

Chapter 5

Understanding larger graphs

This chapter presents especially illuminating findings and approaches from a broader set that GS and others derive in the Supplement of Reference [1].

In the preceding chapter, we illustrated how to use diagrammatic reasoning to learn about the nonequilibrium response of a small, extremely pervasive network motif. In this expository chapter, we use the Matrix Tree Theorem to quantify the behaviors of a panoply of important regulatory scenarios beyond the simple square graph, in or out of equilibrium. The same approaches deployed on the square graph yield insights into the output behaviors of much larger graphs and more complex networks. We organize this discussion as follows. First, as a motivating preview, we curtly summarize algebraic results developed for networks later in the chapter. Second, we recall the need for understanding such networks by parsing and visualizing the distribution of currently known regulatory interactions in *E. coli* and the fly, extracting revealing statistical summaries of how promiscuously transcription factors and promoters appear to interact. Last, we give a conceptual exposition on how the tree theorem is a tool for thinking even about such complex scenarios and derive the regulatory response functions for arbitrary hypercubic

graphs by understanding their structural constraints using an arboreal lens.

5.1 Summary of nonequilibrium generalization of important transcriptional motifs

We summarize key new network calculations in the Table [5.1](#) that follows.

regulatory architecture	control parameters	network topology	equilibrium behavior, $\langle r \rangle^{\text{eq}}$	general nonequilibrium behavior, $\langle r \rangle$	attained powers in $\langle r \rangle$
simple (strict) repression [cf. Bintu [2] case 1]	one: $[X]$		$\frac{A+B[X]}{C+D[X]}$	$\frac{A+B[X]}{C+D[X]}$	
simple "activation" (this study's emphasis) [cf. Bintu [2] case 2]	one: $[X]$		$\frac{A+B[X]}{C+D[X]}$	$\frac{A+B[X]+C[X]^2}{D+E[X]+F[X]^2}$	
activator recruited by a helper (symmetric) [cf. Bintu [2] cases 3 and 7]	two: $[X_1], [X_2]$		$\frac{A+B[X_1]+C[X_2]+D[X_1][X_2]}{E+F[X_1]+G[X_2]+H[X_1][X_2]}$	$\frac{\sum_{i \leq 4, j \leq 4} A_{ij} [X_1]^i [X_2]^j}{\sum_{i \leq 4, j \leq 4} B_{ij} [X_1]^i [X_2]^j}$	
strict repressor recruited by a helper [cf. Bintu [2] case 4]	two: $[R], [H]$		$\frac{A+B[H]+C[R]+D[H][R]}{E+F[H]+G[R]+H[H][R]}$	$\frac{\sum_{i \leq 3, j \leq 2} A_{ij} [H]^i [R]^j}{\sum_{i \leq 3, j \leq 2} B_{ij} [H]^i [R]^j}$	
dual strict repressors [cf. Bintu [2] cases 5 and 6]	two: $[X_1], [X_2]$		$\frac{A+B[X_1]+C[X_2]+D[X_1][X_2]}{E+F[X_1]+G[X_2]+H[X_1][X_2]}$	$\frac{\sum_{i \leq 2, j \leq 2} A_{ij} [X_1]^i [X_2]^j}{\sum_{i \leq 2, j \leq 2} B_{ij} [X_1]^i [X_2]^j}$	
dual activators cooperating via looping [cf. Bintu [2] case 8]	two: $[X_1], [X_2]$		$\frac{A+B[X_1]+C[X_2]+D[X_1][X_2]}{E+F[X_1]+G[X_2]+H[X_1][X_2]}$	$\frac{\sum_{i \leq 4, j \leq 5} A_{ij} [X_1]^i [X_2]^j}{\sum_{i \leq 4, j \leq 5} B_{ij} [X_1]^i [X_2]^j}$	
repressor with two DNA binding units and DNA looping [cf. Bintu [2] case 9]	one: $[X]$		$\frac{A+B[X]+C[X]^2}{D+E[X]+F[X]^2}$	$\frac{A+B[X]+C[X]^2+D[X]^3+E[X]^4}{F+G[X]+H[X]^2+I[X]^3+J[X]^4}$	
m independent transcription factors [cf. Spirit of Bintu [2] case 10]	$m, \{[X_i]\}$		$\frac{\sum_{i_1, i_2, \dots, i_m \in \{0,1\}} \alpha_{(i_1)} \prod_{l=1}^m [X_l]^{i_l}}{\sum_{i_1, i_2, \dots, i_m \in \{0,1\}} \beta_{(i_1)} \prod_{l=1}^m [X_l]^{i_l}}$	$\frac{\sum_{i_{(1)} = i_{(2)} = \dots = i_{(m)} \leq 2^{m-1}} A_{(i_1)} \prod_{i=1}^m [X_i]^{i_i}}{\sum_{i_{(1)} = i_{(2)} = \dots = i_{(m)} \leq 2^{m-1}} B_{(i_1)} \prod_{i=1}^m [X_i]^{i_i}}$	

Table 5.1: Summary of response functions of a variety of important transcriptional regulatory scenarios, generalized here to apply in or out of equilibrium. These examples include generalizing those formerly analyzed by equilibrium statistical mechanics in Reference [2]. The rightmost column visualizes the attained powers of the response function (namely, the domains of the sums in the middle two columns); square boxes denote the constrained equilibrium response. Note that the denominator and the numerator of the response $\langle r \rangle$ have the same algebraic form; they differ only on the coefficients of each monomial term. By construction, the coefficients of the polynomial in the denominator (the sum of all rooted and directed spanning trees, behaving like a partition function) are all positive. However, some of the coefficients of the numerator can be zero for a particular observable.

Specifically, later in this chapter we will focus in detail on our derivations of the regulatory response functions for the hypercubic graphs and their structural derivatives. (We leave details about our additional derivations of DNA looping interactions to be found in the Supplementary Information of Reference [1].)

Next, we review the state of knowledge about regulation in *E. coli*, and *D. melanogaster*, by extracting fresh up-to-date empirical histograms of regulatory architectures and their incidence. The data of interest are how many distinct transcription factors (whether distinct in biological identity or not) can simultaneously occupy each promoter or gene; this gives a putative upper bound on the dimensionality of the state space of the underlying regulatory graph.

5.2 Census of known regulatory interactions in *E. coli*

To understand what state space topologies are most common in *E. coli*, we reanalyzed the bipartite¹ network of all known regulatory interactions between transcription factors and promoters, as reported in the most recent release 11.2 of the canonical database *RegulonDB* [3]. These data report 216 unique transcription factors and 1152 unique regulated promoters. During data processing, we parsed all interactions (conservatively retaining all interactions, reported as strong or weak evidence; this choice does not significantly the incidences observed).

¹A network is called *bipartite* if its nodes can be divided into two classes, and all edges in the network connect nodes of one class to the other class. This is naturally the case for how promoters and transcription factors interact.

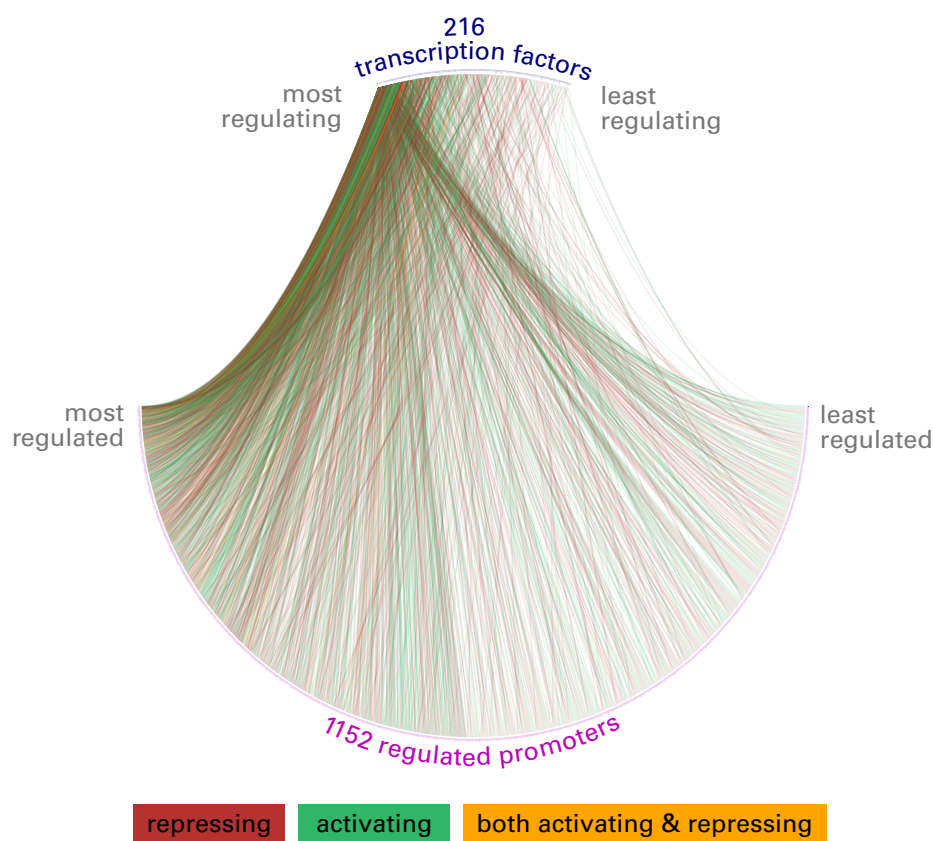


Figure 5.1: The bipartite network of all known regulatory interactions between **transcription factors** (top, navy nodes) and genomic **promoters** (bottom, magenta nodes) in the bacterium *E. coli*, constructed from raw data scraped from the RegulonDB database release 11.2 as of October 2023 [3]. Each arc visualizes an interaction between a transcription factor node on top and a promoter node on the bottom; repressing interactions are red, activating interactions green; and interactions that can be activating or repressing are orange. Nodes are arranged in decreasing degree from left to right on their respective arcs, such that the most highly regulated promoters (starting with *csgDp*, *flhDp*, *gadAp*, *hdeAp*, ...) appear on the left on the bottom arc, and the most promiscuous/highly regulating transcription factors (starting with *CRP*, *FNR*, *Fis*, *H-NS*, *IHF*, *Arca*, ...) appear on the left on the top arc.

While on aggregate, it is true that the most common regulated architecture across these interactions is putatively the square graph (the subject of the preceding chapter), gazing at the whole network as a gestalt makes clear that plenty of interesting biological promoters are associated with more than one transcription factor.

To gain some emotional and quantitative understandings of this fact, we visualize all the network interactions of *E. coli* in a three separate ways. First, we visualize the network as a bipartite graph, where one set of nodes are transcription factors and the other set of nodes are promoters, and we draw a subtle arc edge between a transcription factor and a promoter when the database reports an interaction. To make this overwhelming network slightly more informative, we also graphically sort nodes in decreasing order of their number of regulated or regulating interactions. The result, shown in Fig. 5.2, gives the indelible impression that at least some promiscuous interactions are central to the regulatory logic of this organism.

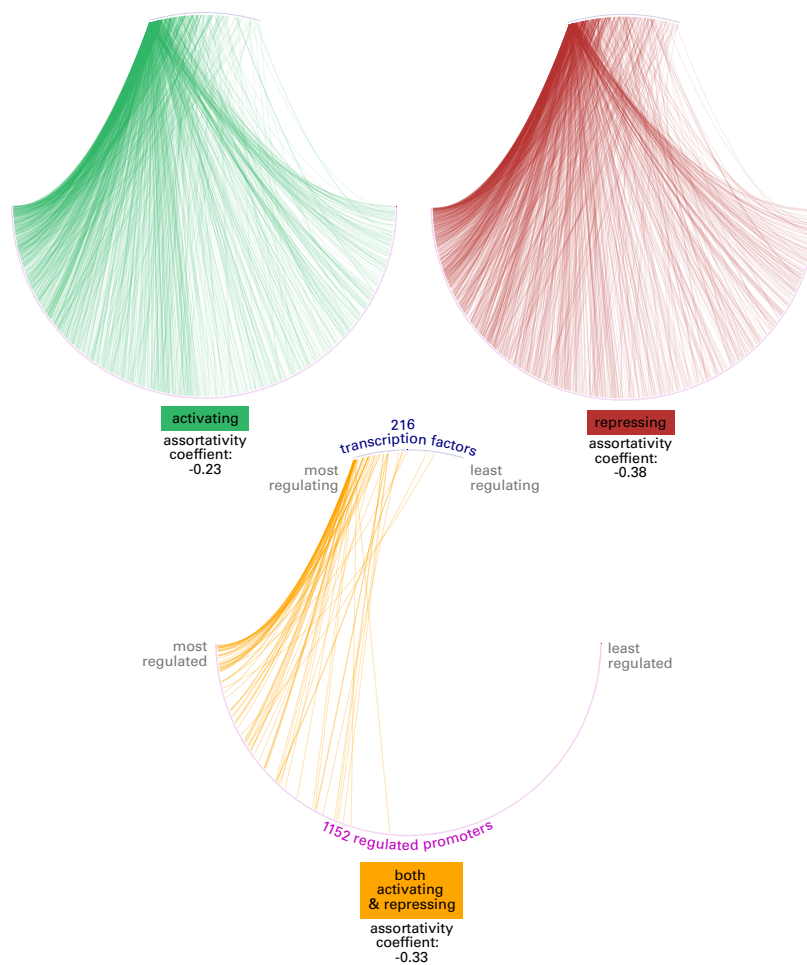


Figure 5.2: *E. coli*'s regulatory network as in Fig. 5.2, explicitly partitioned by the type of regulatory interaction (whether a transcription factor represses a promoter, activates it, or both).

