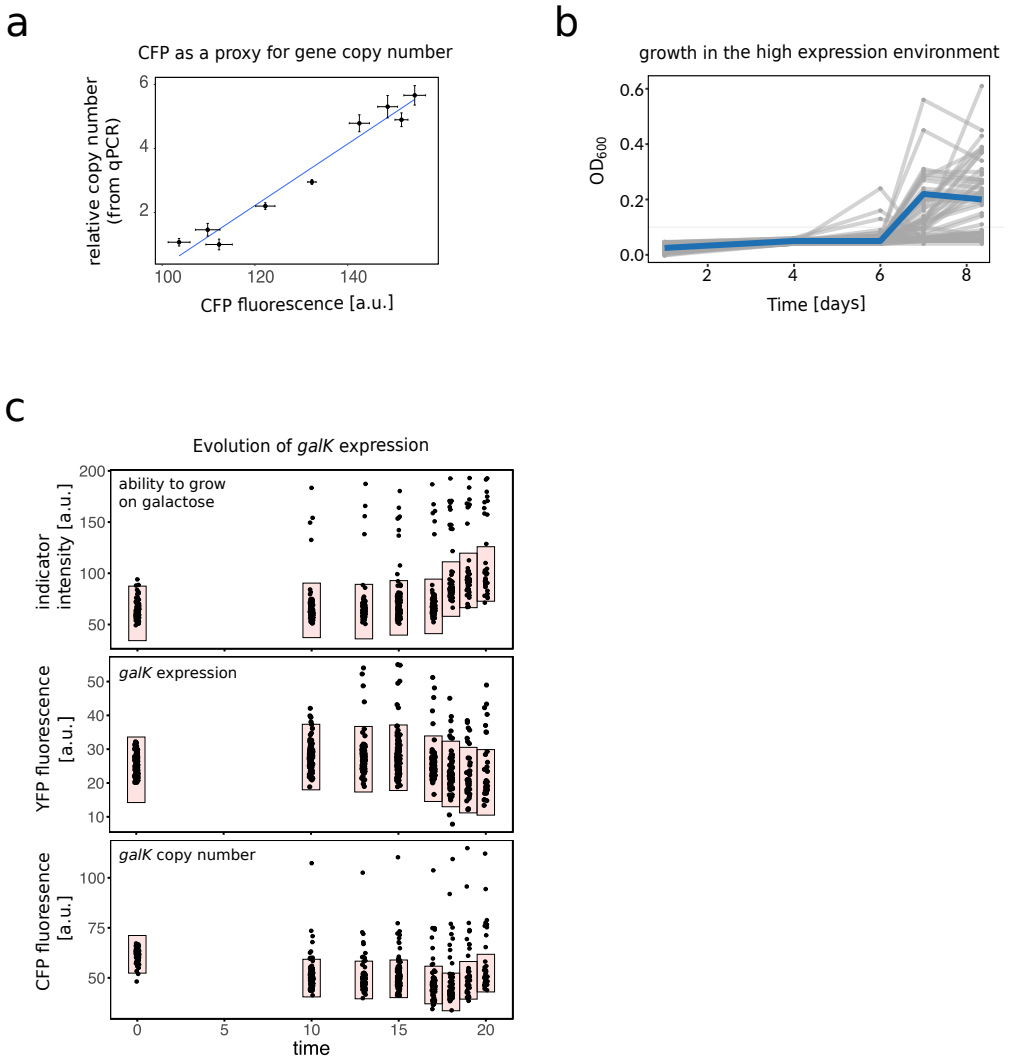

Figure 5.6: **Experimental evolution of galK expression. (a)**, CFP fluorescence of bacterial colonies as a proxy for copy number. Copy number relative to a single copy control strain as determined by qPCR is plotted for eight populations with varying levels of CFP fluorescence. Error bars represent the standard deviation of three and four replicates for copy number and CFP fluorescence, respectively. Linear fit: Adjusted $R^2 = 0.9558$, p-value=3.3510^{-6}. **(b)**, OD_{600} of 95 replicate populations of the ancestral strain each evolving in 200μl minimal galactose medium (high expression environment). Plot shows the initial continuous cultivation phase of the evolution experiment prior to the first transfer to fresh medium. Blue line shows the population of the amplified strain. **(c)**, MacConkey agar pins (as shown in Fig 5.1b, – right part) of the 95 replicate populations shown in b during 21 days of evolution in the high expression environment. Evolving populations were pinned onto MacConkey agar at the beginning of the evolution experiment and prior to each transfer into fresh medium to monitor their phenotypic changes: ability to grow on galactose (apparent from pH indicator color shift to pink) - top panel, colony YFP fluorescence (as a proxy for *galK* expression) - middle panel and colony CFP fluorescence (as a proxy for *galK* copy number) - bottom panel. Area shaded in red corresponds to population median $\pm 3\sigma$ of the ancestral population.

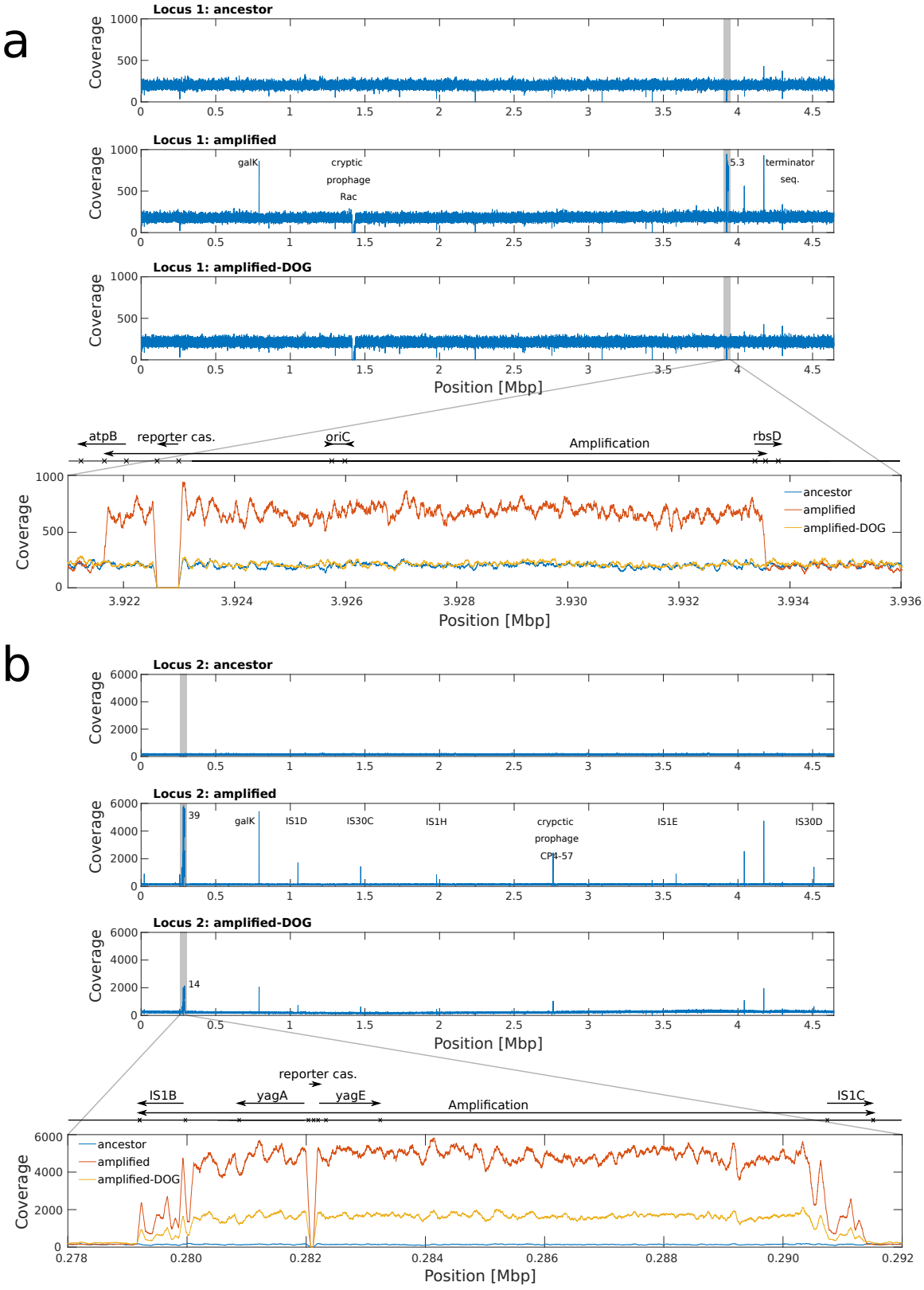

Figure 5.7: (Continued on the following page.)

Figure 5.7: **Coverage plot of ancestral and evolved strains of locus 1 and locus 2.** Read-depth is shown for the whole genome of **(a)**, Locus 1: (**top**) ancestral strain, (**middle**) amplified strain isolated after evolution in the high expression environment (Fig 5.6c), (**bottom**) amplified strain after 24h in the low expression environment (clone from experiment shown in Fig 5.7c). **(b)**, Locus 2 (**top**) ancestral strain, (**middle**) amplified strain isolated after evolution in the high expression environment (Fig 5.9b), (**bottom**) amplified strain after 24h in the low expression environment. The number next to the amplified region indicates the fold change in coverage as compared to the respective ancestral strain. Additional regions with increased coverage (labeled in the middle panels of a) are caused by sequence reads of the synthetic reporter cassette mapping to homologous sequences within the E.coli genome: endogenous *galK*, terminators downstream of *yfp* and *cfp* (4.1 and 4.2 Mbp, resp.). Prophage Rac is absent in the evolved strains of locus 1.ÂăFor locus 2, additional regions of increased coverage (labeled in the middle panel of b) are caused by homologies with the amplified region, especially insertion sequence (IS) element 1.

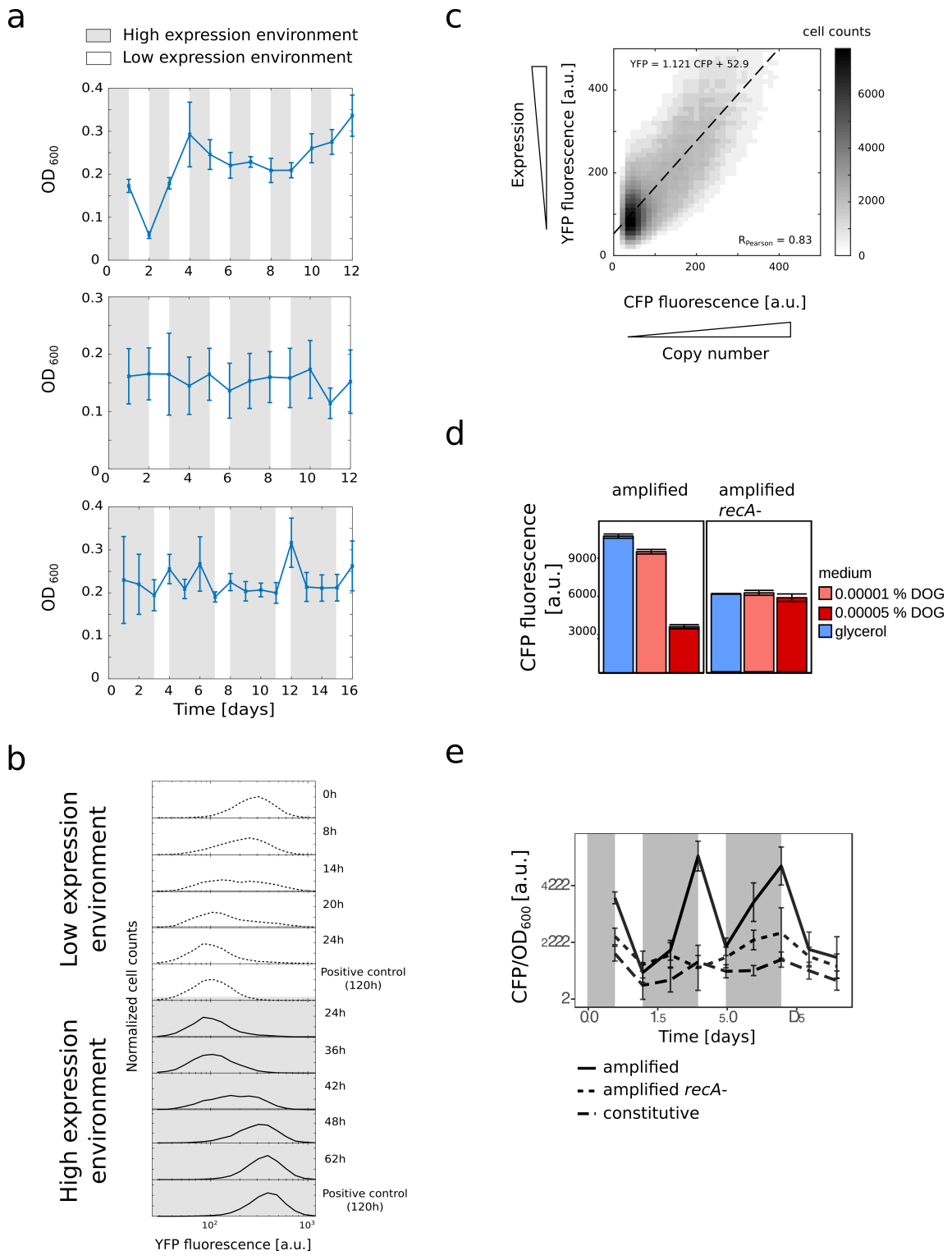

Figure 5.8: (Continued on the following page.)

Figure 5.8: **Amplification-mediated gene expression tuning (AMGET) allows growth in alternating environments and is dependent on recA. (a)**, Growth of the amplified strain during alternating selection (see also Fig 5.2b). OD_{600} is shown for alternating selection following the scheme of 1 day - 1 day, 2 days - 1 day and 3 days - 1 day in high and low expression environment, respectively. Error bars represent the standard deviation (SD) of 60 populations. **(b)**, Flow cytometry histogram (one of six replicates from two independent experiments) following the adaptation of an amplified bacterial population to low and high expression environments. Population was inoculated from a single colony and selected for two days in the high expression environment prior to the two transitions shown here. When switched from high to low expression environment, YFP fluorescence as a proxy for *galK* expression is decreasing within 24h to reach the steady state level of the same population after 5 days in the low environment (positive control). When shifted back to the high expression environment, the amplified population increases in CFP fluorescence to the level reached by the same population after 5 days in the high expression environment (positive control). **(c)**, Plot shows CFP fluorescence as a proxy for galK copy number and YFP fluorescence as a proxy for *galK* expression of the evolving population (data from the experiment shown Fig 5.2c and Fig 5.8b, respectively). **(d)**, Mean steady state CFP fluorescence of amplified populations with (left) and without (right) functional *recA* allele grown in 0%, $5 \cdot 10^{-5}$% and 10^{-5}% DOG. **(e)**, During alternating selection, CFP levels of the amplified strain tracks fluctuating environments. CFP levels of neither the *recA*-derivative of the amplified strain nor a constitutive, single-copy derivative of the amplified strain follows the environments. The constitutive strain evolved serendipitously in an overnight culture as a clone that lost its amplification but gained a point mutation in p_0 of the chromosomal cassette allowing for *galK* expression in the absence of amplification.

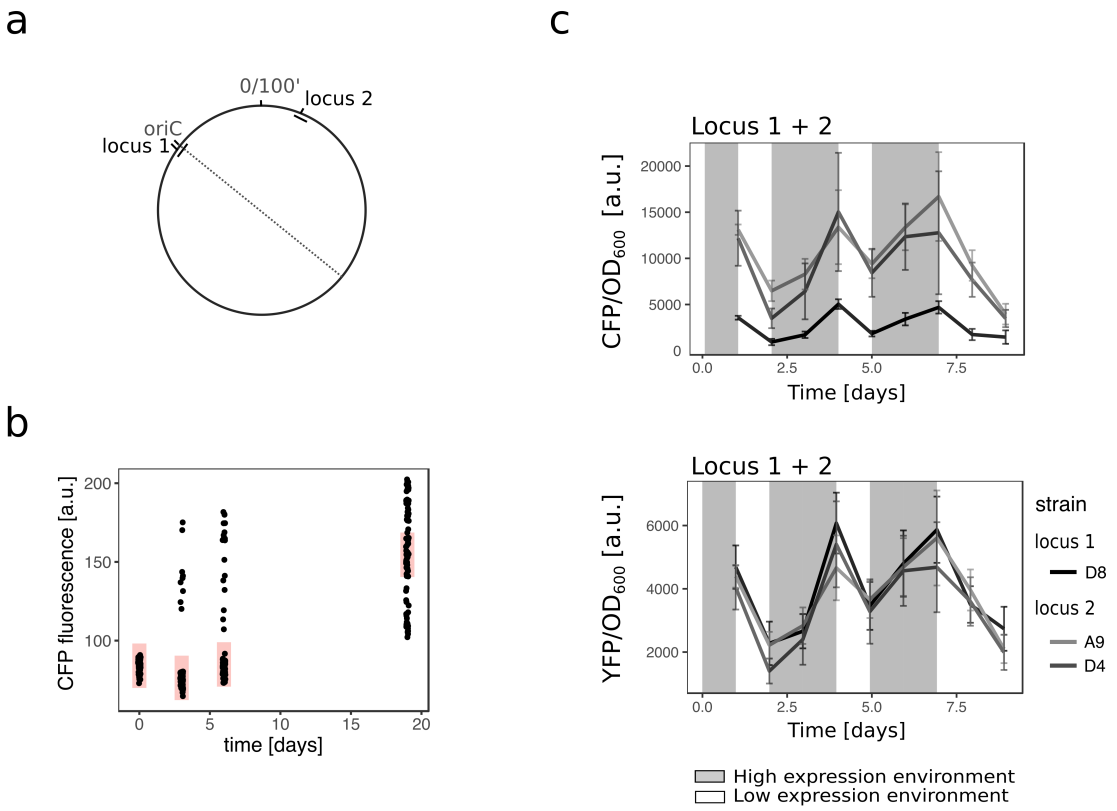

Figure 5.9: **AMGET occurs at a different genomic locus. (a)**, *E. coli* chromosome map showing positions of locus 1 (downstream of *rsmG*) and locus 2 (inside cryptic prophage CP4-6 and flanked by two identical IS elements) relative to the origin of replication (*oriC*). **(b)**, Amplifications readily evolve in locus 2. Colony CFP fluorescence as a proxy for gene copy number of 95 replicate populations pinned onto agar before and during evolution in the high expression environment. Red shaded area represents the median $\pm 3\sigma$ of the ancestral population. **(c)**, Normalized CFP fluorescence of strains with gene amplification in locus 2 (”A9”, ”D4”) tracks fluctuating environments like the strain with a gene amplification in locus 1 (”D8”). Although absolute CFP levels are higher in locus 2 than locus 1 (top panel), fold change of CFP and YFP is similar between both loci (bottom panel).

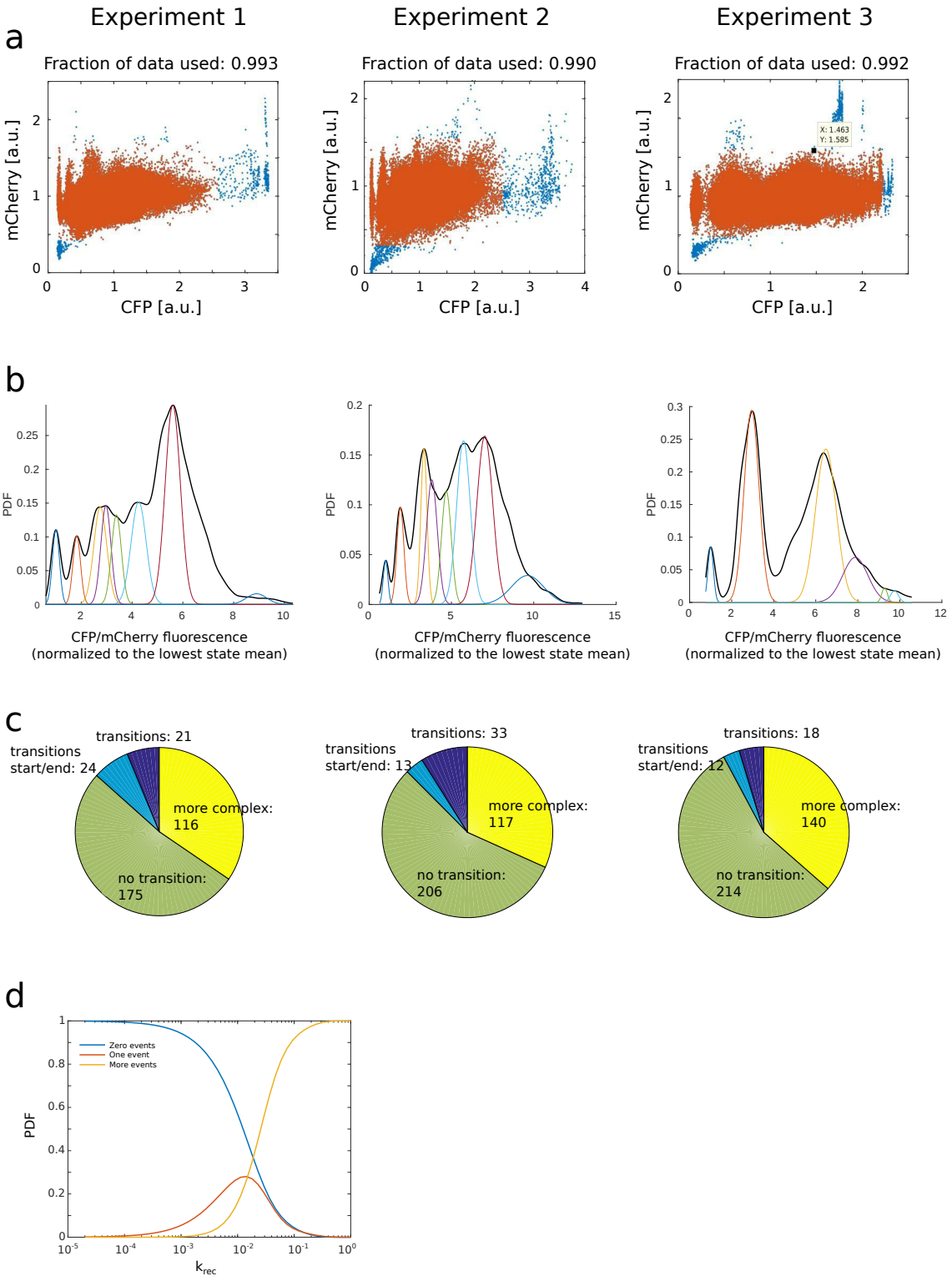

Figure 5.10: (Continued on the following page.)

Figure 5.10: **Microfluidics data analysis. (a)**, Scatter of fluorescence in mother cell for constitutive mCherry and copy number marker CFP in three replicate experiments. Orange data points are included in the further analysis, whereas blue points were manually excluded (for further details, see Section 5.4). **(b)**, Probability density function of orange data points in a are shown in black. Colored lines represent gene copy number estimates that were calculated using a Gaussian mixture model (for further details, see Section 5.4). **(c)**, The time series of the amplification marker fluorescence (growth normalized) for each mother cell was automatically classified into four categories. **Green** - no transition. **Light blue** – transition, but the transition was too close to the start or end of the experiment in order to determine if it was transient or not. **Dark blue** – transition considered to be stable. This number was verified by inspecting microfluidics movies and used to calculate the lower bound of the recombination rate. **Yellow** âĂŞ more complex behavior, multiple fast transitions, oscillations. **(d)**, Probability distribution of observing zero, one, or more independent recombination events, which lead to a change in copy number (see Section 5.4).