Notice that while highly-regulating transcription factors are densely connected to highly-regulated promoters—as expected by chance alone given that highly-connected nodes are highly represented among edges—there is also a significant density of connections between highly regulating transcription factors and sparsely-regulated promoters, and vice versa (streaks from upper left to lower right, and from lower left to upper right). In fact, this tendency to connect highly-connected nodes to sparsely-connected nodes is quantitatively significant beyond what is expected by chance. This is reflected by appreciably negative “assortativity coefficients” that measure the average correlation between a node’s degree and the degrees of nodes it is connected to. Assortativity coefficient values (here, consistently of order -0.3) are normalized to vary from 1 (completely assortative; highly-connected nodes completely prefer connecting to other highly-connected nodes) to -1 (completely disassortative; highly-connected nodes completely prefer connecting to sparsely-connected nodes) [4]. Interestingly, activating interactions (assortativity coefficient -0.23) are detectably less disassortative than repressing interactions (assortativity coefficient -0.38). That is, while both activators and repressors tend to regulate promoters whose connectivity is different than their own, repressing transcription factors that regulate many promoters are even more likely to regulate promoters that are not already very degenerately repressed, compared to activating transcription factors. Also, note the curious fact that all of the (small number of) transcription factors that can act either as repressors or activators (shown in the orange network at bottom) appear to regulate only the top half of highly-regulated promoters.

Such hairball visualizations gives one gestalt of this network, but taken alone give an incomplete impression of this regulation. This is true because the promiscuity of some central interactions (obviously on display as dense node connections) nonetheless coexists with the fact that most promoters are regulated only by one transcription factor. To better appreciate how these facts coexist, second, we visualize the sparsity of this interaction

matrix using an explicit heatmap, as shown by Fig. 5.3. We order the rows (promoters) and columns (transcription factors) of this biadjacency matrix's heatmap in descending order of their promiscuity, and also annotate their empirical cumulative distribution functions as marginal histograms. These visualizations (as in the vast swaths of white (no-interaction) space in the heatmap) reflect the sparsity of this regulatory network, ultimately reflecting the sharply decaying probability of interactions reflected by these distribution functions. See https://www.its.caltech.edu/~gsalmon/gs069_Ecoli_biadjacency_RegulonDB11pt2.html for an interactive version of a heatmap where individual interactions can be gleaned by hovering with a mouse.

To understand the rapid decays in these probabilities of interactions with their promiscuity, we also consider them on a log-log axis, as shown by Fig. 5.4, and find they are acceptably described by power law behaviors. These plots affirm that, for instance, observing promoters regulated by k transcription factors is strongly (superlinearly) unlikely in k . This behavior, along with the large total number of possible interactions, reconciles the fact that the most common regulatory motif remains one transcription factor per promoter, but that with long tails in these distributions, there are some rare promoters regulated by many transcription factors.



Figure 5.3: *E. coli*'s regulatory network as in Fig. 5.2, now visualized as a biadjacency matrix where a pixel (i, j) is white if no interaction is reported between promoter i and transcription factor j , and colored otherwise if this regulation is reported. Rows and columns are arranged in decreasing degree from top to bottom and left to right; this means the total number of transcription factors associated with a given promoter, or the total number of promoters associated with a given transcription factor, trace out the empirical cumulative distributions of the respective degrees, visualized as marginal plots. Note the striking sparsity of the regulatory network; only 2,355 out of the 248,832 possible (transcription factor, promoter) interactions in this table are present (0.95%), and if all the other putatively “unregulated” promoters (not visualized in this table) are included this fraction is only 0.27% (2,355 out of 868,968).

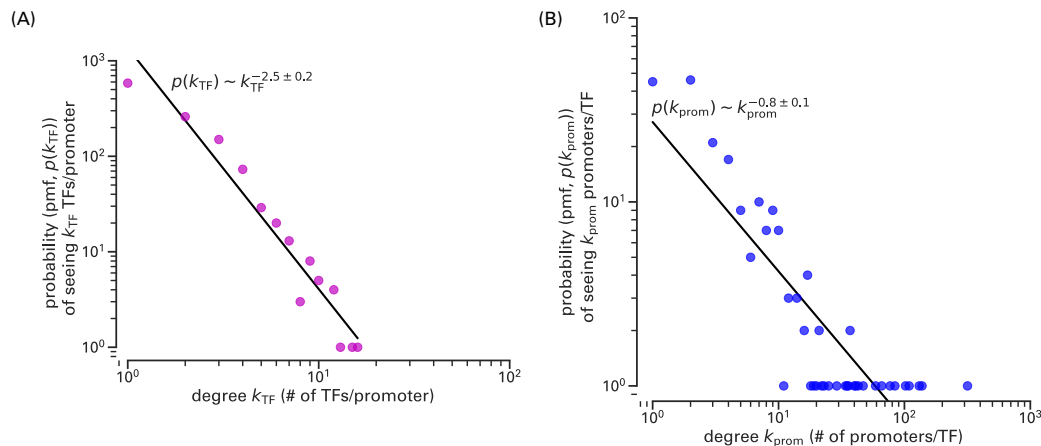


Figure 5.4: The two bipartite degree distributions in *E. coli* are acceptably well described by power laws. (A) The probability $p(k_{\text{TF}})$ that k_{TF} transcription factors regulate a promoter scales as $\sim k_{\text{TF}}^{-2.5}$ (black line; given by a regression minimizing relative error). This strongly-superlinearly decaying relationship says, for instance, that seeing a promoter with twice a given number of associated transcription factors is $2^{-2.5} \approx$ over 5.6 times less likely. (B) The probability of seeing a given number of promoters associated with a transcription factor is weakly sublinear in this number, $p(k_{\text{prom}}) \sim k_{\text{prom}}^{-0.8}$.

5.2.1 Natural statistics of regulatory architectures in *E. coli*

For a more coarse-grained understanding of the statistics of this regulation, we summarized the raw interaction data just discussed in a yet separate way. We counted each distinct binding site according to whether the transcription factor is reported to be repressing, activating, or more rarely both at each distinctly-positioned binding site on the genome. Fig. 5.5 schematizes this count; the indicated layers stratify both the type of interactions (indicated by layer color) and whether the interactions are annotated with the required granularity of the number of binding sites between the interacting transcription factor and promoter (indicated by dotted or solid layer outline).

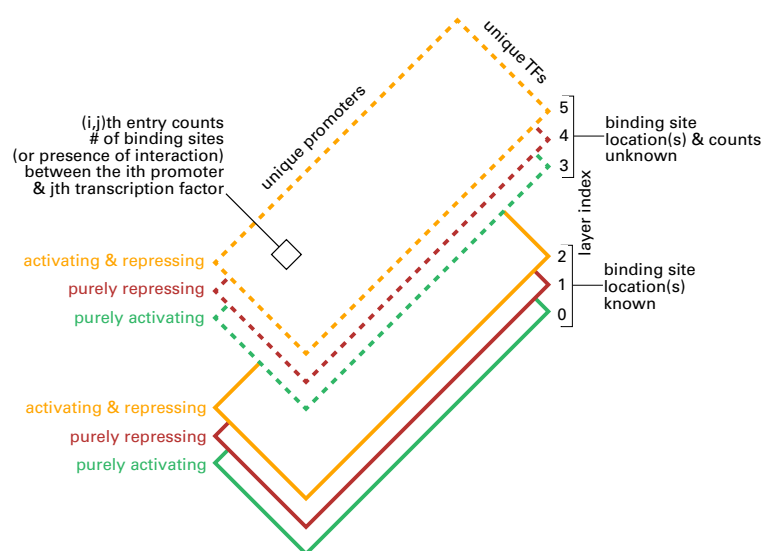


Figure 5.5: Categorical classification of counts of binding sites between promoters and transcription factors, separately acknowledging regulatory interactions where the locations or counts of binding site are not known/reported (dashed outlined layers). Since interactions without any localized binding sites cannot report how many individual transcription factors plausibly reside on the given promoter, the entries without specific binding site locations reported (layers 3-5) were omitted from future counts.

Next, we chose and followed a convention for counting the regulatory architectures achievable by the rare but present cases of binding sites identified as both repressing and activating.

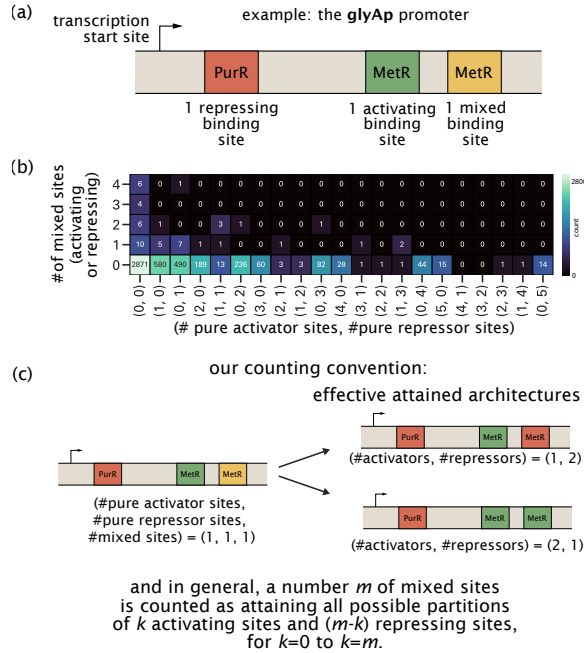


Figure 5.6: Accommodating cases in the regulatory network where individual binding sites (on a promoter for a particular transcription factor) are reported to be both activating and repressing. (A) One example of this type of mixed binding site is found in the *glyAp* promoter, where repressing, activating and mixed binding sites are respectively represented in red, green and yellow. (B) Heatmap showing counts of these mixed binding sites as rows, against columns depicting different counts of binding sites annotated as pure activating or repressing. Mixed sites are uncommon, but present. (C) Since such mixed sites can act as either repressors or activators, we adopt the convention of counting all possible effective architectures where the mixed binding sites assume either repressing roles (e.g. top) and/or activating role (e.g. bottom)s. Mathematically, this means that when there are m mixed binding sites, k sites can act as extra activating sites and $(m - k)$ can act as extra repressing sites, for $0 \leq k \leq m$. This convention allows us to collapse the architectures to represent examples of some number of activators and some number of repressors on the promoter.

As illustrated in Fig. 5.6, when we see a promoter with r purely repressive binding sites; a purely activating binding sites; and m sites that can be either repressing or activating, we count this scenario as attaining multiple architectures, namely the set $\{(a + k, r + (m - k))\}_{0 \leq k \leq m}$. Note that this convention only (very mildly) shapes the observed empirical distributions of architectures, since appreciably less than a percent of

all binding sites are annotated as both repressing and activating. Further, by construction, this convention does not affect the total copy number of transcription factors that are counted per promoter, the number of interest in defining what state space topologies promoters have. The resulting raw counts of regulatory architectures are summarized in Fig. 5.7. We also note that this observed regulatory architectures from an aggregating database is also highly compatible with the results of individual concerted experimental surveys.

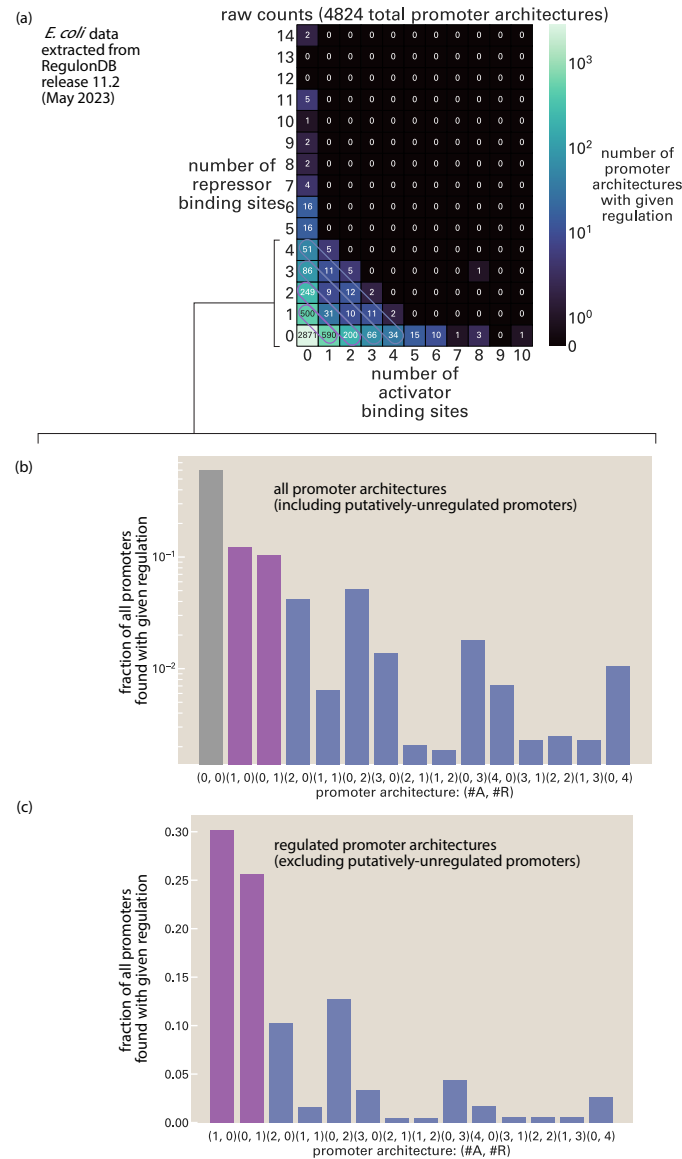


Figure 5.7: Incidence of regulatory architectures reported in *E. coli*, according to data in the RegulonDB database [3] release 11.2 (May 2023), organized by the joint numbers of activating and repressing transcription factors associated with each promoter. Architectures are regarded as the copy number of possibly-non-distinct proteins that can be bound to sites on each promoter. (A) The joint empirical distribution of the number of repressors and activators is largely symmetric (that is, $p(\#A, \#R) \approx p(\#R, \#A)$). (B) The joint distribution projected on a one-dimensional histogram on a logarithmic y -scale, including putatively-unregulated promoters with no known regulatory interactions (in grey) that comprise the majority ($\sim 71\%$) of promoters. Purple bars highlight promoters reported to have one associated transcription factor. (B) The incidence of regulatory architectures among those promoters with at least one known regulatory interaction. Over half of these regulated promoters are reported to be regulated by the most common architecture of a single transcription factor (purple bars), and thus are plausibly describable by the square graph of states analyzed most thoroughly in the preceding chapter's main text. (This figure forms the basis of Fig. 2A in the preceding chapter's main text.)