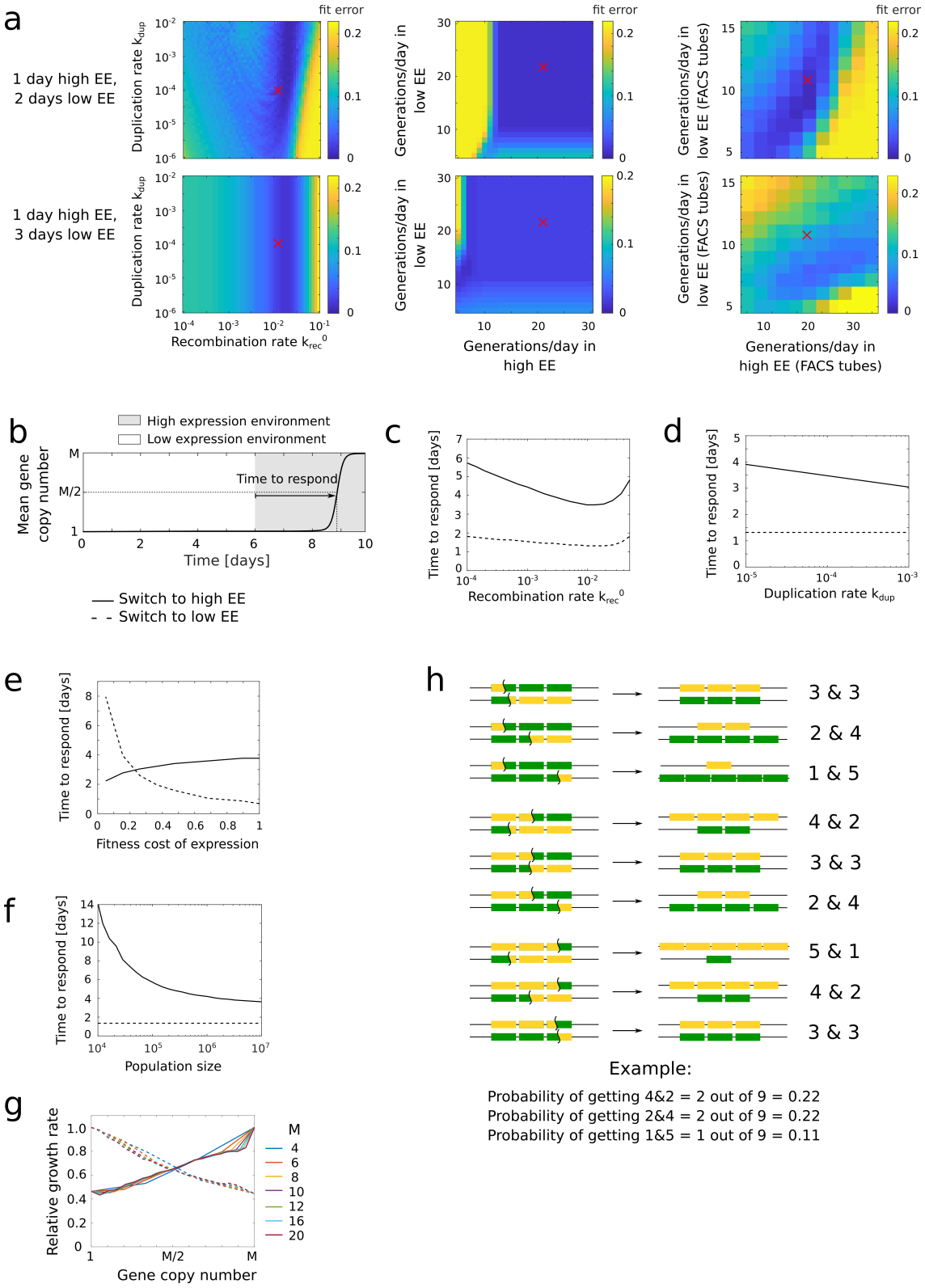

Figure 5.11: (Continued on the following page.)

Figure 5.11: **Mathematical model is not very sensitive to experimentally measured parameters. (a)**, Error of fitting for varying different parameters: gene amplification and duplication rate (left); growth rates, shown as generations per day, in high- and low expression environment (high EE, low EE) (middle), and in FACS tubes (right). When a set of two parameters is varied, all other parameters remain fixed. Error of fit of two different experiments is shown in top and bottom. The error of fitting is defined as the average squared difference between experimental and simulated data point. Values that we measured in independent experiments and are used in our simulations are marked by a red x. **(b)**, An example of a rare environment where the preceding environment is long enough such that the gene copy number distribution does not change. The time to response is defined as the time needed by the population after environmental switch to achieve response R=M/2=5. **(c-f)**, Time to respond as a function of amplification- c, and duplication d, rate, fitness costs of expression e, and population size f, for either switching from low to high expression environment (full line), or from high to low expression environment (dashed line). **(g)**, Relative growth rate for different choices of maximum number of gene copies, M, for low expression environment (dashed lines), and high expression environment (full line). **(h)**, An example of all combinations of two sister chromosomes undergoing homologous recombination and splitting six gene copies among themselves. In all plots, unless stated otherwise, we use relative growth rates as shown in Fig 5.2d, and amplification and duplication rates of $k^0_{\text{rec}} = 1.34 \cdot 10^{-2}$ and $k_{\text{dup}} = 10^{-4}$ per cell per generation, respectively.

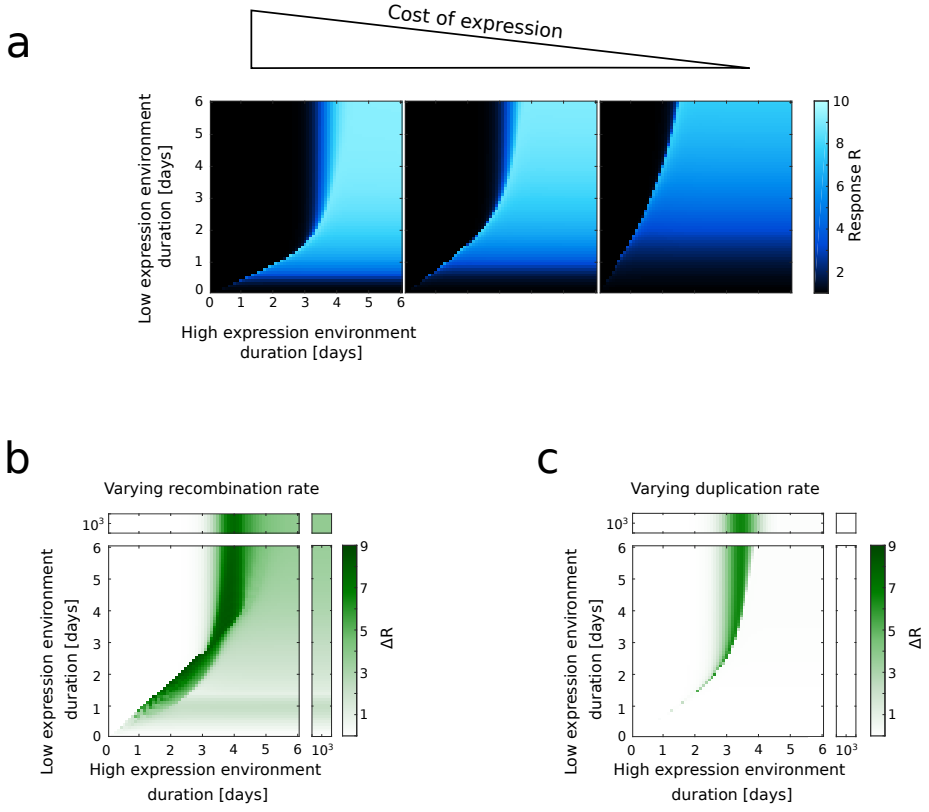

Figure 5.12: (Continued on the following page.)

Figure 5.12: **Robustness of AMGET with respect to varying model parameters. (a)**, The response R as the function of the two environment durations for three different expression costs (from left to right: 0.8, 0.5, 0.2). With decreasing cost of expression, AMGET effectively slows down and increases the environmental durations required to observe a visible response increase. This behavior leads to a predictable outcome in the limit of vanishing expression cost, where the population remains in the high expression state forever and thus no regulation via AMGET is needed. **(b-c)**, Population response generated by AMGET is robust to large variations in the recombination and duplication rate. Maximal variation in response (color scale), defined as ΔR=max(R)-min(R), for varying recombination rate b, and duplication rate c, for a set of environmental durations. We densely sample the parameter ranges for basal recombination rate, k^0_{rec} (see Section 5.4), in the range of $10^{-4} - 5 \cdot 10^{-2}$ and duplication rate in the range of $10^{-5} - 10^{-3}$, and find the largest and smallest response within this range to compute ΔR. The recombination rate mostly affects R around the narrow range of environment durations near the switch from no response to full response. For shorter environment durations, amplifications do not have enough time to sweep through a population and hence no response is achieved for any realistic recombination rate. Conversely, for longer durations enough time has passed in each environment that a response will always be maximal, except for recombination rates above 10-2, which dampen the response as mutation decreases the efficacy of selection. The duplication rate only affects the response for environmental durations close to the switch from no response to full response, and for low expression environments of a long duration. This is because the emergence of new duplications becomes rate-limiting after the low expression environment switches back to the high expression environment. In all plots, unless stated otherwise, we use recombination and duplication rates $k^0_{\text{rec}} = 1.34 \cdot 10^{-2}$ and $k_{\text{dup}} = 10^{-4}$, respectively. All rates have units per cell per generation. In our setup, one-day timescale is equivalent to between 10 and 23 generations (lower and upper bound, respectively; the bounds are estimated from the minimum and maximum growth rate of the least and best adapted copy number types, Table 5.2, Fig 5.2d).

Parameter values	**Symbol**	**Value**	**Obtained from**
Max number of copies	M	10	qPCR & microfluidics (Section 5.4)
LEE FACS tubes gen. per day	T20	10.4	growth rate in culture tube
HEE FACS tubes gen. per day	T20	14.7	growth rate in culture tube
LEE gen. per day	T2	23.0	dilution series in 96-well plates
HEE gen. per day	T1	22.0	dilution series in 96-well plates
Recombination rate (cell^{-1} gen.$^{-1}$)	k^0_{rec}	0.0134	microfluidics (Section 5.4)
Duplication rate (cell^{-1} gen.$^{-1}$)	k_{dup}	10^{-4}	[Anderson and Roth, 1981; Reams *et al.*, 2010; Pettersson *et al.*, 2009; Sun *et al.*, 2012]
Relative growth rates in HEE	s_1^{HEE}	0.46	Flow cytometry experiment
	s_2^{HEE}	0.45	(fitness landscape; Fig 5.2d)
	s_3^{HEE}	0.51	in combination with growth
	s_4^{HEE}	0.57	rate measurement of
	s_5^{HEE}	0.62	the fittest copy number (s_{10}^{HEE})
	s_6^{HEE}	0.68	(Section 5.4)
	s_7^{HEE}	0.74	
	s_8^{HEE}	0.78	
	s_9^{HEE}	0.81	
	s_{10}^{HEE}	1	
Relative growth rates in LEE	s_1^{LEE}	1	Flow cytometry experiment
	s_2^{LEE}	0.94	(fitness landscape; Fig 5.2d)
	s_3^{LEE}	0.84	in combination with growth
	s_4^{LEE}	0.74	rate measurement of
	s_5^{LEE}	0.67	the fittest copy number (s_{10}^{HEE})
	s_6^{LEE}	0.62	(Section 5.4)
	s_7^{LEE}	0.57	
	s_8^{LEE}	0.53	
	s_9^{LEE}	0.50	
	s_{10}^{LEE}	0.44	

Table 5.2: **Model parameter values.** LEE – low expression environment, HEE – high expression environment. Reported values represent the mean of triplicate experiments.