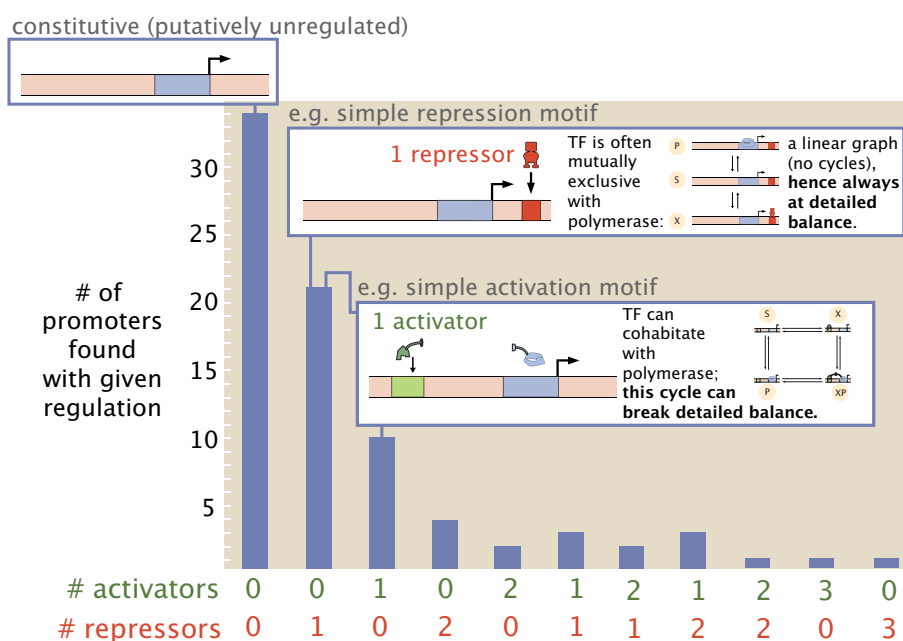


Figure 5.8: An experimental histogram of empirically-observed gene regulatory motifs in *E. coli* [5] reveals that many promoter sites are regulated by a single repressor or activator. A single repressor can often implement the simple repression motif, where the repressor excludes the polymerase from binding, allowing just three states in a linear graph. (Raw histogram data are courtesy of Reference [5].) The reason that the “simple activation motif” is schematized as linked to both the (0 activator, 1 repressor) and (1 repressor, 0 activator) histogram bar is that while steric exclusion commonly occurs for repressors, often making single repressors well described by a linear graph of three states, some repressors do not completely exclude the polymerase, permitting a cycle motif too.

Reference [5] is among the widest experimental censuses discovering regulatory interactions in *E. coli* in the recent literature. This study also found that transcriptional architectures with one activator or repressor are the most commonly observed regulated transcriptional architectures.

Thanks to common steric overlaps between the repressor binding site and polymerase binding site [6], a repressor is often—though not necessarily—mutually exclusive with the polymerase (see Fig. 5.8). In this case, the repressor implements a simple repression motif, a graph which lacks a cycle [7]. However, when the repressor does not sterically exclude the polymerase, a cycle of four states emerges. The same cycle of four states emerges with activators, whose binding sites rarely directly overlap with the polymerase binding site [6]; this produces a “simple activation motif.” These observations affirm that a single cycle of four states is a common motif in prokaryotic gene regulation, but other motifs are also widely observed.

5.3 Incidence of known transcriptional architectures in *Drosophila* from Redfly

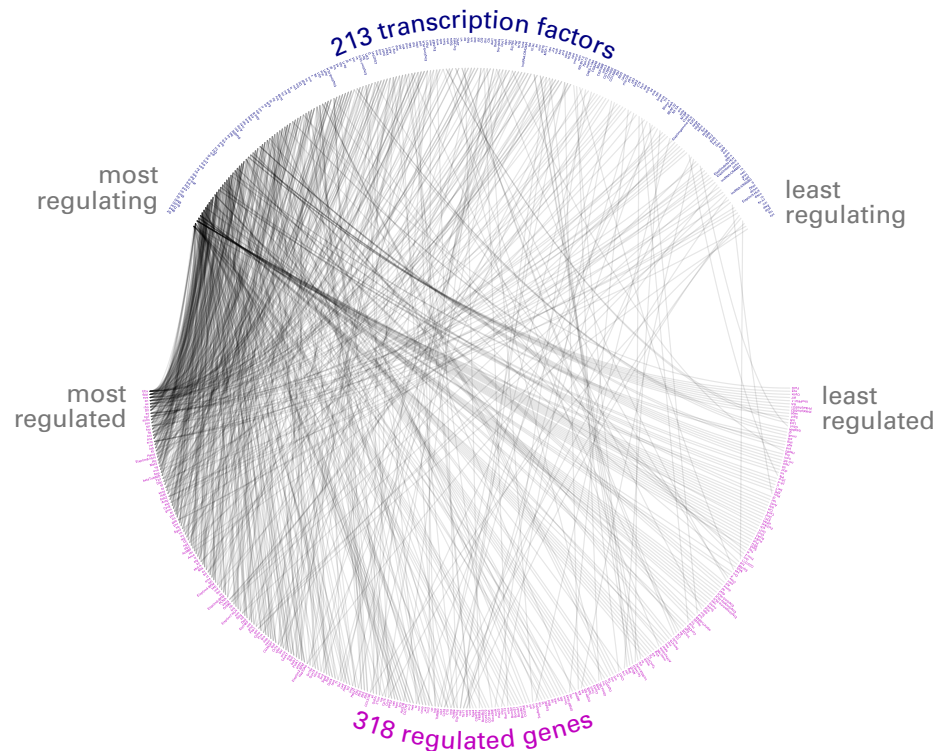


Figure 5.9: The bipartite network of all known regulatory interactions between **transcription factors** (top, navy nodes) and **genes** (bottom, magenta nodes) in the fruit fly eukaryote *D. melanogaster*, constructed from raw data scraped from the Redfly database v9.6.2 as of October 2023 [8] (http://redfly.ccr.buffalo.edu/about_redfly.php). Each black arc visualizes an interaction between a transcription factor node on top and a gene node on the bottom. Nodes are arranged in decreasing degree from left to right on their respective arcs, such that the most highly regulated genes (starting with Ubx, dpp, eve, en...) appear on the left on the bottom arc, and the most promiscuous/highly regulating transcription factors (starting with Dif, Mad, Derf, Su(h), ...) appear on the left on the top arc. As hinted at by the surprisingly dense connections from upper left to lower right or lower right to upper left, *Drosophila*'s network is also detectably disassortative, with an apparent assortative coefficient of -0.18; this is moderately less disassortative than *E. coli*'s regulatory network.

Having just analyzed how common regulatory interactions of different promiscuity (and so, different dimensionality of the underlying graph of state spaces) appear to be in a prokaryote, it is interesting to contrast this with that in a eukaryotic context. Fig. 5.9 visualizes the regulatory network in the fruit fly we parsed from the *Redfly* database; we visualized these interactions according to a similar scheme as in our hairball network visualization of *E. coli* in Fig. 5.2.

Similarly, Fig. 5.10 reports the biadjacency matrix of this regulatory network; Fig. 5.11 depicts power law behavior to regulatory degrees, as in *E. coli*; and Figs. 5.12 and 5.12 summarize these networks in terms of the incidence of certain types of regulatory architecture.

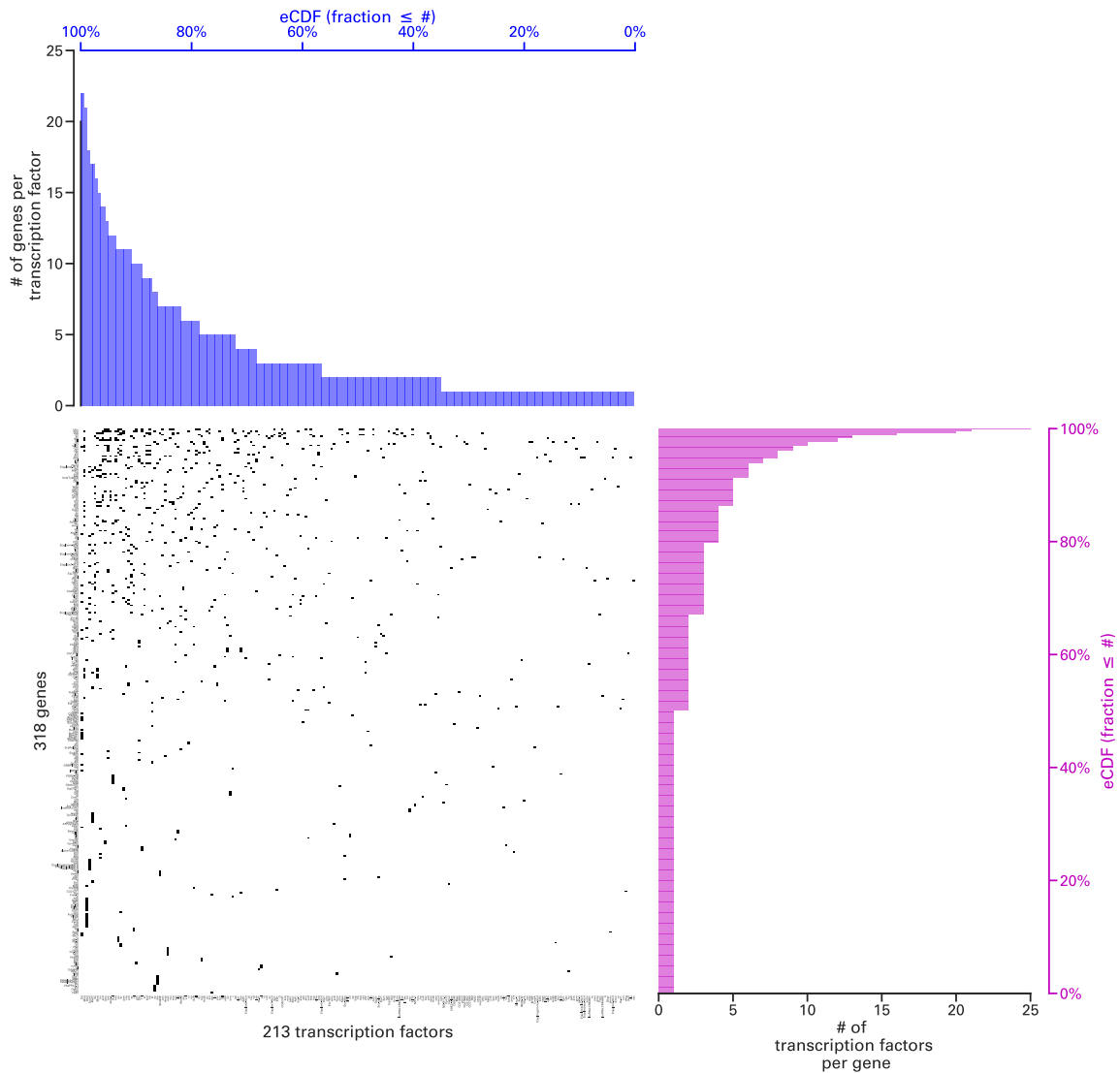


Figure 5.10: Like *E. coli*'s regulatory network, *Drosophila*'s regulatory network is very sparse: only 831 of 67,734 (regulated TF, gene) possible pairs are genuinely reported to be regulatory ($\sim 1.2\%$). If we also count the 17,410 additional putatively-unregulated genes, the proportion drops to 831 of 3,776,064 (or just 0.02%). See https://www.its.caltech.edu/~gsalmon/gs069_Dmelanogaster_biadjacency_Redfly202310.html for an interactive version of the heatmap where individual interactions can be gleaned by hovering with a mouse.

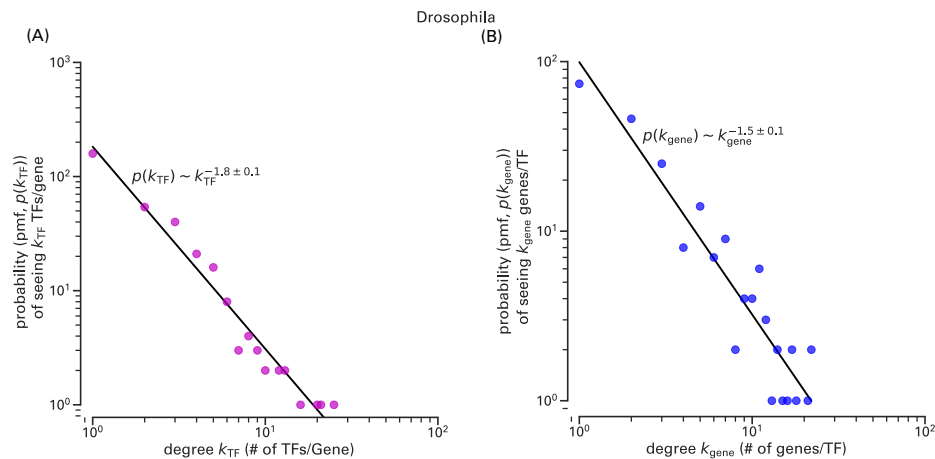


Figure 5.11: The two bipartite degree distributions in *Drosophila* are acceptably well described by power laws. (A) The probability $p(k_{\text{TF}})$ that k_{TF} transcription factors regulate a gene scales as $\sim k_{\text{TF}}^{-1.8}$ (black line; given by a regression minimizing relative error). This superlinearly decaying relationship says that seeing a gene with twice a given number of associated transcription factors is $2^{-1.8} \approx 3.5$ times as unlikely. (B) The probability of seeing a given number of promoters associated with a transcription factor also superlinearly decays with degree, $p(k_{\text{prom}}) \sim k_{\text{prom}}^{-1.5}$.