Name	**Purpose**	**Source**
pZA21-yfp	source for yfp in pMS6*	[Lutz and Bujard, 1997]
pKD13	kan template for recombineering	[Datsenko and Wanner, 2000]
pMS7	starting point for construction of the gene cassette for evolution, pir dependent replication	[Steinrueck and Guet, 2017]
pMS6*	gene cassette template for recombineering, pir dependent replication, pir dependent replication	this study
pMS1	template for constitutive galP promoter (fragment J23100) based on pKD13	lab collection
pBAD24	basis for pIT07	[Guzman *et al.*, 1995]
pIT07	gene cassette template for recombineering, based on pBAD24	this study

Table 5.3: **Plasmids.**

Strain name	Genotype	Purpose	Source
MG1655	F- λ- ilvG- rfb-50 rph-1	strain background for all experiments except testing experiment	lab collection
BW27784	lacIq rrnB3 ΔlacZ4787 hsdR514 DE(araBAD)567 DE(rhaBAD)568 DE(araFGH) Φ(ΔaraEp PCP4AâĂŞaraE)	background for testing experiment (Fig 5.1b) strain construction	[Khlebnikov *et al.*, 2001]
IT013	BW27784, JA23100::galP, mglBAC::FRT, galK::FRT	background for testing experiment (Fig 5.1b) strain construction	this study
IT013-TCD	BW27784, JA23100::galP, mglBAC::FRT, galK::FRT, locus1::pBAD-galK	strain background for testing experiment (Fig 5.1b)	this study
MS022	MG1655, JA23100::galP, mglBAC::FRT, galK::FRT	background for evolution experiment strain construction	lab collection
JW0740-3	F-, Δ(araD-araB)567, ΔlacZ4787(::rrnB-3), λ-, ΔtolC732::kan, rph-1, Δ(rhaD-rhaB)568, hsdR514 ΔgalK729::kan	source for galK deletion	[Baba *et al.*, 2006]
BW25142	lacIq rrnB3 (lacZ4787 hsdR514 DE(araBAD)567 DE(rhaBAD)568 (phoBR580 rph-1 galU95 (endA9 uidA((MluI)::pir-116 recA1	host for pir plasmids pMS6* and pMS7	[Haldimann and Wanner, 2001]
IT028	MS022 locus1::p0-RBS-galK-RBS-yfp-FRT-pR-cfp	ancestor strain for evolution experiment (Fig 5.6b,c)	this study
IT030	MS022 locus2::p0-RBS-galK-RBS-yfp-FRT-pR-cfp	ancestor strain for evolution experiment (Fig 5.9b)	this study
IT028-EE1-D8	IT028 dup(atpB-rsbD), rho (S265>A)	amplified strain locus 1, evolved in evolution experiment (Fig 5.6b,c)	this study
IT028-EE1-D8-recA	IT028 dup(atpB-rsbD), rho (S265>A), ΔrecA	ΔrecA-stabilzed version of amplified strain IT028-EE1-D8 (Fig 5.8d,e)	this study
IT028-EE11-D4	IT030 dup(IS1B-IS1C)	amplified strain locus 2, evolved in evolution experiment (Fig 5.9b)	this study
IT028-EE1-D8-pRmCherry	MS022 dup(locus1::p0-RBS-galK-RBS-yfp-FRT-pR-cfp), attP21::pR-mCherry	amplified strain locus 1, for microfluidics (Fig 5.3b,c)	this study
IT034	IT028 attP21::pR-mCherry	ancestral strain in co-culture experiments (Fig 5.4b,d)	this study
IT028-H5r	MS022 locus1::pconst-RBS-galK-RBS-yfp-FRT-pR-cfp	constitutive strain in co-culture experiments (Fig 5.4b,d)	this study

Table 5.4: **Bacterial strains.**

6 Conclusion

The thesis addressed four questions relating to gene regulation across different scales and their evolutionary consequences. We started with a broader, systems-level problem of crosstalk and how crosstalk influences the type of regulation. We have shown that global crosstalk, which takes into account the whole network, could play a role in determining one form of regulation above the other.
Next, we focused on eukaryotic gene regulation, in particular, on the architecture of eukaryotic enhancers. We showed that the normative approach (i.e., an approach that postulates the optimal biological function) can be successful in controlling the complexity of eukaryotic models. We demonstrated how a simple generalization of equilibrium models allows us to escape equilibrium bounds and access optimal regulatory phenotypes, while remaining consistent with the reported phenomenology and simple enough to be inferred from upcoming experiments.
In Chapter 4 we focused on prokaryotic gene regulation by addressing one of the central questions of evolutionary biology: that of genotype-phenotype mapping. With our model, which accurately predicted the genotype-phenotype map, we explored the constraints and mechanisms of promoter function on this map. This is arguably the first exploration of evolutionary consequences for promoter evolution in a model that is biophysically realistic enough to fit dynamic gene expression data. Furthermore, we have also shown that underlying mechanisms – and not the whole detailed GP mapping – determine general trends in promoter evolution.
We concluded with an evolutionary aspect of gene regulation. In Chapter 5, we demonstrated under what conditions intrinsic instability of gene duplication and amplification provides a generic alternative to canonical gene regulation. Using modeling we showed that this alternative can work in a wide range of environments,

including those where transcription factor-based schemes are hard to evolve or maintain.

These four questions addressed gene expression and regulation across different time scales. First two chapters described processes that are related to organisms responding to environmental cues. These are biological processes, such as production and binding of transcription factors, which ensure the survival of the organism. These occur on short time scales. The opposite, the evolutionary time scales, are required to evolve new regulatory binding sites. We have shown in Chapter 4 that even though both regulation and expression need evolutionary time scales to evolve, the regulation requires much longer to evolve than expression. Therefore, what can a cell with constitutive expression but no available gene regulation do when this regulation is required? If selection is strong, such organisms wouldn't survive unless an alternative to regulation is found. We explored this in the last chapter where we showed that amplification is a viable alternative to (evolving) gene regulation. Therefore, we concluded the thesis with this mechanism which bridges the necessity of fast response to environmental cues together with long evolutionary times required to evolve the machinery to do that.

As we have seen in Chapter 4, thermodynamic-based models of gene regulation are interpretable in mechanistic terms and have high predictive power, making them perfectly suited as components in models of promoter and network evolution. What could be the future directions that such models explore, and how could we address them with our existing framework?
Theoretical studies of evolution often focus on point mutations and, at best, insertions and deletions, whereas experimental results point to a much broader repertoire of mutational moves. There is evidence that regulatory evolution, in practice, proceeds mainly by these alternative moves [Steinrueck and Guet, 2017]. This agrees with what we have shown in Chapter 5: that duplication and amplifications play an important role in gene regulation. Therefore, we could extend our biophysical model to include these types of mutations, thus allowing amplification of individual

binding sites. We could investigate if such mechanisms, which do not leave any genomic signature, could be a factor in significantly influencing the TF-BS evolution. Given their high rates, this idea appears plausible.

Next, understanding how other types of gene regulation change the biophysical constraints could prove important. Therefore, by studying regulation by activation, and how it influences evolutionary trajectories, we could gain new insights into evolutionary preferences of difference regulatory networks. Two extreme scenarios can be imagined. First, the main conclusions of biophysical constraints would not be changed when looking at different type of gene regulation, showing that main properties of such constraints do not originate from network properties. In the second scenario, most of the biophysical constraints would be strongly influenced by the type of regulation, showing a strong need to further understand how network characterstics affect the biophysical properties. Reality most likely lies somewhere between the two extremes. Furthermore, in such study any combination of regulators could be included, addressing questions of biophysically realistic evolution of whole regulatory networks.

When studying transcription factor binding site evolution, a common simplification is that transcription factor properties are fixed and are not changing. Knowing that transcription factor representation is captured by an energy matrix, we can study how evolution of transcription factor binding sites is affected by changing transcription factors (i.e., energy matrices). This would allow us to investigate, for example, the relationship between changing transcription factors and their binding sites, answering under which conditions binding sites can or cannot evolve. In other words, we could ask how fast and in what way can transcription factors change such that evolution of binding sites can follow these changes? Overall, connecting the mutational effects in the coding region of a protein to the *correct representation* of its binding affinity poses a crucial step in correctly understanding transcription factor binding site evolution.

All of the above extensions represent different facets of the same general problem: namely, of using biophysical realism to construct, simulate, and understand the evolution of genetic regulatory networks. This approach would offer the ability to

remain firmly grounded to known molecular mechanisms and study their individual influences, while simultaneously allowing us to consider systems- and network-level effects, such as crosstalk.

This thesis represents over four years of work at IST Austria. Of the four projects described, two of them started already during my rotations which then transformed into full projects. The other two projects are a result of an interdisciplinary and encouraging environment – an environment which shaped me not only as a scientist but also as a person. To conclude, I believe that I took advantage of the interdisciplinary environment at IST which, I hope, is seen at the broadness of my projects and this thesis.

Bibliography

[Aakre *et al.*, 2015] C. D. Aakre, J. Herrou, T. N. Phung, B. S. Perchuk, S. Crosson, and M. T. Laub, "Evolving New Protein-Protein Interaction Specificity through Promiscuous Intermediates," *Cell*, 163(3):594–606, 2015.

[Adam *et al.*, 2015] R. C. Adam, H. Yang, S. Rockowitz, S. B. Larsen, M. Nikolova, D. S. Oristian, L. Polak, M. Kadaja, A. Asare, D. Zheng, and E. Fuchs, "Pioneer factors govern super-enhancer dynamics in stem cell plasticity and lineage choice.," *Nature*, 521(7552):366–370, 2015.

[Afek *et al.*, 2014] A. Afek, J. L. Schipper, J. Horton, R. Gordân, and D. B. Lukatsky, "Protein-DNA binding in the absence of specific base-pair recognition," *Proceedings of the National Academy of Sciences*, page 201410569, 2014.

[Aguilar-Rodríguez *et al.*, 2017] J. Aguilar-Rodríguez, J. L. Payne, and A. Wagner, "A thousand empirical adaptive landscapes and their navigability," *Nature Ecology & Evolution*, 1(2), 2017.

[Aguilar-Rodríguez *et al.*, 2018] J. Aguilar-Rodríguez, L. Peel, M. Stella, A. Wagner, and J. L. Payne, "The architecture of an empirical genotype-phenotype map," *Evolution*, 2018.

[Alberch, 1991] P. Alberch, "From Genes to Phenotypes – Dynamical Systems and Evolvability," *Genetica*, 84(1):5–11, 1991.

[Albertson, 2006] D. G. Albertson, "Gene amplification in cancer," *Trends in Genetics*, 22(8):447–455, 2006.

[Alon, 2007] U. Alon, "Network motifs: theory and experimental approaches," *Nature Reviews Genetics*, 8(6):450–461, 2007.

[Anderson and Roth, 1981] P. Anderson and J. Roth, "Spontaneous tandem genetic duplications in Salmonella typhimurium arise by unequal recombination between rRNA (rrn) cistrons.," *Proceedings of the National Academy of Sciences*, 78(5):3113–3117, 1981.

[Anderson and Roth, 1977] R. P. Anderson and J. R. Roth, "Tandem Genetic Duplications in Phage and Bacteria," *Annual Review of Microbiology*, 31(1):473–505, 1977.

[Andersson and Hughes, 2009] D. I. Andersson and D. Hughes, "Gene Amplification and Adaptive Evolution in Bacteria," *Annual Review of Genetics*, 43(1):167–195, 2009.

[Arbeitman, 2002] M. N. Arbeitman, "Gene Expression During the Life Cycle of Drosophila melanogaster," *Science*, 297(5590):2270–2275, 2002.

[Arnosti and Kulkarni, 2005] D. N. Arnosti and M. M. Kulkarni, "Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?," *Journal of cellular biochemistry*, 94(5):890–898, 2005.

[Baba *et al.*, 2006] T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori, "Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection," *Molecular Systems Biology*, 2:2006.0008, 2006.

[Babu and Teichmann, 2003] M. M. Babu and S. A. Teichmann, "Evolution of transcription factors and the gene regulatory network in Escherichia coli.," *Nucleic Acids Res*, 31(4):1234–1244, 2003.

[Bar-Joseph *et al.*, 2012] Z. Bar-Joseph, A. Gitter, and I. Simon, "Studying and modelling dynamic biological processes using time-series gene expression data," *Nature Reviews Genetics*, 13(8):552–564, 2012.

[Barkan *et al.*, 2011] D. Barkan, C. L. Stallings, and M. S. Glickman, "An improved counterselectable marker system for mycobacterial recombination using galK and 2-deoxy-galactose," *Gene*, 470(1-2):31–36, 2011.

[Barnes *et al.*, 2019] S. L. Barnes, N. M. Belliveau, W. T. Ireland, J. B. Kinney, and R. Phillips, "Mapping DNA sequence to transcription factor binding energy in vivo," *PLOS Computational Biology*, 15(2):e1006226, 2019.

[Bartman *et al.*, 2016] C. Bartman, S. Hsu, C.-S. Hsiung, A. Raj, and G. Blobel, "Enhancer Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping," *Molecular Cell*, 62(2):237–247, 2016.

[Bass and Field, 2011] C. Bass and L. M. Field, "Gene amplification and insecticide resistance.," *Pest management science*, 67(8):886–90, 2011.

[Bataillon and Bailey, 2014] T. Bataillon and S. F. Bailey, "Effects of new mutations on fitness: insights from models and data: Effects of new mutations on fitness," *Annals of the New York Academy of Sciences*, 1320(1):76–92, 2014.

[Bayliss, 2009] C. D. Bayliss, "Determinants of phase variation rate and the fitness implications of differing rates for bacterial pathogens and commensals," *FEMS Microbiology Reviews*, 33(3):504–520, 2009.

[Belliveau *et al.*, 2018] N. M. Belliveau, J. B. Kinney, and R. Phillips, "Systematic approach for dissecting the moleclar mechanisms of transcriptional regulation in bacteria," *PNAS*, page 10, 2018.

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

[Berg *et al.*, 2004] J. Berg, S. Willmann, and M. Lässig, "Adaptive evolution of transcription factor binding sites.," *BMC evolutionary biology*, 4:42, 2004.

[Berg and von Hippel, 1987] O. G. Berg and P. H. von Hippel, "Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters," *Journal of Molecular Biology*, 193(4):723–743, 1987.