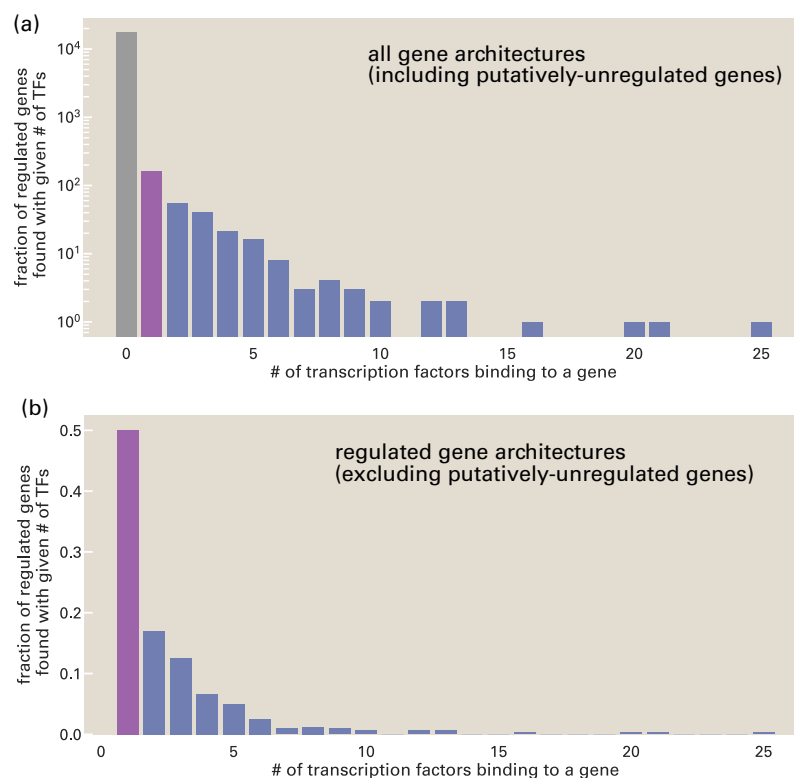


Figure 5.12: Histograms of number of distinct transcription factors interacting with each gene in *Drosophila*, according to data reported the Redfly database [8] v9.6.2 as of October 2023. (A) Histogram of the number of distinct transcription factors reported to bind a gene, including the putatively-unregulated genes in *Drosophila* (gray bar). (B) Histogram of the number of transcription factors associated with genes with at least one reported regulatory transcription factor; the purple bar represents the large proportion (about half) of regulated genes reported to be regulated by a single transcription factor. These data (up to five transcription factors) form the inset of Fig. 2A in the preceding chapter.

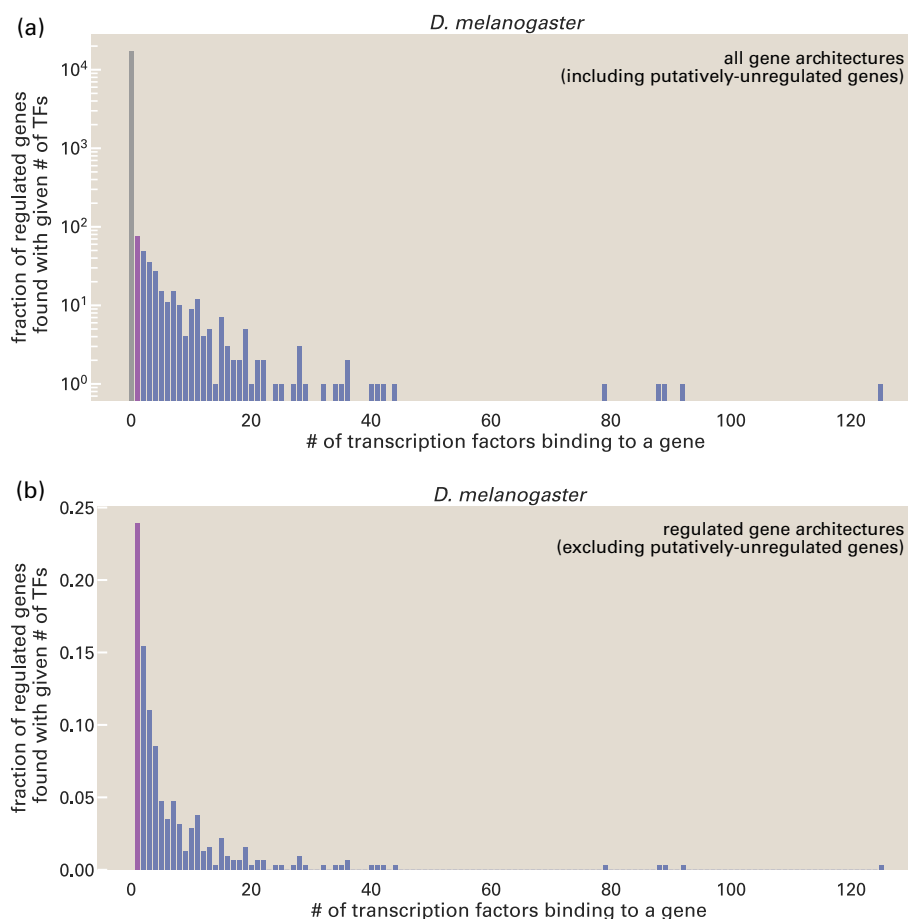


Figure 5.13: Histograms of the copy numbers of transcription factors capable of interacting with each gene in *Drosophila*, equivalent to the numbers of transcription factor reported binding sites per gene, according to data reported the Redfly database [8] v9.6.2 as of October 2023. (A) Histogram of the number of distinct transcription factors reported to bind a gene, including the putatively-unregulated genes in *Drosophila* (gray bar). (B) Histogram of the number of transcription factors associated with genes with at least one reported regulatory transcription factor; the purple bar represents the large proportion (about half) of regulated genes reported to be regulated by a single transcription factor. These data (up to five transcription factors) form the inset of Fig. 2A in the preceding chapter. (Incidence of regulatory architectures regarded by number of binding sites.)

5.4 Understanding the hypercubic graph: regulation by N transcription factors

The empirical distributions of known regulatory architectures for *E. coli* and *D. melanogaster* summarized in §5.2, §5.3, and the preceding chapter's Fig. 2A show that fully $\sim$ a quarter to a half of promoters or genes are reported to be regulated by the most common case of just one transcription factor. These scenarios are often plausibly described by a square graph of states and subject to the restrictions and behaviors we explore thoroughly in the preceding chapter. Yet clearly, promoters and genes can also be under the control of multiple transcription factors. The presence or absence of these transcription factors on the genome form hypercubic state spaces; some simple examples appear as annotations in Fig. 2A of the preceding chapter.

In this section, we derive the mathematical behavior of any response observable defined by these hypercubic networks, in or out of equilibrium. This derivation follows by understanding the strong structural constraints on the possible spanning trees in a hypercube. Following an illustration with small hypercubes and general arguments, Equations 5.6 and 5.8 report these generic nonequilibrium and equilibrium behaviors, respectively, thus fully generalizing Equations 1 and 3 of the preceding chapter to arbitrary combinatorial binding topologies. Obviously, such more complex network topologies can show more complex and higher-dimensional mathematical behaviors than simple regulatory cycles. However, their flexibility remains constrained, and later in §5.4.2 we also discuss how biophysical constraints—such as the common scenario of overlapping binding sites—can significantly simplify responses (see Figures 5.26 and 5.27).

As new experimental measurements become available, these equations can be used for quantitative fits to gene expression input-output curves, accommodating promoters known to be under the control of many transcription factors. As the contrast between nonequilibrium and equilibrium behaviors grows in higher dimension, these results give

tools to discern whether more complicated networks show signatures of operating out of equilibrium. These mathematical behaviors also prelude more specific investigations of attainable sensitivity, newly possible curve shapes, and specific access to interesting regulatory phenotypes, like those found for the square graph in later sections of the preceding chapter.

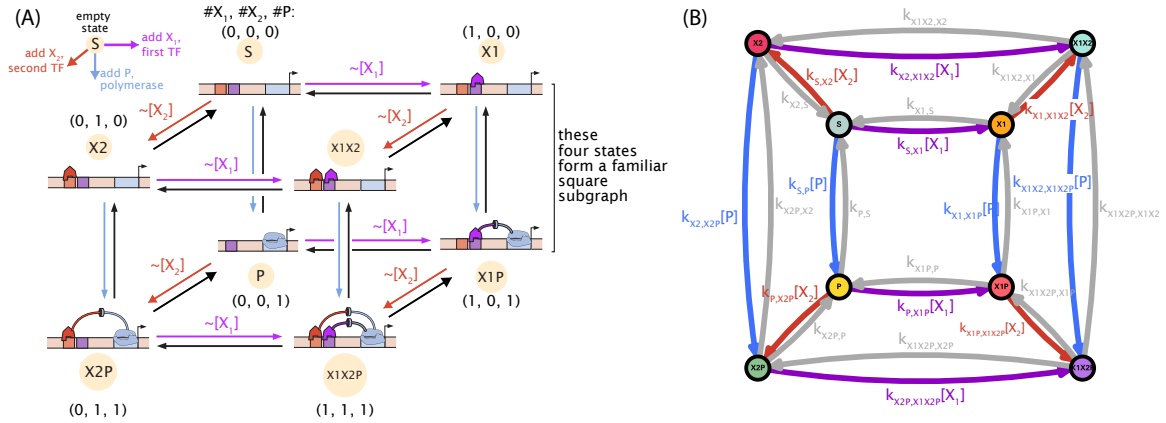


Figure 5.14: Regulation with two transcription factor binding sites: a cubic graph. (A): Allowing another transcription factor extends the state space from a square to a cube. In general, for m transcription factors and one polymerase, the state space will become a $n = (m + 1)$ dimensional hypercube, where the $2^n = 2^{m+1}$ vertices encode the presence or absence of each of the transcription factors or polymerase. This description assumes that every transcription factor is mutually compatible with the presence of other transcription factors; these transcription factors need not be distinct, if some proteins each have multiple binding sites, in which case some concentrations control multiple sets of orthogonal edges. Here we visualize the $m = 2$ transcription factors case, giving a $n = (2 + 1)$ -dimensional cube. Purple rightwards arrows $\rightarrow$ illustrate binding with transcription factor X_1 ; red leftwards arrows $\leftarrow$ denote binding with transcription factor X_2 ; and downwards vertical $\downarrow$ arrows denote binding with polymerase P . Each state is annotated with the appropriate binary code (a vector of length $n = (m + 1)$ whose i th entry is 1 if transcription factor i is present in that state, and 0 otherwise, with the convention that the final $(m + 1)$ th entry refers to the polymerase). We refer this code to help derive how algebraic responses collapse at detailed balance. (B): The same $n = 2 + 1$ hypercube-graph, visualized as a planar Tutte embedding, with rates associated labeled.

As a hint to the impending complexity, while we only had sixteen directed spanning trees on the square ($n = 2$ -cube), many more trees confront us for hypercube graphs in higher dimensions. In general, the number of undirected spanning trees on the n -cube is,

$$T_u(n) = \frac{1}{2^n} \prod_{i=1}^n (2i)^{\binom{n}{i}}. \quad (5.1)$$

This enumeration result follows from nontrivial combinatorial arguments [9] or from finding the eigenvalues of the Laplacian by analyzing the vertices of the hypercube as an Abelian group via representation theory [10, § 5.5, p. 62]. (These counts of the number of undirected trees on the n -cube are reported in the Online Encyclopedia of Integer Sequences as terms of the [sequence OEIS A006237](#).) Specifically, there are $T_u = 384$ undirected spanning trees on the 3-cube. Since any undirected spanning tree can be rooted at each vertex and generate a unique directed spanning tree, there are $N_{\text{vertices}} = 2^n$ rooted spanning trees per undirected spanning tree, giving

$$T_d = N_{\text{vertices}} T_u(n) = 2^n T_u(n) = \prod_{i=1}^n (2i)^{\binom{n}{i}}$$

directed spanning trees. Or, for $n = 3$, there are $T_d = 2^n T_u = 8 \times 384 = 3,072$ directed spanning trees on the 3-cube. This complexity explodes rapidly: the next largest cube, the tesseract ($n = 4$), has a whopping 42,467,328 undirected spanning trees (so 679,477,248 directed spanning trees), and so on. We plot this explosion of rooted spanning trees in Fig. 5.15.

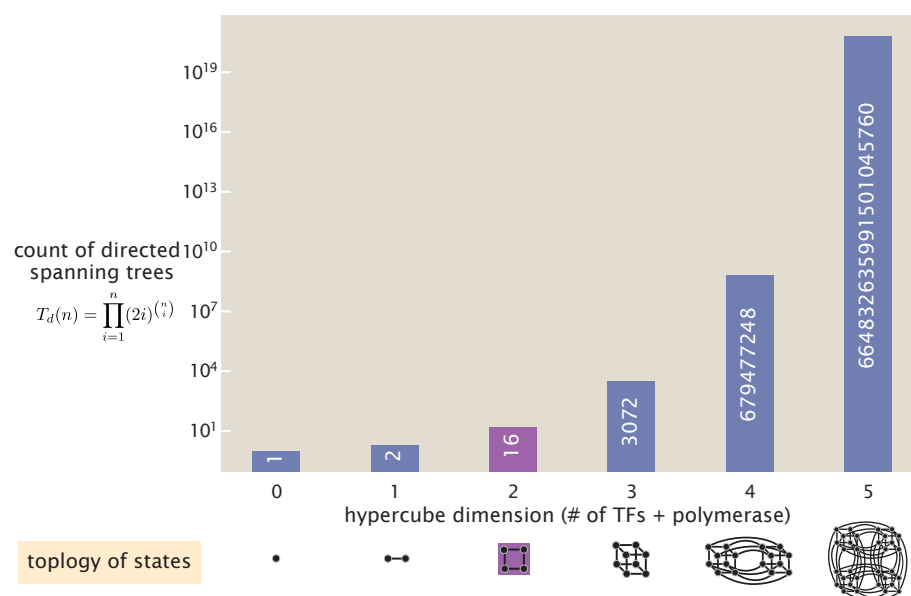


Figure 5.15: Superexponential explosion of the number in rooted spanning trees with respect to the hypercube dimension.

Notice that by the time the network is governed by only four transcription factors (forming a hypercube in five dimensions), the number of rooted spanning trees has exploded to over 664 quintillion. This number is over a “millimole” of spanning trees, a supercosmological number: if we managed to speedily sketch 1 spanning tree per second, it would take us $\sim 1500\times$ the current age of the universe to find all these trees. Clearly, brute force must be replaced by the foresight of structural reasoning about these trees.

5.4.1 Illustration with the 3-cube case

We first illustrate the consequences of multiple transcription factors with the tractable case of the $n = 3$ dimensional cube, capturing the impact of two transcription factors (call them X_1 and X_2) which can be present at up to one copy each, in addition to the polymerase. This architecture, where appropriate transitions scale with each of these control variables, is depicted in Figures 5.14.

To enumerate the spanning trees, we exploit the Matrix Tree Theorem: removing row and column i in the Laplacian matrix and taking the determinant gives, up to a sign, the sum of all spanning trees rooted in state i , a procedure amenable to symbolic computation. Figure 5.16 illustrates the 384 trees rooted in just the state S , where the node positions are the same as those in Fig. 5.14.

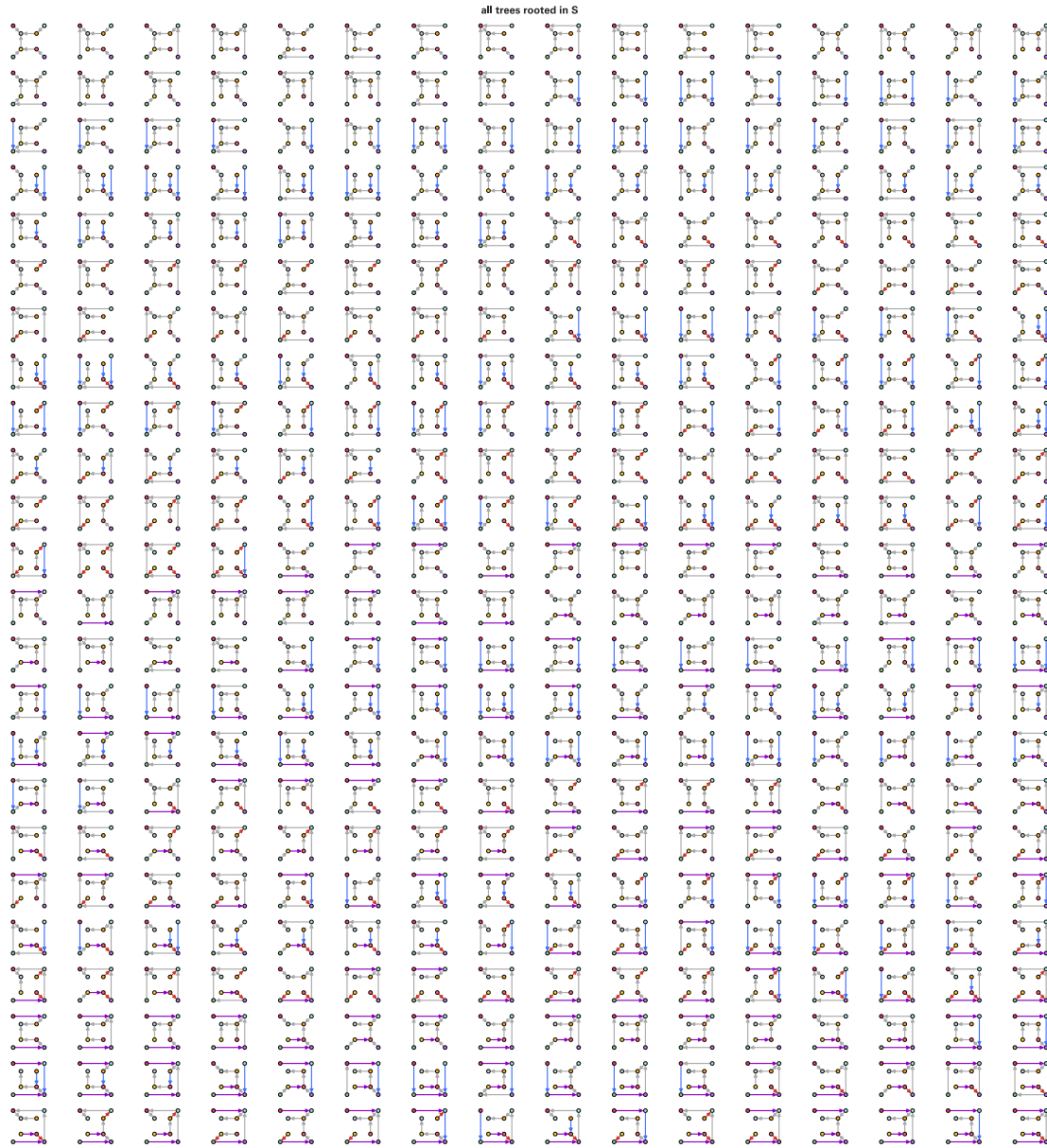


Figure 5.16: All directed spanning trees that are rooted in state S , computed by symbolic determinants, on the $n = 3$ cube graph of states. The powers of transcription factors $[X_1]$ and $[X_2]$ found in each spanning tree increases from the top left to the bottom right in this grid.

A subset of these trees can be organized into a two-variable analog of our main text's tabular figure 3B, showing how different root states adopt varied dependencies on powers of the control variable(s). We map examples of all attained mixed powers of the transcription factor concentrations for just one representative state, for illustration, in Figure 5.16. (These trees, being just those rooted in state S , are one eighth of all the trees; those rooted in the seven other states are shown in full in e.g. our Reference [1], SI.) Accumulated over states, these dependencies give ρ_{tot} , and therefore any response $\langle r \rangle$, up to a quartic dependence on each control variable, since it is possible to form rooted trees that use all four $[X_i]$ -controlled edges present in the graph. However, the graph's structure enforces that not all combinations of powers of the transcription factors are accessible, as we now discuss.

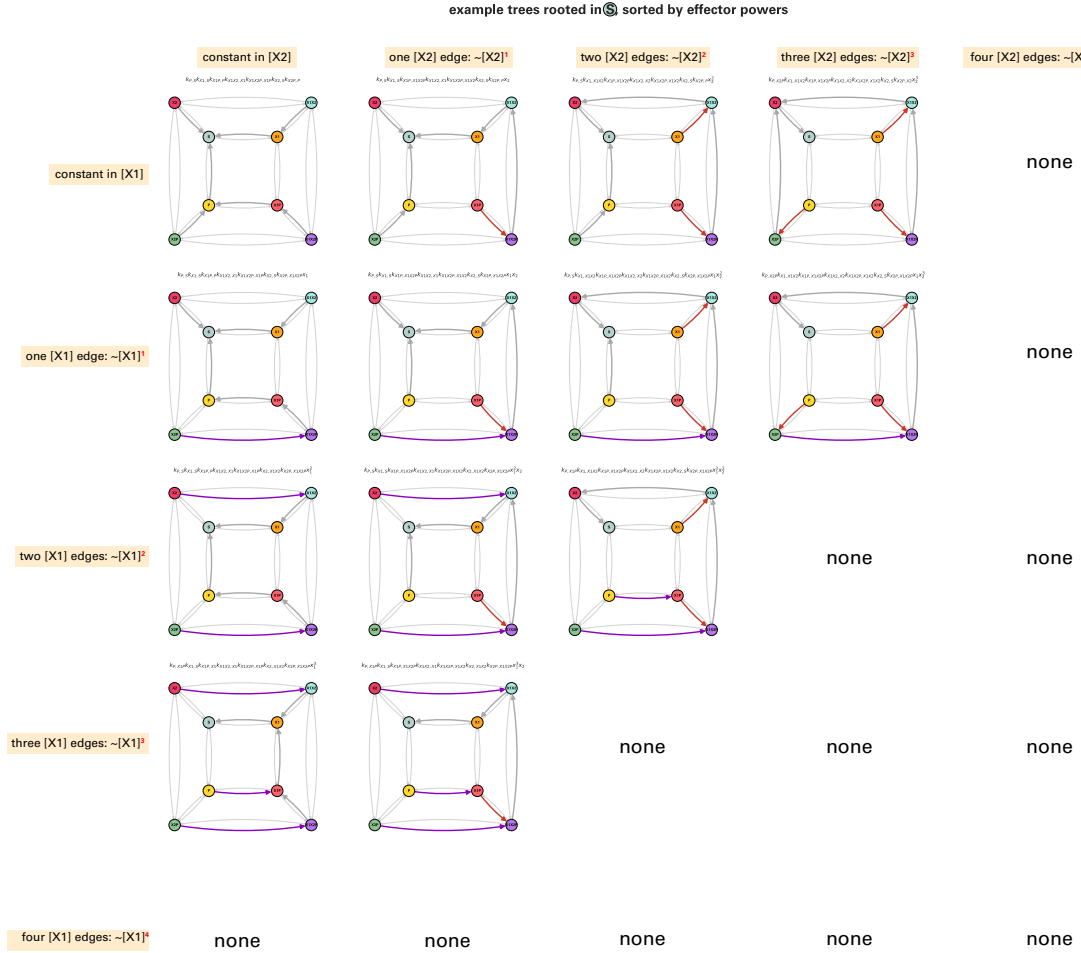


Figure 5.17: Examples of spanning trees rooted in the (empty) state S that contribute different combined powers of each transcription factor. Specifically, the entry at row m and column l of the grid exemplifies a spanning tree (when it exists) that has m edges dependent on transcription factor $[X_1]$ and l edges dependent on transcription factor $[X_2]$, contributing a monomial term $\sim [X_1]^m [X_2]^l$ to any response function. This establishes that, for some constants A_{ij} entirely determined by rate constants,

$$\rho_S = A_{00} + \sum_{i=1}^3 A_{i0} x_1^i + x_2 \sum_{i=0}^3 A_{i1} x_1^i + x_2^2 \sum_{i=0}^2 A_{i2} x_1^i + x_2^3 \sum_{i=0}^2 A_{i3} x_1^i. \quad (5.3)$$

(State P attains the same possible powers of $[X_1]$ and $[X_2]$ shown here.)

In summary, these allowed spanning tree structures over all roots give any response (in or out of detailed balance) for the $m = 2$ transcription factors case (two control variables $[X_1], [X_2]$) of the form,

$$\langle r \rangle = \frac{A_{00} + \sum_{i=1}^4 A_{i0}[X_1]^i + [X_2] \sum_{i=1}^4 A_{i1}[X_1]^i + [X_2]^2 \sum_{i=1}^4 A_{i2}[X_1]^i + [X_2]^3 \sum_{i=1}^3 A_{i3}[X_1]^i + [X_2]^4 \sum_{i=0}^2 A_{i4}[X_1]^i}{B_{00} + \sum_{i=1}^4 B_{i0}[X_1]^i + [X_2] \sum_{i=1}^4 B_{i1}[X_1]^i + [X_2]^2 \sum_{i=1}^4 B_{i2}[X_1]^i + [X_2]^3 \sum_{i=1}^3 B_{i3}[X_1]^i + [X_2]^4 \sum_{i=0}^2 B_{i4}[X_1]^i}. \quad (5.4)$$

(Note that while these terms are organized into groups by constant powers of $[X_2]$ for convenience, this ordering is arbitrary: this response is indeed symmetric under interchange of $[X_1]$ and $[X_2]$. The “democratic” structure of the hypercube graph enforces that no transcription factor is special.)

Slicing the response function when one transcription factor is fixed

Notice that if say the second transcription factor were held fixed at some concentration $[X_2] := [X_2]_0$, the response function is a ratio of quartic polynomials in the remaining transcription factor control parameter. Clearly this response, even with respect to just one control parameter, has acquired new flexibility. Whether a complex experimentally-observed response is attributable to a simple system operating out of equilibrium, or a more architecturally-complex system operating near equilibrium, will depend on specific biological details and separate experimental and biological context—namely, whether many transcription factors’ binding sites are indeed found in a promoter and whether these proteins are actually expressed. Nevertheless, such details are often known—or knowable—for genes and promoters of interest.

Full generalization to the hypercube in n dimensions: cubes within cubes

What explains which transcription factor powers are attained (or not) in this cube graph, and how does this structure generalize to an arbitrary number of transcription factors? To answer this question, it is instructive to consider how a spanning tree could maximize both the numbers of edges controlled by one transcription factor x_1 , and those controlled by another x_2 , simultaneously.