[Bergmiller *et al.*, 2017] T. Bergmiller, A. M. Andersson, K. Tomasek, E. Balleza, D. J. Kiviet, R. Hauschild, G. Tkačik, and C. C. Guet, "Biased partitioning of the multidrug efflux pump AcrAB-TolC underlies long-lived phenotypic heterogeneity," *Science*, 356(6335):311–315, 2017.

[Bialek *et al.*, 2019] W. Bialek, T. Gregor, and G. Tkačik, "Action at a distance in transcriptional regulation," *arXiv preprint arXiv:1912.08579*, 2019.

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

[Biggin, 2011] M. D. Biggin, "Animal transcription networks as highly connected, quantitative continua.," *Developmental cell*, 21(4):611–626, 2011.

[Bintu *et al.*, 2005a] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips, "Transcriptional regulation by the numbers: applications," *Current Opinion in Genetics & Development*, 15(2):125–135, 2005.

[Bintu *et al.*, 2005b] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips, "Transcriptional regulation by the numbers: models," *Current Opinion in Genetics & Development*, 15(2):116–124, 2005.

[Carter *et al.*, 2013] H. Carter, M. Hofree, and T. Ideker, "Genotype to phenotype via network analysis," *Current Opinion in Genetics & Development*, 23(6):611–621, 2013.

[Carter *et al.*, 2005] M. G. Carter, A. A. Sharov, V. VanBuren, D. B. Dudekula, C. E. Carmack, C. Nelson, and M. S. Ko, "Transcript copy number estimation using a mouse whole-genome oligonucleotide microarray," *Genome Biology*, 6:R61, 2005.

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

[Chait *et al.*, 2010] R. Chait, S. Shrestha, A. K. Shah, J. B. Michel, and R. Kishony, "A differential drug screen for compounds that select against antibiotic resistance," *PLoS ONE*, 5(12):e15179, 2010.

[Changeux, 2012] J.-P. Changeux, "Allostery and the Monod-Wyman-Changeux model after 50 years," *Annual review of biophysics*, 41:103–133, 2012.

[Chen *et al.*, 2018] H. Chen, M. Levo, L. Barinov, M. Fujioka, J. B. Jaynes, and T. Gregor, "Dynamic interplay between enhancer–promoter topology and gene activity," *Nature genetics*, 50(9):1296–1303, 2018.

[Chen *et al.*, 2014] J. Chen, Z. Zhang, L. Li, B.-C. Chen, A. Revyakin, B. Hajj, W. Legant, M. Dahan, T. Lionnet, E. Betzig, R. Tjian, and Z. Liu, "Single-Molecule Dynamics of Enhanceosome Assembly in Embryonic Stem Cells," *Cell*, 156(6):1274–1285, 2014.

[Cho *et al.*, 2018] W.-K. Cho, J.-H. Spille, M. Hecht, C. Lee, C. Li, V. Grube, and I. I. Cisse, "Mediator and RNA polymerase II clusters associate in transcription-dependent condensates.," *Science*, 361(6400):412–415, 2018.

[Ciliberti *et al.*, 2007] S. Ciliberti, O. C. Martin, and A. Wagner, "Innovation and robustness in complex regulatory gene networks," *Proceedings of the National Academy of Sciences*, 104(34):13591–13596, 2007.

[Coulon *et al.*, 2013] A. Coulon, C. C. Chow, R. H. Singer, and D. R. Larson, "Eukaryotic transcriptional dynamics: from single molecules to cell populations," *Nature Reviews Genetics*, 14(8):572–584, 2013.

[Datsenko and Wanner, 2000] K. A. Datsenko and B. L. Wanner, "One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products," *Proceedings of the National Academy of Sciences*, 97(12):6640–6645, 2000.

[Datta *et al.*, 2006] S. Datta, N. Costantino, and D. L. Court, "A set of recombineering plasmids for gram-negative bacteria," *Gene*, 379:109–115, 2006.

[de Visser and Krug, 2014] J. A. G. M. de Visser and J. Krug, "Empirical fitness landscapes and the predictability of evolution," *Nature Reviews Genetics*, 15(7):480–490, 2014.

[Dean and Thornton, 2007] A. M. Dean and J. W. Thornton, "Mechanistic approaches to the study of evolution: the functional synthesis," *Nature Reviews Genetics*, 8(9):675–688, 2007.

[Dekel and Alon, 2005] E. Dekel and U. Alon, "Optimality and evolutionary tuning of the expression level of a protein," *Nature*, 436(7050):588–592, 2005.

[Dhar *et al.*, 2014] R. Dhar, T. Bergmiller, and A. Wagner, "Increased gene dosage plays a predominant role in the initial stages of evolution of duplicate TEM-1 beta lactamase genes," *Evolution*, 68(6):1775–1791, 2014.

[Dobrindt *et al.*, 2004] U. Dobrindt, B. Hochhut, U. Hentschel, and J. Hacker, "Genomic islands in pathogenic and environmental microorganisms.," *Nature reviews. Microbiology*, 2(5):414–24, 2004.

[Donovan *et al.*, 2019] B. T. Donovan, A. Huynh, D. A. Ball, H. P. Patel, M. G. Poirier, D. R. Larson, M. L. Ferguson, and T. L. Lenstra, "Live-cell imaging reveals the interplay between transcription factors, nucleosomes, and bursting," *The EMBO Journal*, 38(12):e100809–18, 2019.

[Drake, 1991] J. W. Drake, "A constant rate of spontaneous mutation in DNA-based microbes.," *Proceedings of the National Academy of Sciences*, 88(16):7160–7164, 1991.

[Dubnau and Losick, 2006] D. Dubnau and R. Losick, "Bistability in bacteria," *Molecular Microbiology*, 61(3):564–572, 2006.

[Durão *et al.*, 2018] P. Durão, R. Balbontín, and I. Gordo, "Evolutionary Mechanisms Shaping the Maintenance of Antibiotic Resistance," *Trends in Microbiology*, 26(8):677–691, 2018.

[Duveau *et al.*, 2017] F. Duveau, W. Toubiana, and P. J. Wittkopp, "Fitness Effects

of Cis-Regulatory Variants in the Saccharomyces cerevisiae TDH3 Promoter," *Molecular Biology and Evolution*, 34(11):2908–2912, 2017.

[Ehrensberger *et al.*, 2013] A. H. Ehrensberger, G. P. Kelly, and J. Q. Svejstrup, "Mechanistic interpretation of promoter-proximal peaks and RNAPII density maps.," *Cell*, 154(4):713–715, 2013.

[Elde *et al.*, 2012] N. C. Elde, S. J. Child, M. T. Eickbush, J. O. Kitzman, K. S. Rogers, J. Shendure, A. P. Geballe, and H. S. Malik, "Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses.," *Cell*, 150(4):831–41, 2012.

[Elez *et al.*, 2010] M. Elez, A. W. Murray, L.-J. Bi, X.-E. Zhang, I. Matic, and M. Radman, "Seeing Mutations in Living Cells," *Current Biology*, 20(16):1432–1437, 2010.

[Elliott *et al.*, 2013] K. T. Elliott, L. E. Cuff, and E. L. Neidle, "Copy number change: evolving views on gene amplification," *Future Microbiology*, 8(7):887–899, 2013.

[Elowitz and Leibler, 2000] M. B. Elowitz and S. Leibler, "A synthetic oscillatory network of transcriptional regulators.," *Nature*, 403(6767):335–8, 2000.

[Eme *et al.*, 2017] L. Eme, E. Gentekaki, B. Curtis, J. M. Archibald, and A. J. Roger, "Lateral Gene Transfer in the Adaptation of the Anaerobic Parasite Blastocystis to the Gut," *Current Biology*, 27(6):807–820, 2017.

[Estrada *et al.*, 2016] J. Estrada, F. Wong, A. DePace, and J. Gunawardena, "Information Integration and Energy Expenditure in Gene Regulation," *Cell*, 166(1):234–244, 2016.

[Eydallin *et al.*, 2014] G. Eydallin, B. Ryall, R. Maharjan, and T. Ferenci, "The nature of laboratory domestication changes in freshly isolated Escherichia coli strains," *Environmental Microbiology*, 16(3):813–828, 2014.

[Eyre-Walker and Keightley, 2007] A. Eyre-Walker and P. D. Keightley, "The distribution of fitness effects of new mutations," *Nature Reviews Genetics*, 8(8):610–618, 2007.

[Fontana and Buss, 1994] W. Fontana and L. W. Buss, "What would be conserved if "the tape were played twice"?," *Proceedings of the National Academy of Sciences*, 91(2):757–761, 1994.

[Forcier *et al.*, 2018] T. L. Forcier, A. Ayaz, M. S. Gill, D. Jones, R. Phillips, and J. B. Kinney, "Measuring cis-regulatory energetics in living cells using allelic manifolds," *Elife*, 7:e40618, 2018.

[Fordyce *et al.*, 2010] P. M. Fordyce, D. Gerber, D. Tran, J. Zheng, H. Li, J. L. DeRisi, and S. R. Quake, "*De novo* identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis," *Nature Biotechnology*, 28(9):970–975, 2010.

[Friedlander *et al.*, 2013] T. Friedlander, A. E. Mayo, T. Tlusty, and U. Alon, "Mutation Rules and the Evolution of Sparseness and Modularity in Biological Systems," *PLoS ONE*, 8(8), 2013.

[Friedlander *et al.*, 2017] T. Friedlander, R. Prizak, N. H. Barton, and G. Tkačik, "Evolution of new regulatory functions on biophysically realistic fitness landscapes," *Nature Communications*, 8(1):216, 2017.

[Friedlander *et al.*, 2016] T. Friedlander, R. Prizak, C. C. Guet, N. H. Barton, and G. Tkačik, "Intrinsic limits to gene regulation by global crosstalk," *Nature Communications*, 7:12307, 2016.

[Garcia *et al.*, 2012] H. G. Garcia, A. Sanchez, J. Q. Boedicker, M. Osborne, J. Gelles, J. Kondev, and R. Phillips, "Operator sequence alters gene expression independently of transcription factor occupancy in bacteria," *Cell reports*, 2(1):150–161, 2012.

[Gasch *et al.*, 2000] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown, "Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes," *Molecular Biology of the Cell*, 11(12):4241–4257, 2000.

[Gebhardt *et al.*, 2013] J. C. M. Gebhardt, D. M. Suter, R. Roy, Z. W. Zhao, A. R. Chapman, S. Basu, T. Maniatis, and X. S. Xie, "Single-molecule imaging of transcription factor binding to DNA in live mammalian cells," *Nature Methods*, 10(5):421–426, 2013.

[Gerland and Hwa, 2009] U. Gerland and T. Hwa, "Evolutionary selection between alternative modes of gene regulation," *Proceedings of the National Academy of Sciences*, 106(22):8841–8846, 2009.

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

[Gertz *et al.*, 2009] J. Gertz, E. D. Siggia, and B. A. Cohen, "Analysis of Combinatorial cis-Regulation in Synthetic and Genomic Promoters," *Nature*, 457(7226):215–218, 2009.

[Ghaemmaghami *et al.*, 2003] S. Ghaemmaghami, W.-K. Huh, K. Bower, R. W. Howson, and others, "Global analysis of protein expression in yeast," *Nature*, 425(6959):737, 2003.

[Gil *et al.*, 2006] R. Gil, B. Sabater-Muñoz, V. Perez-Brocal, F. J. Silva, and A. Latorre, "Plasmids in the aphid endosymbiont Buchnera aphidicola with the smallest genomes. A puzzling evolutionary story," *Gene*, 370:17–25, 2006.

[Gillespie, 2007] D. T. Gillespie, "Stochastic simulation of chemical kinetics.," *Annual Review of Physical Chemistry*, 58:35–55, 2007.

[Gladman *et al.*, 2015] S. L. Gladman, T. Seemann, W. Gao, T. P. Stinear, B. P. Howden, I. R. Monk, and N. J. Tobias, "Large tandem chromosome expansions facilitate niche adaptation during persistent infection with drug-resistant Staphylococcus aureus," *Microbial Genomics*, 1(2):e000026, 2015.

[Golding *et al.*, 2005] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox, "Real-Time Kinetics of Gene Activity in Individual Bacteria," *Cell*, 123(6):1025–1036, 2005.

[Greenblum *et al.*, 2015] S. Greenblum, R. Carr, and E. Borenstein, "Extensive Strain-Level Copy-Number Variation across Human Gut Microbiome Species," *Cell*, 160(4):583–594, 2015.