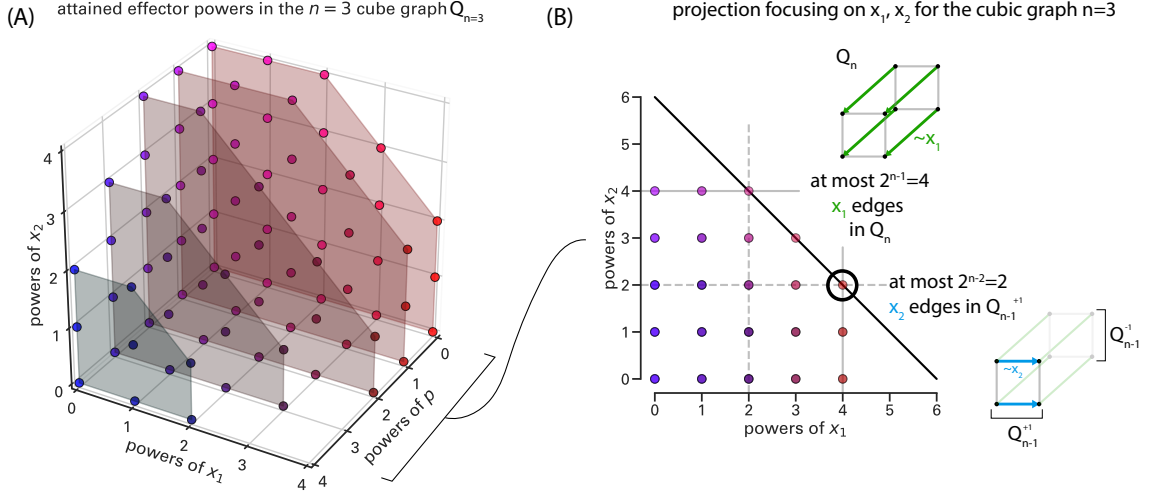


Figure 5.18: Powers attained by the transcription factors x_1 and x_2 , in addition to the polymerase p , over all spanning trees in the cubic $n = 3$ graph. (A) Powers of $x_1^i x_2^j p^k$ involved in the numerator and denominator of the output function (5.4). If there is a dot at position (i, j, k) in space, then there is a term proportional to $x_1^i x_2^j p^k$ in both numerator and denominator of (5.4). The color of a dot corresponds to a power of p . Pink dots are term with a contribution of order 0 in p , the blue ones of order 5 in p . The plotted planes with the corresponding colors are the planes are constant power of p . (B) Powers of $x_1^i x_2^j$ involved in the numerator and denominator of the output function (5.4), not taking into account the polymerase P . (Indeed, this concentration of polymerase is often taken to be constant in regulatory models.) Q_n represents the full hyper-cubic graph of dimension n ; in this example, $n = 3$. The edges having an x_1 dependency are represented in green. In a graph representing the binding of two transcription factors, there are at most 2^{n-1} edges depending on the concentration of one of the transcription factors. (In dimension n the hypercube represents the binding of two transcription factors and of the polymerase to the DNA, 2^{n-1} states don't have x_1 bound to the DNA and could acquire it, creating 2^{n-1} edges with a x_1 dependency). In dimension 3 the number of this type of edges is 4. The circled dot represents a spanning tree in which, these 4 edges are participating. As there are 6 nodes in a cube, two edges are still needed to create a spanning tree. The two edges presented in blue, depending on x_2 , complete the four green edges, depending on x_1 , in a spanning tree. In fact, to complement the 4 green edges into a spanning tree, the two edges have to be in Q_{n-1}^{+1} . Q_{n-1}^{+1} and Q_{n-1}^{-1} are the $n - 1$ dimensional cubes formed by states in which X_1 is respectively bound or unbound to the DNA and the edges connecting them. Indeed, the 2 edges have to depart from a node in which X_1 is bound to the DNA (included in Q_{n-1}^{+1} by definition), because the participation edges in Q_{n-1}^{-1} would create nodes with two outgoing edges. In Q_{n-1}^{+1} , there are up to 2^{n-2} edges depending on x_2 . Therefore, complementing the 4 green edges, with a x_1 linear dependency, by the two blue edges, with a x_2 linear dependency, creates a spanning tree with a quartic dependency in x_1 and quadratic in x_2 , corresponding to the circled dot.

Figure 5.18 summarizes the attainment of powers of x_1 , x_2 , and polymerase p , where the powers attained by x_1 and x_2 are emphasized by a projection over the attained polymerase powers at right. A tree on the full (hyper)cube can easily contain all of the possible edges driven by a particular transcription factor, say x_i : there are 2^{n-1} such x_i -driven edges in the whole graph (since half of the 2^n states in a full hypercube lack this particular transcription factor, and from each of these x_i -deprived states we count one x_i -driven edge), explaining how trees can attain powers of $2^{n-1} = 4$ in either x_1 or x_2 individually. (This boundary is visualized by the unbroken grey line in Fig. 5.18.) However, these trees that contain all of the possible x_i -driven edges cannot simultaneously contain all x_j -driven edges in the graph. This is because, by construction, all of the x_i -deprived states already have exactly one outgoing edge (precisely to assure each of the x_i -driven edges belong to the tree), so the remaining edges that build the spanning tree must be found wholly among the half of the graph's states where x_i is *present*. This restriction halves the number of possible x_j edges that can now participate in the same tree.

In other words, if Q_n represents the full n -dimensional hypercube, then it can always be partitioned into two sets of states, one set deprived of transcription factor x_i and the other enriched by x_i . (In these sets of states, we also consider the interconnecting edges among them.) These sets of states and edges each form an $(n - 1)$ -dimensional hypercube; call Q_{n-1}^{+i} the subcube where X_i is present, and Q_{n-1}^{-i} the subcube where X_i is absent. The full n -dimensional hypercube is formed by joining these smaller hypercubes together by the perfect matching formed by the set of edges describing the binding of transcription factor X_i . Inside each smaller hypercube Q_{n-1} , there are only 2^{n-2} possible edges driven by the other transcription factor X_j ; and demanding that all of the X_i edges appear in a tree restricts the remaining edges to be taken from the Q_{n-1}^{+i} cube of states (where x_i is present). This scenario is illustrated by the example of the power combination circled in black in Fig. 5.18.

If we relax our demand that *all* 2^{n-1} edges of some transcription factor X_i participate in

a tree by allowing one fewer x_i -driven edge to appear, we can find a tree that now includes up to one extra power of another transcription factor X_j ; this extra power will come from one of the edges in the Q_{n-1} formed by the X_i -deprived states (formerly forbidden from contributing to the powers of x_j that could participate). Continuing to remove X_i edges continues to add new possible x_j edges in valid spanning trees, up to the 2^{n-1} allowed (by symmetric reasoning). This explains the diagonal boundary of attained powers marked by a black line with slope -1 in the right of Fig. 5.18; in general, this maximal boundary is a hyperplane assuring that the total powers of transcription factors (and polymerase) is at most

$$\# \text{ of total regulated edges in any tree (hence powers in response)} \leq \sum_{k=0}^n 2^{n-k} = 2^n - 1. \quad (5.5)$$

For instance, Figure 5.19 visualizes this planar constraint that powers of $[X_1]$, $[X_2]$, and P must ultimately sum to less than or equal to $2^3 - 1 =$ seven regulated edges in the $n = 3$ -cube.

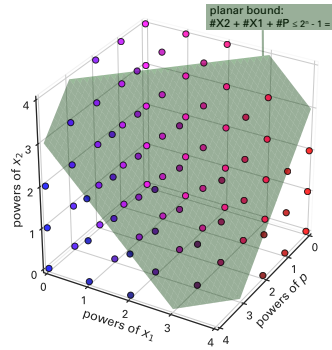


Figure 5.19: Visualization of how the plane $\#X_2 + \#X_1 + \#P \leq 2^n - 1 = 7$ (plotted in green) bound the attainable powers in any tree (hence response), in the $n = 3$ -cube case. $\#X_2$, $\#X_1$ and $\#P$ respectively correspond to exponent of x_2 , x_1 and p . Powers of $x_1^i x_2^j p^k$ involved in the numerator and denominator of the output function (5.4). If there is a dot in the (i, j, k) space, then there is a term proportional to $x_1^i x_2^j p^k$ in both numerator and denominator of (5.4). The color of a dot corresponds to a power of p . Pink dots are term with a contribution of order 0 in p , the blue ones of order 5 in p .

Furthermore, this logic applies recursively to each *subsequent* transcription factor $x_k \notin \{x_i, x_j\}$ —for instance, this structure gives trees of maximal power $x_i^{2^{n-1}} x_j^{2^{n-2}} x_k^{2^{n-3}} \cdots$. This reasoning is agnostic to the dimensionality of the state space (total number of transcription factors). Since the maximal power over all transcription factors is at most 2^{n-1} , the second highest power over transcription factors is 2^{n-2} , and so on, we use the notation of order statistics to write the general form. Recall that the i th *order statistic* $y_{(i)}$ of a set of values $\{Y_1, Y_2, \dots, Y_n\}$ refers to the value that occupies the i th position when the set of values is arranged in ascending order. In other words, it represents the i th smallest value in the set, so the i th largest value of a set is $y_{(n-i)}$. Using this notation, the generic response of any observable (in or out of detailed balance) formed by the hypercubic graph from n transcription factors can be written as,

$$\langle r \rangle = \frac{\sum_{\substack{\max\{i_l\}=i_{(n)} \leq 2^{n-1}, \ i_{(n-1)} \leq 2^{n-2}, \\ \dots \ i_{(1)} = \min\{i_l\} \leq 2^0}} A_{\{i_l\}} \prod_{i=1}^m [X_i]^{i_l}}{\sum_{\substack{\max\{i_l\}=i_{(n)} \leq 2^{n-1}, \ i_{(n-1)} \leq 2^{n-2}, \\ \dots \ i_{(1)} = \min\{i_l\} \leq 2^0}} B_{\{i_l\}} \prod_{i=1}^m [X_i]^{i_l}}, \quad (5.6)$$

where we have temporarily adopted the convention that one of the “transcription factor”s (say the last one $[X_n]$) represents the polymerase. (Note that this response form obeys the interchange symmetry required of all the transcription factors.)

Further illustration with $n = 4$ tesseract graph

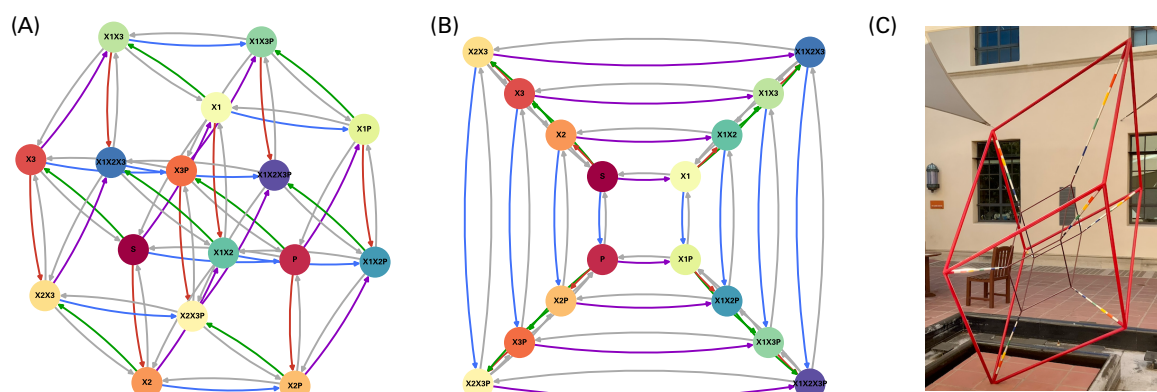


Figure 5.20: Adding a third transcription factor X_3 forms a hypercubic state space in $n = 4$ dimensions, known as a tesseract, represented in a concentric layout in panel (A). Panel (B) is a sculpture showing another popular arrangement of nodes in the tesseract, from artist Ronald Linde, pictured at Caltech.

Explicit enumeration of all 679,477,248 *directed* spanning trees for the 4-cube (tesseract) graph (giving the state space of three transcription factors acting in concert with the polymerase) is not very computationally tractable while preserving all symbolic dependencies on the 32 rate constants associated with the graph. However, we restrict our ambition to discerning just the algebraic form of the x_i dependencies, by replacing all of the rate constant prefactors by 1 but preserving the x_i -dependencies on the rates. This simplifies the symbolic computation of the determinant substantially, enough that *Mathematica* (but still no longer *Python* via *sympy* on consumer computing hardware) can now perform the computation, and reveal the powers attained for all transcription factors x_i .

We visualize the powers attained in this graph in Figure 5.21. These terms are consistent with the structure of the generic response form we indicate in Eq. 5.6.

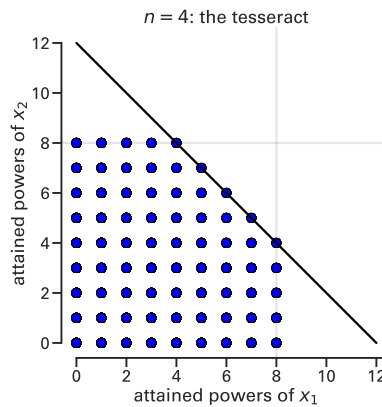


Figure 5.21: Powers of $x_1^i x_2^j$ involved in the numerator and denominator of the output function of a tesseract graph (hypercube for $n = 4$), representing a regulation unit with three transcription factor binding sites and depicted in Figure 5.20. x_2 and x_1 are the respective concentrations of two transcription factor. If there is a dot in the (i, j) plane, then there is a term proportional to $x_1^i x_2^j$ in both numerator and denominator of the output function, at constant x_3 and p (concentration of the third transcription factor and the polymerase). In this graph there are $\sim 6.8 \times 10^8$ spanning trees dependent on x_1, x_2, x_3, p .

Collapse of the output function at detailed balance

Earlier, in the square graph ($n = 2$ -cube), we originally arrived at the response's demotion to a ratio of linear polynomials at detailed balance by explicitly using the cycle condition (namely, that the product of rate constants in one direction around the cycle equals that around the opposite direction at equilibrium). However, now for the $n = 3$ -cube, we have no fewer than twelve coupled cycles, complicating a direct translation of our earlier tact to such more complicated graphs. Instead of explicit appeals to cycle conditions, we use the alternative structural reasoning we discuss in Reference [1], SI.

By emphasizing the powers of control parameter accumulated along paths between states that are disconnected, this reasoning helps us understand how the $n = 3$ -cube demotes its response at detailed balance. As before, we are essentially interested in expressing a given state j 's statistical weight ρ_j relative to the statistical weight of a reference state (call it P again without loss of generality). The state P is directly connected to states $\{S, X_1P, X_2P\}$, so these terms contribute at most a pure linear power $[X_1]$ or $[X_2]$ to any responses. This leaves the remaining, disconnected states: $\{X_2, X_1, X_1X_2P\}$ are two edges away from P in the graph, and state X_1X_2 is a full graph diameter of three edges away from P . While it is possible to create paths between P and these other states that involve two edges dependent on an individual control parameter, actually this does *not* imply that the response is demoted to only a quadratic polynomial in that parameter, since all of these paths also must contain edges scaling as the inverse of that control parameter, canceling at least one power of this dependence.

This cancellation immediately becomes more clear when we return to the binary encoding of states we mentioned in Fig. 5.14. Every state in the hypercube graph, no matter how many edges apart, differs from any other state by at most a negation in each of its binary digits. When comparing two states in the graph, a negation in a particular binary coordinate i is exactly a difference in whether or not transcription factor i is present or

absent between the two states. Since every edge at most changes one binary coordinate at a time, this just says that the net number of binding versus unbinding events of transcription factor i along any path connecting two states is at most the maximum binary distance along a binary coordinate possible allowed by the graph, namely one. Summarizing in other words, even if two states are connected by a long path of possibly $[X_i]$ -dependent arrows, the simple state space requires that every such path ultimately contribute at most a factor $[X_i]^{\pm 1}$ to the ratio of the statistical weights of the two states.