[Gusev *et al.*, 2014] O. Gusev, Y. Suetsugu, R. Cornette, T. Kawashima, M. D. Logacheva, A. S. Kondrashov, A. a. Penin, R. Hatanaka, S. Kikuta, S. Shimura, H. Kanamori, Y. Katayose, T. Matsumoto, E. Shagimardanova, D. Alexeev, V. Govorun, J. Wisecaver, A. Mikheyev, R. Koyanagi, M. Fujie, T. Nishiyama, S. Shigenobu, T. F. Shibata, V. Golygina, M. Hasebe, T. Okuda, N. Satoh, and T. Kikawada, "Comparative genome sequencing reveals genomic signature of extreme desiccation tolerance in the anhydrobiotic midge.," *Nature communications*, 5:4784, 2014.

[Guzman *et al.*, 1995] L. M. Guzman, D. Belin, M. J. Carson, and J. Beckwith, "Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter.," *Journal of bacteriology*, 177(14):4121–30, 1995.

[Hahn *et al.*, 2003] M. W. Hahn, J. E. Stajich, and G. A. Wray, "The Effects of Selection Against Spurious Transcription Factor Binding Sites," *Molecular Biology and Evolution*, 20(6):901–906, 2003.

[Haldane *et al.*, 2014] A. Haldane, M. Manhart, and A. V. Morozov, "Biophysical Fitness Landscapes for Transcription Factor Binding Sites," *PLOS Computational Biology*, 10(7):e1003683, 2014.

[Haldimann and Wanner, 2001] A. Haldimann and B. L. Wanner, "Conditional-replication, integration, excision, and retrieval plasmid-host systems for gene structure-function studies of bacteria.," *Journal of bacteriology*, 183(21):6384–93, 2001.

[Hammar *et al.*, 2014] P. Hammar, M. Walldén, D. Fange, F. Persson, Ö. Baltekin, G. Ullman, P. Leroy, and J. Elf, "Direct measurement of transcription factor dissociation excludes a simple operator occupancy model for gene regulation," *Nature genetics*, 46(4):405, 2014.

[Hansen, 2006] T. F. Hansen, "The Evolution of Genetic Architecture," *Annual Review of Ecology, Evolution, and Systematics*, 37(1):123–157, 2006.

[Hershberg and Margalit, 2006] R. Hershberg and H. Margalit, "Co-evolution of transcription factors and their targets depends on mode of regulation," *Genome Biology*, 7:R62, 2006.

[Hjort *et al.*, 2016] K. Hjort, H. Nicoloff, and D. I. Andersson, "Unstable tandem gene amplification generates heteroresistance (variation in resistance within a population) to colistin in Salmonella enterica," *Molecular Microbiology*, 102(2):274–289, 2016.

[Hnisz *et al.*, 2017] D. Hnisz, K. Shrinivas, R. A. Young, A. K. Chakraborty, and P. A. Sharp, "A Phase Separation Model for Transcriptional Control," *Cell*, 169(1):13–23, 2017.

[Hooper and Berg, 2003] S. D. Hooper and O. G. Berg, "Duplication is more common among laterally transferred genes than among indigenous genes.," *Genome biology*, 4(8):R48, 2003.

[Hopfield, 1974] J. J. Hopfield, "Kinetic Proofreading: A New Mechanism for Reducing Errors in Biosynthetic Processes Requiring High Specificity," *Proceedings of the National Academy of Sciences*, 71(10):4135–4139, 1974.

[Houle *et al.*, 2010] D. Houle, D. R. Govindaraju, and S. Omholt, "Phenomics: the next challenge," *Nature Reviews Genetics*, 11(12):855–866, 2010.

[Igler *et al.*, 2018] C. Igler, M. Lagator, G. Tkačik, J. P. Bollback, and C. C. Guet, "Evolutionary potential of transcription factors for gene regulatory rewiring," *Nature Ecology & Evolution*, 2(10):1633–1643, 2018.

[Islam *et al.*, 2014] S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lönnerberg, and S. Linnarsson, "Quantitative single-cell RNA-seq with unique molecular identifiers," *Nature Methods*, 11(2):163–166, 2014.

[Itzkovitz *et al.*, 2006] S. Itzkovitz, T. Tlusty, and U. Alon, "Coding limits on the number of transcription factors," *BMC Genomics*, 7(1):239, 2006.

[Jacob, 1977] F. Jacob, "Evolution and tinkering," *Science*, 196(4295):1161–1166, 1977.

[Jacob, 1994] F. Jacob, *The Possible and the Actual*, The Jessie and John Danz lectures. University of Washington Press, 1994.

[Jacquier *et al.*, 2013] H. Jacquier, A. Birgy, H. Le Nagard, Y. Mechulam, E. Schmitt, J. Glodt, B. Bercot, E. Petit, J. Poulain, G. Barnaud, P.-A. Gros, and O. Tenaillon, "Capturing the mutational landscape of the beta-lactamase TEM-1," *Proceedings of the National Academy of Sciences*, 110(32):13067–13072, 2013.

[Johnson *et al.*, 2005] J. M. Johnson, S. Edwards, D. Shoemaker, and E. E. Schadt, "Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments," *Trends in Genetics*, 21(2):93–102, 2005.

[Jones *et al.*, 2014] D. L. Jones, R. C. Brewster, and R. Phillips, "Promoter architecture dictates cell-to-cell variability in gene expression," *Science*, 346(6216):1533–1536, 2014.

[Juhas *et al.*, 2009] M. Juhas, J. R. Van Der Meer, M. Gaillard, R. M. Harding, D. W. Hood, and D. W. Crook, "Genomic islands: Tools of bacterial horizontal gene transfer and evolution," 2009.

[Kafri *et al.*, 2016] M. Kafri, E. Metzl-Raz, G. Jona, and N. Barkai, "The Cost of Protein Production," *Cell Reports*, 14(1):22–31, 2016.

[Kaizu *et al.*, 2014] K. Kaizu, W. De Ronde, J. Paijmans, K. Takahashi, F. Tostevin, and P. R. Ten Wolde, "The berg-purcell limit revisited," *Biophysical journal*, 106(4):976–985, 2014.

[Karlebach and Shamir, 2008] G. Karlebach and R. Shamir, "Modelling and analysis of gene regulatory networks," *Nature Reviews Molecular Cell Biology*, 9(10):770–780, 2008.

[Kassen and Bataillon, 2006] R. Kassen and T. Bataillon, "Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria," *Nature Genetics*, 38(4):484–488, 2006.

[Keightley, 2000] P. D. Keightley, "Deleterious Mutations and the Evolution of Sex," *Science*, 290(5490):331–333, 2000.

[Kemble *et al.*, 2019] H. Kemble, P. Nghe, and O. Tenaillon, "Recent insights into the genotype-phenotype relationship from massively parallel genetic assays," *Evolutionary Applications*, 12(9):1721–1742, 2019.

[Khlebnikov *et al.*, 2001] A. Khlebnikov, K. A. Datsenko, T. Skaug, B. L. Wanner, and J. D. Keasling, "Homogeneous expression of the P(BAD) promoter in Escherichia coli by constitutive expression of the low-affinity high-capacity AraE transporter.," *Microbiology (Reading, England)*, 147(Pt 12):3241–7, 2001.

[Kim *et al.*, 2009] H. D. Kim, T. Shay, E. K. O'Shea, and A. Regev, "Transcriptional Regulatory Circuits: Predicting Numbers from Alphabets," *Science*, 325(5939):429–432, 2009.

[Kinney and McCandlish, 2019] J. B. Kinney and D. M. McCandlish, "Massively Parallel Assays and Quantitative Sequence-Function Relationships," *Annual Review of Genomics and Human Genetics*, 20(1):annurev–genom–083118–014845, 2019.

[Kinney *et al.*, 2010] J. B. Kinney, A. Murugan, C. G. Callan, and E. 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.

[Klemm *et al.*, 2019] S. L. Klemm, Z. Shipony, and W. J. Greenleaf, "Chromatin accessibility and the regulatory epigenome.," *Nature Reviews Genetics*, 20(4):207–220, 2019.

[Koch, 1983] A. L. Koch, "The protein burden oflac operon products," *Journal of Molecular Evolution*, 19(6):455–462, 1983.

[Kryazhimskiy *et al.*, 2009] S. Kryazhimskiy, G. Tkačik, and J. B. Plotkin, "The dynamics of adaptation on correlated fitness landscapes," *Proceedings of the National Academy of Sciences*, 106(44):18638–18643, 2009.

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

[Kumar Prajapat *et al.*, 2016] M. Kumar Prajapat, K. Jain, D. Choudhury, N. Raj, and S. Saini, "Revisiting demand rules for gene regulation," *Molecular BioSystems*, 12(2):421–430, 2016.

[Kurland and Dong, 1996] C. G. Kurland and H. Dong, "Bacterial growth inhibition by overproduction of protein," *Molecular Microbiology*, 21(1):1–4, 1996.

[Kussell and Laibler, 2005] E. Kussell and Laibler, "Phenotypic Diversity, Population Growth, and Information in Fluctuating Environments," *Science*, 309(5743):2075–2078, 2005.

[Lagator *et al.*, 2017a] M. Lagator, T. Paixão, N. H. Barton, J. P. Bollback, and C. C. Guet, "On the mechanistic nature of epistasis in a canonical cis-regulatory element," *eLife*, 6, 2017.

[Lagator *et al.*, 2017b] M. Lagator, S. Sarikas, H. Acar, J. P. Bollback, and C. C. Guet, "Regulatory network structure determines patterns of intermolecular epistasis," *eLife*, 6, 2017.

[Lagator *et al.*, 2020] M. Lagator, S. Sarikas, M. Steinrück, D. Toledo-Aparicio, J. P. Bollback, G. Tkacik, and C. C. Guet, "Structure and Evolution of Constitutive Bacterial Promoters," *bioRxiv*, 2020.

[Landman *et al.*, 2017] J. Landman, R. C. Brewster, F. M. Weinert, R. Phillips, and W. K. Kegel, "Self-consistent theory of transcriptional control in complex regulatory architectures," *PLOS ONE*, 12(7):e0179235, 2017.

[Larson *et al.*, 2013] D. R. Larson, C. Fritzsch, L. Sun, X. Meng, D. S. Lawrence, and R. H. Singer, "Direct observation of frequency modulated transcription in single cells using light activation.," *eLife*, 2:e00750, 2013.

[Lässig, 2007] M. Lässig, "From biophysics to evolutionary genetics: statistical aspects of gene regulation," *BMC Bioinformatics*, 8(6):1–21, 2007.

[Latorre *et al.*, 2005] A. Latorre, R. Gil, F. J. Silva, and A. Moya, "Chromosomal stasis versus plasmid plasticity in aphid endosymbiont Buchnera aphidicola," *Heredity*, 95(5):339–347, 2005.

[Lehner, 2013] B. Lehner, "Genotype to phenotype: lessons from model organisms for human genetics," *Nature Reviews Genetics*, 14(3):168–178, 2013.

[Lercher and Pál, 2008] M. J. Lercher and C. Pál, "Integration of horizontally transferred genes into regulatory interaction networks takes many million years," *Molecular Biology and Evolution*, 25(3):559–567, 2008.

[Lestas *et al.*, 2008] I. Lestas, J. Paulsson, N. E. Ross, and G. Vinnicombe, "Noise in Gene Regulatory Networks," *Automatic Control, IEEE Transactions on*, 53:189–200, 2008.

[Lewin, 2007] B. Lewin, *Genes IX 9th edition by Lewin, Benjamin published by Jones & Bartlett Publishers Hardcover*, Jones & Bartlett Publishers, 2007.

[Longo and Hasty, 2006] D. Longo and J. Hasty, "Dynamics of single-cell gene expression," *Molecular Systems Biology*, 2(1):64, 2006.

[López-Maury *et al.*, 2008] L. López-Maury, S. Marguerat, and J. Bähler, "Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation," *Nature Reviews Genetics*, 9(8):583–593, 2008.

[Lutz and Bujard, 1997] R. Lutz and H. Bujard, "Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements.," *Nucleic acids research*, 25(6):1203–10, 1997.

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

[Maerkl and Quake, 2009] S. J. Maerkl and S. R. Quake, "Experimental Determination of the Evolvability of a Transcription Factor," *Proceedings of the National Academy of Sciences*, 106(44):18650–18655, 2009.

[Martin *et al.*, 2016] O. C. Martin, A. Krzywicki, and M. Zagorski, "Drivers of structural features in gene regulatory networks: From biophysical constraints to biological function," *Physics of Life Reviews*, 17:124–158, 2016.

[Mato *et al.*, 2017] L. Mato, P. Tiago, N. H. Barton, J. P. Bollback, and C. C. Guet, "On the mechanistic nature of epistasis in a canonical cis-regulatory element," *eLife; Cambridge*, 6, 2017.

[Metzger *et al.*, 2016] B. P. H. Metzger, F. Duveau, D. C. Yuan, S. Tryban, B. Yang, and P. J. Wittkopp, "Contrasting Frequencies and Effects of *cis* - and *trans* -Regulatory Mutations Affecting Gene Expression," *Molecular Biology and Evolution*, 33(5):1131–1146, 2016.

[Milocco and Salazar-Ciudad, 2020] L. Milocco and I. Salazar-Ciudad, "Is evolution predictable? Quantitative genetics under complex genotype-phenotype maps," *Evolution*, 74(2):230–244, 2020.

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

[Molina *et al.*, 2013] N. Molina, D. M. Suter, R. Cannavo, B. Zoller, I. Gotic, and F. Naef, "Stimulus-induced modulation of transcriptional bursting in a single mammalian gene," *Proceedings of the National Academy of Sciences*, 110(51):20563–20568, 2013.

[Morisaki *et al.*, 2014] T. Morisaki, W. G. Müller, N. Golob, D. Mazza, and J. G. McNally, "Single-molecule analysis of transcription factor binding at transcription sites in live cells.," *Nature Communications*, 5(1):4456, 2014.

[Moxon *et al.*, 1994] E. R. Moxon, P. B. Rainey, M. A. Nowak, and R. E. Lenski, "Adaptive evolution of highly mutable loci in pathogenic bacteria," *Current Biology*, 4(1):24–33, 1994.

[Nagai *et al.*, 2002] T. Nagai, K. Ibata, E. S. Park, M. Kubota, K. Mikoshiba, and A. Miyawaki, "A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications.," *Nature biotechnology*, 20(1):87–90, 2002.

[Nagaraj *et al.*, 2011] N. Nagaraj, J. R. Wisniewski, T. Geiger, J. Cox, M. Kircher, J. Kelso, S. Pääbo, and M. Mann, "Deep proteome and transcriptome mapping of a human cancer cell line," *Molecular Systems Biology*, 7(1):548, 2011.

[Nagelkerke and Postma, 1978] F. Nagelkerke and P. W. Postma, "2-Deoxygalactose, a specific substrate of the Salmonella typhimurium galactose permease: Its use for the isolation of galP mutants," *Journal of Bacteriology*, 133(2):607–613, 1978.

[Näsvall *et al.*, 2012] J. Näsvall, L. Sun, J. R. Roth, and D. I. Andersson, "Supplement: Real-time evolution of new genes by innovation, amplification, and divergence.," *Science (New York, N.Y.)*, 338(6105):384–7, 2012.

[Neve, 2007] P. Neve, "Challenges for herbicide resistance evolution and management: 50 years after Harper: Herbicide resistance challenges," *Weed Research*, 47(5):365–369, 2007.

[Nguyen and Saier, 1995] C. C. Nguyen and M. H. Saier, "Phylogenetic, structural and functional analyses of the LacI-GalR family of bacterial transcription factors," *FEBS Letters*, 377(2):98–102, 1995.

[Nguyen *et al.*, 1989] T. N. Nguyen, Q. G. Phan, L. P. Duong, K. P. Bertrand, and R. E. Lenski, "Effects of carriage and expression of the Tn10 tetracycline-resistance

operon on the fitness of Escherichia coli K12.," *Molecular biology and evolution*, 6(3):213–25, 1989.

[Nicolas *et al.*, 2018] D. Nicolas, B. Zoller, D. M. Suter, and F. Naef, "Modulation of transcriptional burst frequency by histone acetylation," *Proceedings of the National Academy of Sciences*, page 201722330, 2018.

[Nicoloff *et al.*, 2019] H. Nicoloff, K. Hjort, B. R. Levin, and D. I. Andersson, "The high prevalence of antibiotic heteroresistance in pathogenic bacteria is mainly caused by gene amplification," *Nature Microbiology*, 4(3):504–514, 2019.

[Novick and Weiner, 1957] A. Novick and M. Weiner, "Enzyme Induction as an All-or-None Phenomenon," *Proceedings of the National Academy of Sciences*, 43(7):553–566, 1957.

[Orr, 2003] H. A. Orr, "The Distribution of Fitness Effects Among Beneficial Mutations," *Genetics*, 163(4):1519–1526, 2003.

[Otwinowski and Nemenman, 2013] J. Otwinowski and I. Nemenman, "Genotype to Phenotype Mapping and the Fitness Landscape of the E. coli lac Promoter," *PLoS ONE*, 8(5):e61570, 2013.

[Pál *et al.*, 2005] C. Pál, B. Papp, and M. J. Lercher, "Adaptive evolution of bacterial metabolic networks by horizontal gene transfer.," *Nature genetics*, 37(12):1372–5, 2005.

[Park *et al.*, 2019] J. Park, J. Estrada, G. Johnson, B. J. Vincent, C. Ricci-Tam, M. D. Bragdon, Y. Shulgina, A. Cha, Z. Wunderlich, J. Gunawardena, and A. H. DePace, "Dissecting the sharp response of a canonical developmental enhancer reveals multiple sources of cooperativity.," *eLife*, 8:2787, 2019.

[Paulsson, 2004] J. Paulsson, "Summing up the noise in gene networks.," *Nature*, 427(6973):415–418, 2004.

[Payne and Wagner, 2014] J. L. Payne and A. Wagner, "The Robustness and Evolvability of Transcription Factor Binding Sites," *Science*, 343(6173):875–877, 2014.

[Perry *et al.*, 2007] G. H. Perry, N. J. Dominy, K. G. Claw, A. S. Lee, H. Fiegler, R. Redon, J. Werner, F. A. Villanea, J. L. Mountain, R. Misra, N. P. Carter, C. Lee, and A. C. Stone, "Diet and the evolution of human amylase gene copy number variation," *Nature Genetics*, 39(10):1256–1260, 2007.

[Petkova *et al.*, 2019] M. D. Petkova, G. Tkačik, W. Bialek, E. F. Wieschaus, and T. Gregor, "Optimal Decoding of Cellular Identities in a Genetic Network," *Cell*, 176(4):844–855.e15, 2019.

[Pettersson *et al.*, 2005] M. E. Pettersson, D. I. Andersson, J. R. Roth, and O. G. Berg, "The Amplification Model for Adaptive Mutation," *Genetics*, 169(2):1105–1115, 2005.

[Pettersson *et al.*, 2009] M. E. Pettersson, S. Sun, D. I. Andersson, and O. G. Berg, "Evolution of new gene functions: simulation and analysis of the amplification model.," *Genetica*, 135(3):309–324, 2009.

[Pfaffl, 2001] M. W. Pfaffl, "A new mathematical model for relative quantification in real-time RT-PCR," *Nucleic Acids Research*, 29(9):45e–45, 2001.

[Phillips *et al.*, 2012] R. Phillips, J. Theriot, J. Kondev, and H. Garcia, *Physical biology of the cell*, Garland Science, 2012.

[Price *et al.*, 2013] M. N. Price, A. M. Deutschbauer, J. M. Skerker, K. M. Wetmore, T. Ruths, J. S. Mar, J. V. Kuehl, W. Shao, and A. P. Arkin, "Indirect and suboptimal control of gene expression is widespread in bacteria," *Molecular Systems Biology*, 9(1):660, 2013.

[Ptashne, 1986] M. Ptashne, *A genetic switch: gene control and phage [lambda]*, Cell Press Cambridge, MA, 1986.

[Ptashne, 2011] M. Ptashne, "Principles of a switch," *Nature Chemical Biology*, 7(8):484–487, 2011.

[Qian and Kussell, 2016] L. Qian and E. Kussell, "Genome-Wide Motif Statistics

are Shaped by DNA Binding Proteins over Evolutionary Time Scales," *Physical Review X*, 6(4):041009, 2016.

[Reams *et al.*, 2010] A. B. Reams, E. Kofoid, M. Savageau, and J. R. Roth, "Duplication Frequency in a Population of Salmonella enterica Rapidly Approaches Steady State With or Without Recombination," *Genetics*, 184(4):1077–1094, 2010.

[Reams and Roth, 2015] A. B. Reams and J. R. Roth, "Mechanisms of gene duplication and amplification.," *Cold Spring Harbor perspectives in biology*, 7(2):a016592, 2015.

[Ren *et al.*, 2017] G. Ren, W. Jin, K. Cui, J. Rodrigez, G. Hu, Z. Zhang, D. R. Larson, and K. Zhao, "CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression.," *Molecular Cell*, 67(6):1049–1058.e6, 2017.

[Rice, 2006] J. A. Rice, *Mathematical Statistics and Data Analysis*, Duxbury Press, Belmont, CA, 3 edition edition, 2006.

[Rieckh and Tkačik, 2014] G. Rieckh and G. Tkačik, "Noise and information transmission in promoters with multiple internal states," *Biophysical journal*, 106(5):1194–1204, 2014.

[Rockel *et al.*, 2012] S. Rockel, M. Geertz, K. Hens, B. Deplancke, and S. J. Maerkl, "iSLIM: a comprehensive approach to mapping and characterizing gene regulatory networks," *Nucleic acids research*, page gks1323, 2012.

[Rodriguez-Beltran *et al.*, 2018] J. Rodriguez-Beltran, J. C. R. Hernandez-Beltran, J. DelaFuente, J. A. Escudero, A. Fuentes-Hernandez, R. C. MacLean, R. Peña-Miller, and A. San Millan, "Multicopy plasmids allow bacteria to escape from fitness trade-offs during evolutionary innovation," *Nature Ecology & Evolution*, 2(5):873–881, 2018.

[Roth *et al.*, 1996] J. R. Roth, N. Benson, T. Galitski, K. Haack, J. G. Lawrence, and L. Miesel, "Rearrangements of the Bacterial Chromosome: Formation and

Applications.," In F. C. Neidhardt, editor, *In Escherichia coli and Salmonella: Cellular and Molecular Biology*, volume 2, chapter 120, pages 2256–76. American Society for Microbiology, Washington, D.C., 2nd edition, 1996.

[Rowland *et al.*, 2017] M. A. Rowland, J. M. Greenbaum, and E. J. Deeds, "Crosstalk and the evolvability of intracellular communication," *Nature Communications*, 8:ncomms16009, 2017.

[Sabari *et al.*, 2018] B. R. Sabari, A. Dall'Agnese, A. Boija, I. A. Klein, E. L. Coffey, K. Shrinivas, B. J. Abraham, N. M. Hannett, A. V. Zamudio, J. C. Manteiga, C. H. Li, Y. E. Guo, D. S. Day, J. Schuijers, E. Vasile, S. Malik, D. Hnisz, T. I. Lee, I. I. Cisse, R. G. Roeder, P. A. Sharp, A. K. Chakraborty, and R. A. Young, "Coactivator condensation at super-enhancers links phase separation and gene control.," *Science*, 361(6400), 2018.

[Salgado *et al.*, 2013] H. Salgado, M. Peralta-Gil, S. Gama-Castro, A. Santos-Zavaleta, L. Muñiz Rascado, J. S. García-Sotelo, V. Weiss, H. Solano-Lira, I. Martínez-Flores, A. Medina-Rivera, G. Salgado-Osorio, S. Alquicira-Hernández, K. Alquicira-Hernández, A. López-Fuentes, L. Porrón-Sotelo, A. M. Huerta, C. Bonavides-Martínez, Y. I. Balderas-Martínez, L. Pannier, M. Olvera, A. Labastida, V. Jiménez-Jacinto, L. Vega-Alvarado, V. del Moral-Chávez, A. Hernández-Alvarez, E. Morett, and J. Collado-Vides, "RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more," *Nucleic Acids Research*, 41(D1):D203–D213, 2013.

[Sanchez and Kondev, 2008] A. Sanchez and J. Kondev, "Transcriptional control of noise in gene expression.," *Proceedings of the National Academy of Sciences*, 105(13):5081–5086, 2008.

[Sanjuán *et al.*, 2004] R. Sanjuán, A. Moya, and S. F. Elena, "The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus," *Proceedings of the National Academy of Sciences of the United States of America*, 101(22):8396–8401, 2004.

[Sarai and Takeda, 1989] A. Sarai and Y. Takeda, "Lambda repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically," *Proceedings of the National Academy of Sciences*, 86(17):6513–6517, 1989.

[Sasson *et al.*, 2012] V. Sasson, I. Shachrai, A. Bren, E. Dekel, and U. Alon, "Mode of Regulation and the Insulation of Bacterial Gene Expression," *Molecular Cell*, 46(4):399–407, 2012.