To wit, the response of a hypercube graph formed by the interaction of N transcription factors $\{X_i\}_{i=1:N}$ must ultimately collapse at equilibrium to the dramatically simplified multilinear form,

$$\langle r \rangle^{\text{eq.}} = \frac{\sum_{z_1, z_2, z_3, \dots, z_N \in \{0,1\}} \alpha_{z_1, z_2, z_3, \dots, z_N} [X_1]^{z_1} [X_2]^{z_2} [X_3]^{z_3} \dots [X_n]^{z_n}}{\sum_{z_1, z_2, z_3, \dots, z_N \in \{0,1\}} \beta_{z_1, z_2, z_3, \dots, z_N} [X_1]^{z_1} [X_2]^{z_2} [X_3]^{z_3} \dots [X_n]^{z_n}} \quad (5.7)$$

$$= \frac{\sum_{z_1, z_2, z_3, \dots, z_N \in \{0,1\}} \alpha_{z_1, z_2, z_3, \dots, z_N} \prod_{i=1}^n [X_i]^{z_i}}{\sum_{z_1, z_2, z_3, \dots, z_N \in \{0,1\}} \beta_{z_1, z_2, z_3, \dots, z_N} \prod_{i=1}^n [X_i]^{z_i}}. \quad (5.8)$$

For our $n = 3$ (two-transcription factor) example, the detailed balance response has the form,

$$\langle r \rangle^{\text{eq., } m=2} = \frac{\alpha_{111}[X_1][X_2]p + \alpha_{110}[X_1][X_2] + \alpha_{100}[X_1] + \alpha_{010}[X_2] + \alpha_{001}p + \alpha_{000}}{\beta_{111}[X_1][X_2]p + \beta_{110}[X_1][X_2] + \beta_{100}[X_1] + \beta_{010}[X_2] + \beta_{001}p + \beta_{000}}. \quad (5.9)$$

In this example, if e.g. $[X_2]$ is held at some constant, then this form collapses dramatically all the way back down to a ratio of linear polynomials in $[X_1]$. This collapse is still true in n dimensions: if all but a specific transcription factor's concentration $[X_i]$ were held fixed,

then the functional form with just this non-fixed control variable is still just a ratio of linear polynomials in $[X_i]$.

In summary, we have shown how the response function attains a restricted subset of possible combinatorial powers of control variables, in or out of equilibrium, on the n -cubic graph. We graphically summarize the algebraic constraints on these responses in Figure 5.22, which shows the attainable powers of any two control variables among the full n -space.

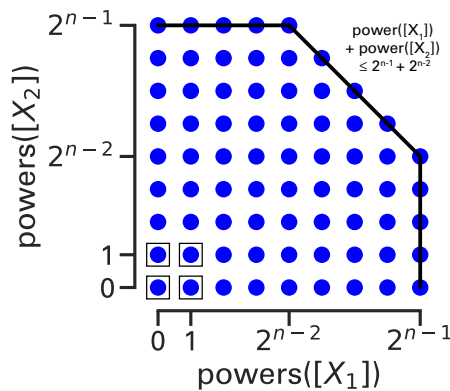


Figure 5.22: Graphical summary of how any two control variables $[X_1]$ and $[X_2]$ participate algebraically in the nonequilibrium or equilibrium response on an n -cubic graph. If a blue dot appears at position (i, j) , then a monomial of the form $[X_1]^i[X_2]^j$ participates in the numerator and denominator of the algebraic form of the response $\langle r \rangle$. The dots outlined by the black boxes are the only powers attainable when the response is at equilibrium. The boundary of the permitted jointly-attained powers is a result of the structural reasoning explained in this section.

Accommodating multiple binding sites for a transcription factor

The general response function for the hypercubic graph, derived in Eq. 5.6 and Eq. 5.8, accommodates equally well the cases where all transcription factors bound to the promoter are molecularly distinct, and the cases where there are multiple copies of a given transcription factor bound to the promoter. For instance, if one transcription factor X_j has two binding sites (and thus can be found at a copy number of zero, one, or two copies on

the promoter), against a backdrop of m other transcription factors with single binding sites, then the response function is given by Eq. 5.6 and Eq. 5.8 for a $(m + 2)$ -dimensional hypercube, except that two concentrations appearing in these response formulas have been set to the same value $[X_j]$.

5.4.2 Common graphs formed by subsets of the hypercube

N strictly exclusive transcription factors: the book/rolodex graph

Often, the binding sites for transcription factors overlap. For instance, four promoters with appreciably-overlapping binding sites in *E. coli* are visualized in Fig. 5.23. These common overlaps suggest that the presence of transcription factors on a promoter might exclude others from also being bound.

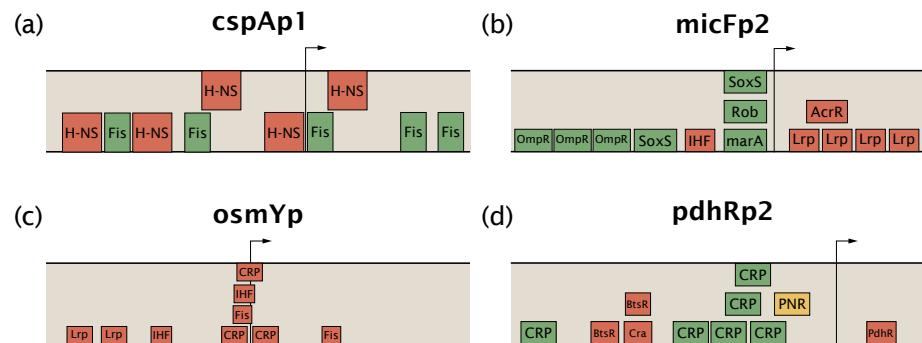


Figure 5.23: Four examples showing how *E. coli* promoters regulated by multiple transcription factors often present extensive overlaps between binding sites. (A) Regulation of the promoter *cspAp1*. (B) Regulation of the promoter *micFp2*. (C) Regulation of the promoter *osmYp*. (D) Regulation of the promoter *pdhRp2*. Binding sites are annotated by their name and colored in green, red, or yellow if the sites are understood to be activating, repressing, or both, respectively. The transcription starting site is indicated by the black arrow. (Note that these schematics reflect the topology/overlaps and relative positions of binding sites, but the size of each binding site's box is not to scale in genomic coordinates.)

If some transcription factors are mutually exclusive with each other and/or the polymerase, the hypercubic graph is pruned to a smaller state space, significantly changing the observed response functional form of the response with respect to the control variables.

Consider the highly biologically plausible case where two transcription factors X_1 and X_2 cannot coexist with each other, but each can coexist with the polymerase. The trees that span the resulting graph have particularly simple structure and give correspondingly simple forms for the ultimate response function of the system.

fig 4a from Lammers, Flamholz, Garcia 2023

<https://doi.org/10.1073/pnas.2211203120>

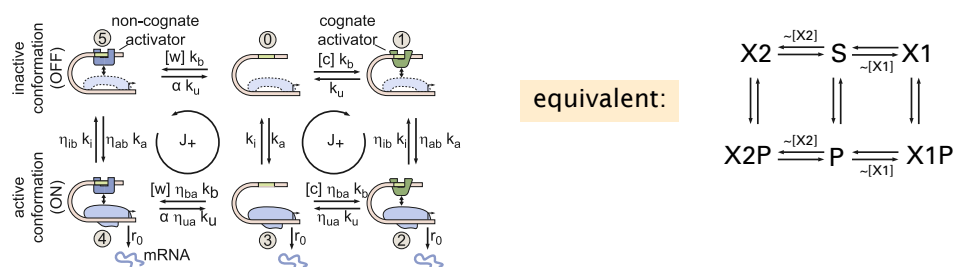


Figure 5.24: Regulatory motif with two transcription factors, whose binding is mutually exclusive, from Reference [11]. When only one transcription factor can bind at a time, but both can bind in the presence of the polymerase, a smaller state space graph is formed. One biological scenario yielding this simple exclusion structure is the setting studied by Reference [11], where both cognate and noncognate enhancers modulate the action of polymerase in cellular decision making.

For instance, consider the biological example invoked by Fig 4A of Lammers, Flamholz and Garcia [11], which they use to model interference by noncognate factors in cellular decision making. The state space collapses from a three-dimensional cubic graph (with eight conjoined cycles) to just two conjoined square cycles, as shown in Fig. 5.24.

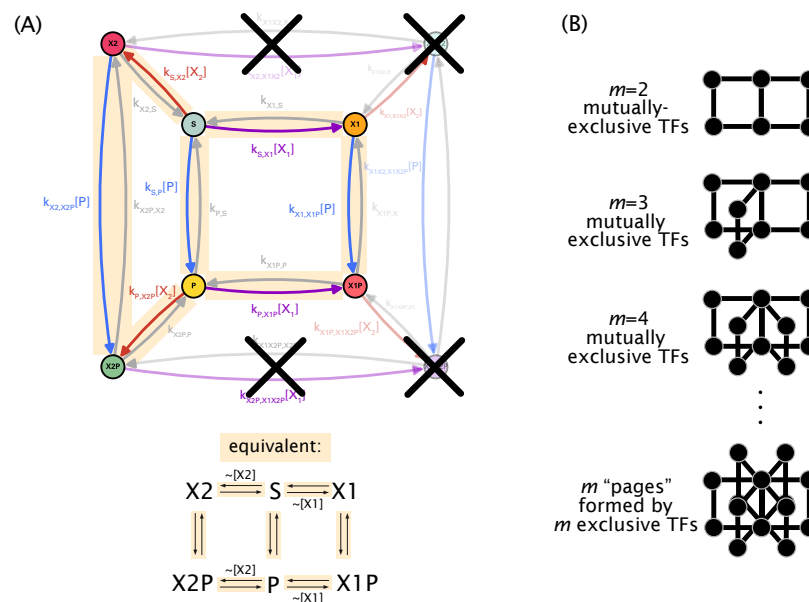


Figure 5.25: Relation between the cubic graph and the “book graph” and a generalisation of this “book graph” in m dimensions corresponding to the binding of m exclusive transcription factors. (A) Schematic highlighting the set of states (outlined in tan) of two mutually-exclusive transcription factors found inside a full hypercube graph (subject to no exclusions). Forbidden interactions and states are crossed out in black. (B) When multiple transcription factors govern a network and all are mutually exclusive (except with the polymerase), the resulting graph is always a set of m square cycle “pages,” each associated with a given transcription factor, conjoined along one central “spine”/edge-pair between the empty S and polymerase-bound P states.

Figure 5.25 localizes the resulting states of this example as a subgraph of the hypercube graph we just investigated in §5.4. This graph also achieves analogous structures in higher dimension (with a larger number of transcription factors, still mutually exclusive except for the polymerase): extra “pages” are added around the “axis” of the $S \leftrightarrow P$ transition, creating a structure rather like a cylindrical rolodex. Each of the independent square pages is a cycle between the empty state S , the polymerase-bound state P , plus the states X_j and X_jP where a particular transcription factor X_j is bound alone or bound along with the polymerase, respectively; the exclusions among transcription factors enforce that these cycles are joined only along the central axis. It transpires that some literature refers to this general simple graph structure as a “book graph” (see e.g. [Wolfram MathWorld](#)).

Another important, biologically-distinct example where this regulatory architecture manifests is the setting of post-translational modifications, such as phosphorylation, made to a standing pool of transcription factors. For instance, if both unphosphorylated and phosphorylated forms of a transcription factor bind (perhaps with distinct affinity) to the same site on the genome, then these species are effectively mutually-exclusive distinct transcription factors, and will produce a book graph like those visualized in Figure 5.25B.

General response function of the book graph from multiple exclusive transcription factors

Having understood that multiple transcription factors form a book graph when they are mutually exclusive, we can deduce the form of any resulting response function. The crucial point is that the states S and P are the only states in the graph which have any outgoing edges dependent on transcription factor concentrations. We conclude that in any spanning tree, at most one outgoing edge can depend on a transcription factor (say X_i) outgoing from S , and at most one edge can depend on a transcription factor (say X_j , possibly with $i = j$) from P . This structural constraint immediately requires that any tree (hence any response function)

can only attain exactly the following dependencies on transcription factor concentrations: $\sim [X_i]^2, \sim [X_j]^2, \sim [X_i][X_j], \sim [X_i]$, or $\sim [X_j]$, for all pairs of distinct transcription factors (i, j) . Accordingly, any response function of a book graph always exhibits the remarkably simple form²,

$$\langle r \rangle^{\text{book}} = \frac{a_0 + \sum_{i=1}^m (a_{i,1}[X_i] + a_{i,2}[X_i]^2) + \sum_{i \neq j \leq m} b_{ij}[X_i][X_j]}{c_0 + \sum_{i=1}^m (c_{i,1}[X_i] + c_{i,2}[X_i]^2) + \sum_{i \neq j \leq m} d_{ij}[X_i][X_j]}. \quad (5.10)$$

Importantly, observe that if only one of the transcription factor concentrations $[X_i]$ is taken to vary (namely, regarding the function above from a hyperplanar slice with all but one transcription factors fixed), then this form collapses all the way down to a ratio of quadratic polynomials in this concentration, exactly the mathematical form studied extensively in preceding chapter.

At equilibrium, detailed balance enforces that every ratio of statistical weights has at most one power of a transcription factor's concentration. This collapse can be understood by the same means as the collapses found for the hypercube, or DNA looping graphs, analyzed earlier: focusing on the possible paths between any pair of states, we see that at most one

²In this work, we have usually absorbed the response function's dependence on polymerase concentrations into coefficients since we regard polymerase as a completely separate (and plausibly much more constrained) biological control variable. However, here it is structurally amusing to notice that the $[X_i]^2$ -dependent terms of the response function present all possible powers of $[P]$ from $[P]^0$ (constant) to $[P]^m$. Why? Each tree delivering a quadratic dependence on some transcription factor $\sim [X_i]^2$ cannot accumulate any further dependencies on other transcription factor concentrations, nor can the $S \rightarrow P$ edge participate in the tree (since both S and P have "spent" their possible outgoing edges exactly to achieve the $\sim [X_i]^2$ dependence). However, every page of the book graph formed by the transcription factors can easily either include a $[P]$ -dependent edge (going from $X_j \rightarrow X_j P$) or not (then involving the unbinding edge $X_j P \rightarrow P$), and still span the graph's states. In addition, if we relax the requirement that $[X_i]^2$ appear, a separate $[P]^{m+1}$ dependency can be attained by taking all $X_j \rightarrow X_j P$ edges with the $S \rightarrow P$ (all downward edges) and including up to one $[X_i]$ -dependent edge, where the tree is completed by all other radial edges being oriented inwards as unbinding edges, $X_j P \rightarrow P$. We propose these structural considerations could be useful for other scenarios where an analog of the book graph appears and the variable P , with which all other transcription factor control variables are compatible, is perhaps instead understood as itself a separate transcription factor.

power of any given transcription factor's concentration $[X_i]$ is accumulated along any simple path between any two states in the graph. Thus any response of a book graph is a linear rational function of transcription factor concentrations at equilibrium, such that Eq. 5.10 collapses to,

$$\langle r \rangle^{\text{book, eq.}} = \frac{a_0 + \sum_{i=1}^m a_i [X_i]}{b_0 + \sum_{i=1}^m b_i [X_i]}. \quad (5.11)$$

5.4.3 All nontrivial, transcriptionally-potent subgraphs of the $n = 4$ hypercube

While we just analyzed the limiting case where all transcription factors are mutually exclusive, more relaxed and complex exclusion topologies (where only some transcription factors or polymerase are incompatible with each other on the genome) are also plausible. In this section, we systematically find the architectures (and associated response functions) of all possible networks with up to three transcription factors. We find these networks by beginning with the tesseract ($n = 3 + 1$ dimensional hypercube) graph and removing all states formed by possible individual, pairwise, triplet, or quadruplet mutual exclusions among the three transcription factors and polymerase. (For instance, the book graph is generated by forbidding states where any pair of transcription factors are present, here specifically excluding states $X_1 X_2$, $X_2 X_3$, and $X_1 X_3$ and their accompanying transitions.) Further, we require that the resulting graphs remain transcriptionally potent (e.g. retain at least one state where polymerase can be bound to the genome) and discard redundant (e.g. literally isomorphic) networks generated during this enumeration. Figs. 5.26 and 5.27 summarize the results of this exercise.

This exercise reveals that a large fraction of networks constrained by exclusions show simple mathematical dependencies in at least one of their transcription factors.

Specifically, the majority (90 out of the total nontrivial 147 networks) show at most a quadratic dependence in at least one transcription factor; such networks are shaded in purple in Figs. 5.26 and 5.27. This quadratic dependence is the same as that analyzed thoroughly in the preceding chapter; when the simply-varying transcription factor is the only transcription factor that varies, the resulting response function is subject to all the precise and quantitative mathematical constraints born by the square graph. The result of this exercise is also summarized in Figure 2B of the preceding chapter.

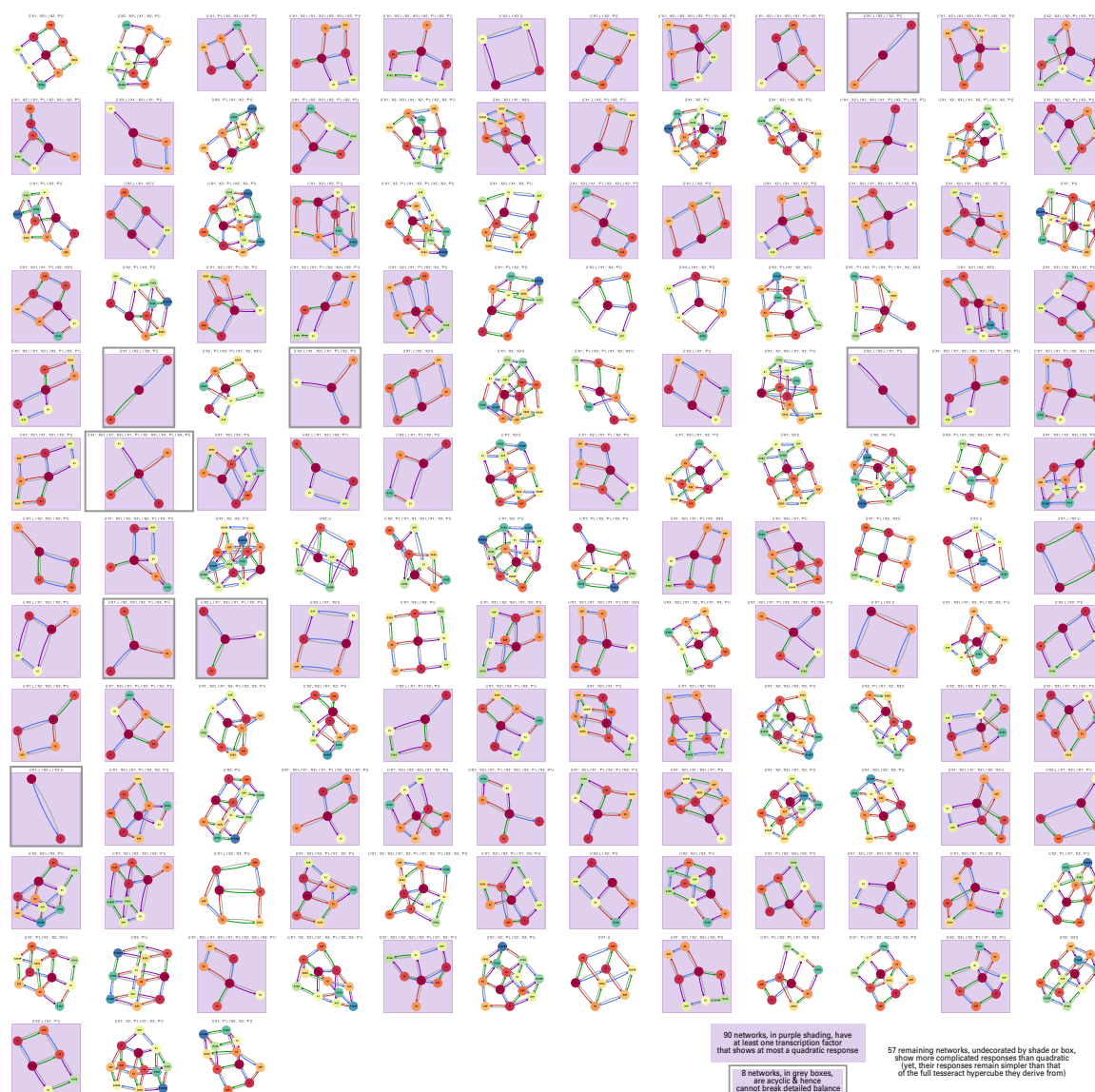


Figure 5.26: Comprehensive census of all subsets of the tesseract cube formed by all possible pairwise, triplet, or individual exclusions among the transcription factors and polymerase, restricted to the set where polymerase can still be present. Each panel depicts the state topology of a distinct possible set of mutual exclusions between transcription factors (among themselves) and/or polymerase, as long as the polymerase can be found bound to the genome in at least one state. (The ordering of panels follows an arbitrary enumeration of while maintaining that the polymerase can bind in at least one state.) Purple shading emphasizes those networks that display at least one transcription factor whose response varies as a quadratic response or simpler (namely, matching the mathematical response of a square graph).

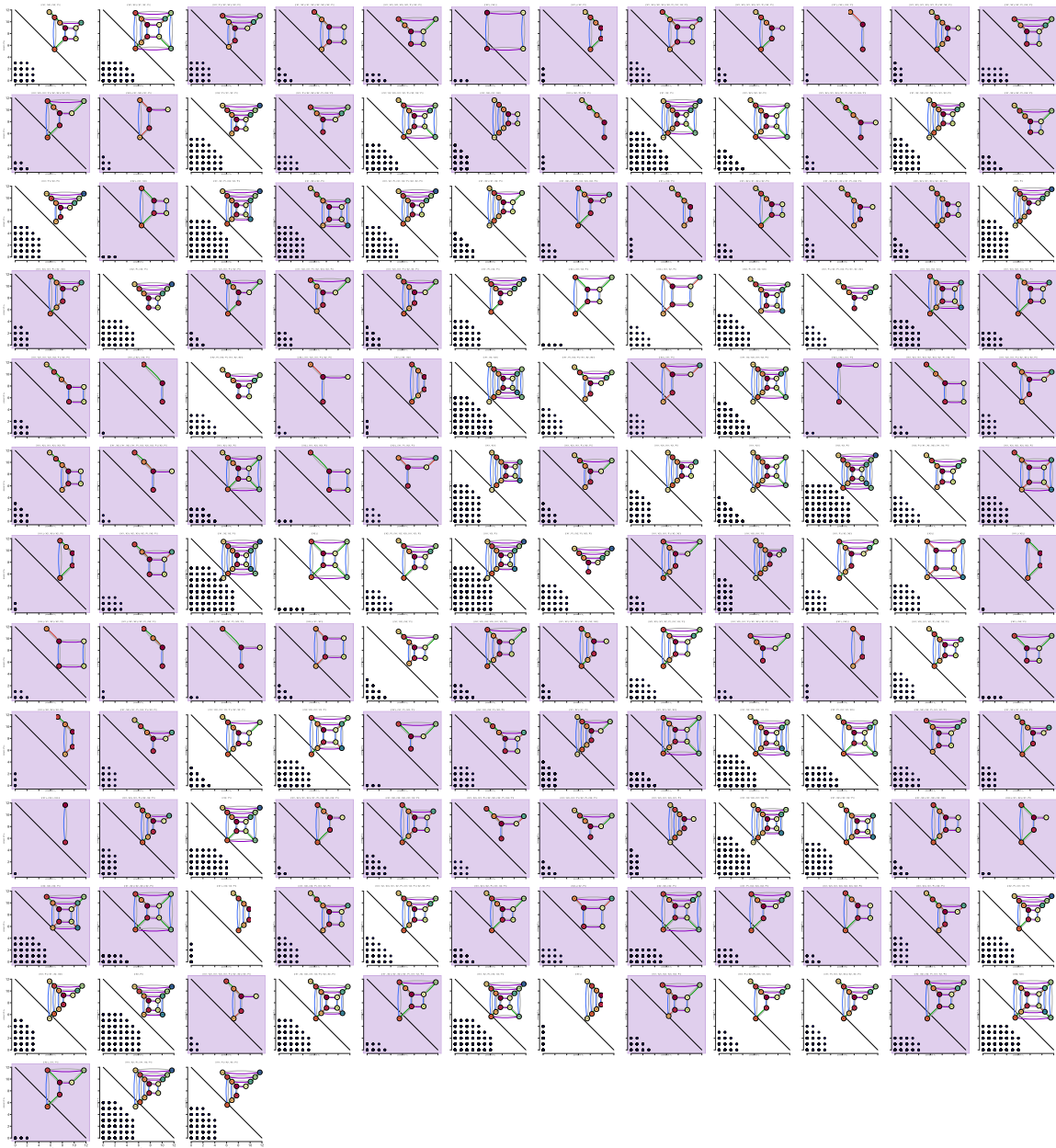


Figure 5.27: The same three-transcription-factor-networks identified in Fig. 5.26, now visualized alongside their attained powers in two of their transcription factor concentrations, $[X_1]$, $[X_2]$. The networks (and their ordering) are the same as those in the earlier Fig. 5.26; the nodes are just arranged according to a different layout to emphasize their nature as a subgraph of the tesseract graph, whose full set of nodes is shown in Fig. 5.20B.

Bibliography

- [1] Sara D. Mahdavi, Gabriel L. Salmon, Patill Daghljan, Hernan G. Garcia, and Rob Phillips. Flexibility and sensitivity in gene regulation out of equilibrium. *Proceedings of the National Academy of Sciences*, 121(46):e2411395121, 2024.
- [2] Lacramioara Bintu, Nicolas E Buchler, Hernan G Garcia, Ulrich Gerland, Terence Hwa, Jané Kondev, and Rob Phillips. Transcriptional regulation by the numbers: models. *Curr. Opin. Genet. Dev.*, 15(2):116–124, 2005.
- [3] Víctor H Tierrafría, Claire Rioualen, Heladia Salgado, Paloma Lara, Socorro Gama-Castro, Patrick Lally, Laura Gómez-Romero, Pablo Peña-Loredo, Andrés G López-Almazo, Gabriel Alarcón-Carranza, et al. Regulondb 11.0: Comprehensive high-throughput datasets on transcriptional regulation in *Escherichia coli* K-12. *Microb. Genom.*, 8(5), 2022.
- [4] Mark EJ Newman. Mixing patterns in networks. *Physical review E*, 67(2):026126, 2003.
- [5] William T Ireland, Suzannah M Beeler, Emanuel Flores-Bautista, Nicholas S McCarty, Tom Röschinger, Nathan M Belliveau, Michael J Sweredoski, Annie Moradian, Justin B Kinney, and Rob Phillips. Deciphering the regulatory genome of *Escherichia coli*, one hundred promoters at a time. *eLife*, 9:e55308, 2020.
- [6] Mattias Rydenfelt, Hernan G Garcia, Robert Sidney Cox III, and Rob Phillips. The influence of promoter architectures and regulatory motifs on gene expression in *Escherichia coli*. *PLoS One*, 9(12):e114347, 2014.
- [7] Muir Morrison, Manuel Razo-Mejia, and Rob Phillips. Reconciling kinetic and thermodynamic models of bacterial transcription. *PLoS Comput. Biol.*, 17(1):e1008572, 2021.
- [8] Soile VE Keränen, Angel Villahoz-Baletta, Andrew E Bruno, and Marc S Halfon. REDfly: an integrated knowledgebase for insect regulatory genomics. *Insects*, 13(7):618, 2022.
- [9] Olivier Bernardi. On the spanning trees of the hypercube and other products of graphs. *arXiv preprint arXiv:1207.0896*, 2012.
- [10] Richard Stanley. *Enumerative Combinatorics: Volume 2*. Cambridge University Press, 1999.

- [11] Nicholas C Lammers, Avi I Flamholz, and Hernan G Garcia. Competing constraints shape the nonequilibrium limits of cellular decision-making. *Proc. Natl. Acad. Sci. U.S.A.*, 120(10): e2211203120, 2023.