[Savageau, 1977] M. A. Savageau, "Design of molecular control mechanisms and the demand for gene expression," *Proceedings of the National Academy of Sciences*, 74(12):5647–5651, 1977.

[Savageau, 1983] M. A. Savageau, "Regulation of differentiated cell-specific functions," *Proceedings of the National Academy of Sciences*, 80(5):1411–1415, 1983.

[Savageau, 1974] M. A. Savageau, "Genetic Regulatory Mechanisms and the Ecological Niche of Escherichia coli," *Proceedings of the National Academy of Sciences*, 71(6):2453–2455, 1974.

[Savageau, 1998] M. A. Savageau, "Demand Theory of Gene Regulation. I. Quantitative Development of the Theory," *Genetics*, 149(4):1665–1676, 1998.

[Schneider *et al.*, 1986] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht, "Information content of binding sites on nucleotide sequences," *Journal of Molecular Biology*, 188(3):415–431, 1986.

[Schuster, 2006] P. Schuster, "Prediction of RNA secondary structures: from theory to models and real molecules," *Reports on Progress in Physics*, 69(5):1419–1477, 2006.

[Schuster *et al.*, 1994] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker, "From Sequences to Shapes and Back: A Case Study in RNA Secondary Structures," *Proceedings of the Royal Society of London B: Biological Sciences*, 255(1344):279–284, 1994.

[Segré *et al.*, 2000] D. Segré, D. Ben-Eli, and D. Lancet, "Compositional genomes: Prebiotic information transfer in mutually catalytic noncovalent assemblies," *Proceedings of the National Academy of Sciences*, 97(8):4112–4117, 2000.

[Senecal *et al.*, 2014] A. Senecal, B. Munsky, F. Proux, N. Ly, F. Braye, C. Zimmer, F. Mueller, and X. Darzacq, "Transcription Factors Modulate c-Fos Transcriptional Bursts," *Cell Reports*, 8(1):75–83, 2014.

[Shachrai *et al.*, 2010] I. Shachrai, A. Zaslaver, U. Alon, and E. Dekel, "Cost of Unneeded Proteins in E. coli Is Reduced after Several Generations in Exponential Growth," *Molecular Cell*, 38(5):758–767, 2010.

[Shea and Ackers, 1985] M. A. Shea and G. 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.

[Shinar *et al.*, 2006] G. Shinar, E. Dekel, T. Tlusty, and U. Alon, "Rules for biological regulation based on error minimization," *Proceedings of the National Academy of Sciences of the United States of America*, 103(11):3999–4004, 2006.

[Shlyueva *et al.*, 2014] D. Shlyueva, G. Stampfel, and A. Stark, "Transcriptional enhancers: from properties to genome-wide predictions," *Nature Reviews Genetics*, 15(4):272, 2014.

[Shultzaberger *et al.*, 2012] R. K. Shultzaberger, S. J. Maerkl, J. F. Kirsch, and M. B. Eisen, "Probing the Informational and regulatory plasticity of a transcription factor DNA–binding domain," *PLoS genetics*, 8(3), 2012.

[Singh, 2014] V. Singh, "Recent advancements in synthetic biology: Current status and challenges," *Gene*, 535(1):1–11, 2014.

[Soskine and Tawfik, 2010] M. Soskine and D. S. Tawfik, "Mutational effects and the evolution of new protein functions," *Nature Reviews Genetics*, 11(8):572–582, 2010.

[Spivak and Stormo, 2012] A. T. Spivak and G. D. Stormo, "ScerTF: a comprehensive database of benchmarked position weight matrices for Saccharomyces species," *Nucleic Acids Research*, 40(D1):D162–D168, 2012.

[Steinrueck and Guet, 2017] M. Steinrueck and C. C. Guet, "Complex chromosomal neighborhood effects determine the adaptive potential of a gene under selection," *eLife*, 6:1–26, 2017.

[Struhl, 2007] K. Struhl, "Transcriptional noise and the fidelity of initiation by RNA polymerase II.," *Nature Structural & Molecular Biology*, 14(2):103–105, 2007.

[Sun *et al.*, 2012] S. Sun, R. Ke, D. Hughes, M. Nilsson, and D. I. Andersson, "Genome-Wide Detection of Spontaneous Chromosomal Rearrangements in Bacteria," *PLoS ONE*, 7(8):e42639, 2012.

[Suter *et al.*, 2011] D. M. Suter, N. Molina, D. Gatfield, K. Schneider, U. Schibler, and F. Naef, "Mammalian genes are transcribed with widely different bursting kinetics.," *Science*, 332(6028):472–474, 2011.

[Szathmáry, 1993] E. Szathmáry, "Do deleterious mutations act synergistically? Metabolic control theory provides a partial answer.," *Genetics*, 133(1):127–132, 1993.

[Tao *et al.*, 1999] H. Tao, C. Bausch, C. Richmond, F. R. Blattner, and T. Conway, "Functional Genomics: Expression Analysis ofEscherichia coli Growing on Minimal and Rich Media," *Journal of Bacteriology*, 181(20):6425–6440, 1999.

[Tawfik, 2010] D. S. Tawfik, "Messy biology and the origins of evolutionary innovations," *Nature Chemical Biology*, 6(10):692–696, 2010.

[Taylor *et al.*, 2015] T. B. Taylor, G. Mulley, A. H. Dills, A. S. Alsohim, L. J. McGuffin, D. J. Studholme, M. W. Silby, M. A. Brockhurst, L. J. Johnson, and R. W. Jackson, "Evolutionary resurrection of flagellar motility via rewiring of the nitrogen regulation system," *Science*, 347(6225):1014–1017, 2015.

[Thomas *et al.*, 2019] C. Thomas, Y. Ji, C. Wu, H. Datz, C. Boyle, B. MacLeod, S. Patel, M. Ampofo, M. Currie, J. Harbin, K. Pechenkina, N. Lodhi, S. J. Johnson, and A. V. Tulin, "Hit and run versus long-term activation of PARP-1 by its different domains fine-tunes nuclear processes," *Proceedings of the National Academy of Sciences*, page 201901183, 2019.

[Tkačik and Bialek, 2016] G. Tkačik and W. Bialek, "Information processing in living systems," *Annual Review of Condensed Matter Physics*, 7:89–117, 2016.

[Tkačik *et al.*, 2008] G. Tkačik, T. Gregor, and W. Bialek, "The role of input noise in transcriptional regulation," *PloS one*, 3(7), 2008.

[Tkačik and Walczak, 2011] G. Tkačik and A. M. Walczak, "Information transmission in genetic regulatory networks: a review," *Journal of Physics: Condensed Matter*, 23(15):153102, 2011.

[Troein *et al.*, 2007] C. Troein, D. Ahrén, M. Krogh, and C. Peterson, "Is transcriptional regulation of metabolic pathways an optimal strategy for fitness?," *PloS one*, 2(9):e855, 2007.

[Trosset and Carbonell, 2013] J.-Y. Trosset and P. Carbonell, "Synergistic Synthetic Biology: Units in Concert," *Frontiers in Bioengineering and Biotechnology*, 1, 2013.

[Tuğrul *et al.*, 2015] M. Tuğrul, T. Paixão, N. H. Barton, and G. Tkačik, "Dynamics of Transcription Factor Binding Site Evolution," *PLoS Genet*, 11(11):e1005639, 2015.

[Ueda *et al.*, 2004] H. R. Ueda, S. Hayashi, S. Matsuyama, T. Yomo, S. Hashimoto, S. A. Kay, J. B. Hogenesch, and M. Iino, "Universality and flexibility in gene expression from bacteria to human," *Proceedings of the National Academy of Sciences*, 101(11):3765–3769, 2004.

[Veening *et al.*, 2008] J.-W. Veening, W. K. Smits, and O. P. Kuipers, "Bistability, Epigenetics, and Bet-Hedging in Bacteria," *Annual Review of Microbiology*, 62(1):193–210, 2008.

[Vilar, 2010] J. M. Vilar, "Accurate Prediction of Gene Expression by Integration of DNA Sequence Statistics with Detailed Modeling of Transcription Regulation," *Biophysical Journal*, 99(8):2408–2413, 2010.

[Von Hippel and Berg, 1986] P. H. Von Hippel and O. G. Berg, "On the specificity of DNA-protein interactions," *Proceedings of the National Academy of Sciences*, 83(6):1608, 1986.

[Von Hippel *et al.*, 1974] P. H. Von Hippel, A. Revzin, C. A. Gross, and A. C. Wang, "Non-specific DNA binding of genome regulating proteins as a biological control mechanism: 1. The lac operon: equilibrium aspects," *Proceedings of the National Academy of Sciences*, 71(12):4808–4812, 1974.

[Wagner, 2008] A. Wagner, "Robustness and evolvability: a paradox resolved," *Proceedings of the Royal Society B: Biological Sciences*, 275(1630):91–100, 2008.

[Wagner, 2005] A. Wagner, "Robustness, evolvability, and neutrality," *FEBS Letters*, 579(8):1772–1778, 2005.

[Wagner *et al.*, 1997] G. P. Wagner, G. Booth, and H. Bagheri-Chaichian, "A population genetic theory of canalization," *Evolution*, 51(2):329–347, 1997.

[Wagner and Zhang, 2011] G. P. Wagner and J. Zhang, "The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms," *Nature Reviews Genetics*, 12(3):204–213, 2011.

[Walczak *et al.*, 2012] A. M. Walczak, A. Mugler, and C. H. WIggins, "Analytic methods for modeling stochastic regulatory networks.," *Methods in molecular biology (Clifton, N.J.)*, 880(Chapter 13):273–322, 2012.

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

[Wang *et al.*, 2010] P. Wang, L. Robert, J. Pelletier, W. L. Dang, F. Taddei, A. Wright,

and S. Jun, "Robust Growth of Escherichia coli," *Current Biology*, 20(12):1099–1103, 2010.

[Wang *et al.*, 2009] Y. Wang, L. Guo, I. Golding, E. C. Cox, and N. P. Ong, "Quantitative Transcription Factor Binding Kinetics at the Single-Molecule Level," *Biophysical Journal*, 96(2):609–620, 2009.

[Watts and Strogatz, 1998] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," 393:3, 1998.

[Wei *et al.*, 2001] Y. Wei, J.-M. Lee, C. Richmond, F. R. Blattner, J. A. Rafalski, and R. A. LaRossa, "High-Density Microarray-Mediated Gene Expression Profiling of Escherichia coli," *Journal of Bacteriology*, 183(2):545–556, 2001.

[Wolf *et al.*, 2015] L. Wolf, O. K. Silander, and E. van Nimwegen, "Expression noise facilitates the evolution of gene regulation," *eLife*, 4:1–48, 2015.

[Wunderlich and Mirny, 2009] Z. Wunderlich and L. A. Mirny, "Different gene regulation strategies revealed by analysis of binding motifs," *Trends in Genetics*, 25(10):434–440, 2009.

[Yi and Dean, 2019] X. Yi and A. M. Dean, "Adaptive Landscapes in the Age of Synthetic Biology," *Molecular Biology and Evolution*, 36(5):890–907, 2019.

[Yona *et al.*, 2018] A. H. Yona, E. J. Alm, and J. Gore, "Random sequences rapidly evolve into de novo promoters," *Nature Communications*, 9(1):1530, 2018.

[Yosef and Regev, 2011] N. Yosef and A. Regev, "Impulse Control: Temporal Dynamics in Gene Transcription," *Cell*, 144(6):886–896, 2011.

[Zenklusen *et al.*, 2008] D. Zenklusen, D. R. Larson, and R. H. Singer, "Single-RNA counting reveals alternative modes of gene expression in yeast.," *Nature Structural & Molecular Biology*, 15(12):1263–1271, 2008.

[Zhou *et al.*, 2017] L. Zhou, Q. Ding, G.-Z. Jiang, Z.-N. Liu, H.-Y. Wang, and G.-R. Zhao, "Chromosome engineering of Escherichia coli for constitutive production of salvianic acid A," *Microbial Cell Factories*, 16(1):84, 2017.

[Zoller *et al.*, 2015] B. Zoller, D. Nicolas, N. Molina, and F. Naef, "Structure of silent transcription intervals and noise characteristics of mammalian genes," *Molecular Systems Biology*, 11(7):823–823, 2015.

[Zoller *et al.*, 2018] B. Zoller, S. C. Little, and T. Gregor, "Diverse Spatial Expression Patterns Emerge from Unified Kinetics of Transcriptional Bursting.," *Cell*, 175(3):835–847.e25, 2018.