A mathematical model clarifies the ABC Score formula used in enhancer-gene prediction

Joseph Nasser[1,2,*], Kee-Myoung Nam[1,3], Jeremy Gunawardena[1,*]

[1]Department of Systems Biology, Harvard Medical School, Boston, MA, USA
[2]Current address: Department of Physics, Brandeis University, Waltham, MA, USA
[3]Current address: Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT, USA
[*]To whom correspondence may be addressed: Joseph Nasser (joseph.nasser21@gmail.com) and Jeremy Gunawardena (jeremy@hms.harvard.edu)

Abstract

Enhancers are discrete DNA elements that regulate the expression of eukaryotic genes. They are important not only for their regulatory function, but also as loci that are frequently associated with disease traits. Despite their significance, our conceptual understanding of how enhancers work remains limited. CRISPR-interference methods have recently provided the means to systematically screen for enhancers in cell culture, from which a formula for predicting whether an enhancer regulates a gene, the Activity-by-Contact (ABC) Score, has emerged and has been widely adopted. While useful as a binary classifier, it is less effective at predicting the quantitative effect of an enhancer on gene expression. It is also unclear how the algebraic form of the ABC Score arises from the underlying molecular mechanisms and what assumptions are needed for it to hold. Here, we use the graph-theoretic linear framework, previously introduced to analyze gene regulation, to formulate the *default model*, a mathematical model of how multiple enhancers independently regulate a gene. We show that the algebraic form of the ABC Score arises from this model. However, the default model assumptions also imply that enhancers act additively on steady-state gene expression. This is known to be false for certain genes and we show how modifying the assumptions can accommodate this discrepancy. Overall, our approach lays a rigorous, biophysical foundation for future studies of enhancer-gene regulation.

Introduction

Much of our current understanding of how genes are regulated arose from classical studies in bacteria of the lac operon and λ-phage [1]. However, the eukaryotic context differs from the bacterial in many significant ways. One key difference is that, in bacteria, regulatory DNA is found proximal to the gene, typically within 1kb upstream of the transcription start site (TSS), whereas eukaryotic regulatory sequences are found in discrete pieces that may be proximal to, or distal from, the TSS. The eukaryotic regulatory elements known as enhancers form a particularly important class. Enhancers were originally defined as DNA sequences which could drive the expression of genes in a location and orientation independent manner [2, 3]. Since these initial discoveries, many native enhancers have been identified which play critical roles in a variety of processes, such as embryonic development [4], physiology [5] and evolution [6]. Genetic variation in enhancers has also been shown to mediate risk for complex disease [7], Mendelian disease [8] and cancer [9]. Based on these and many other studies, we know that enhancers can be located over 1Mb from a target gene TSS, an individual enhancer may regulate multiple genes, some genes are regulated by multiple enhancers and the set of enhancers actively regulating a given gene may depend on cellular context. These properties have made it difficult to identify the rules governing enhancer-gene regulation.

Given their importance, much attention has been given to systematically identifying enhancer sequences and the genes they regulate. An important breakthrough has been the development of high-throughput CRISPR interference (CRISPRi) screens, which enable putative enhancer sequences to be perturbed in cell

culture and the resulting effect on expression of a target gene to be measured [10–13]. These screens typically measure quantitative effects on gene expression as the proportional change in mean gene expression over a cell population. We call this quantity the *fractional change* and, given its importance in this paper, define it formally as follows: let $\psi(g)$ denote the wild-type mean expression level of a gene g, in whatever units are used to measure it, and let $\psi(g, \not{e_q})$ denote the mean expression level of g after an enhancer of g, e_q, has been perturbed. The *fractional change* of e_q is then the non-dimensional quantity,

$$\mathrm{f}(g, e_q) := \frac{\psi(g) - \psi(g, \not{e_q})}{\psi(g)} \,. \tag{1}$$

The fractional change for thousands of putative enhancer-gene connections has been measured and computational methods have assessed whether the observed fractional change is statistically different from zero. Current efforts are now focused on two main questions. First, what can we learn about enhancer biology from these screens? Second, can the results from these screens be used to develop computational methods which can predict which enhancers regulate which genes in arbitrary cellular contexts?

The Activity-by-Contact (ABC) model has been proposed as a way to make progress on both of these questions [11]. The ABC model is based on the mechanistic notion that an enhancer's effect on gene expression depends on the intrinsic strength of the enhancer (activity) and the frequency with which it comes into physical proximity to the gene promoter (contact). The ABC model gives rise to the ABC Score, a quantitative formula which is intended to predict the fractional change observed in an enhancer perturbation experiment. For a gene, g, with N putative enhancers, $e_1, \cdots, e_N$, the ABC Score for a specific enhancer e_q, is given by,

$$\mathrm{ABC}(g, e_q) := \frac{\alpha_q \gamma_q}{\alpha_1 \gamma_1 + \cdots + \alpha_N \gamma_N} \,, \tag{2}$$

where α_i represents the activity of e_i and γ_i represents the contact frequency between e_i and the promoter of g. In [11] a putative enhancer was defined as a chromatin-accessible DNA element of approximately 500 base pairs; α_i was assigned using measures of chromatin state of the enhancer, such as DNase-Seq and H3K27ac ChIP-Seq; γ_i was assigned using the contact frequency between a putative enhancer and the gene promoter, as measured by Hi-C; and the sum in the denominator of Eqn.2 was taken over all putative enhancers within 5Mb of g.

The ABC Score is reasonably effective at predicting the results of CRISPRi screens. When considered as a binary classifier, the ABC Score has achieved a precision of 59% at 70% recall benchmarked against a database of nearly 4,000 putative enhancer-gene connections in the K562 cell line [11]. Similar performance has also been observed in other cell types [11, 14] and in subsequent benchmarking against other CRISPRi screens in K562 cells [15, Fig.S8a]. We emphasize that the ABC Score is computed directly from genomic data orthogonal to the CRISPRi experiment. As such, it has no free parameters and does not require fitting or training. The classification ability of the ABC Score and its modest input data requirements have resulted in its widespread use to interpret non-coding genetic variation [14, 16, 17], identify enhancers in disease related contexts [18, 19] and investigate the dosage effect of transcription factor concentration on gene expression [20].

Despite its practical utility as a binary classifier, the ability of the ABC Score to predict the fractional change is fundamentally limited [11, Fig.3c]. From Eqn. 2, it is clear that the sum of the ABC Scores over all putative enhancers of a given gene is equal to 1,

$$\sum_{i=1}^{N} \mathrm{ABC}(g, e_i) = 1 \,. \tag{3}$$

We define the total fractional change of a gene to be the sum of the fractional changes of all enhancers for the gene, $\mathrm{f}(g, e_1) + \cdots + \mathrm{f}(g, e_N)$. If the ABC model were perfectly reflecting the fractional change, so that $\mathrm{ABC}(g, e_i) = f(g, e_i)$, it would predict that the total fractional change for all genes is equal to 1. However, experimentally, a range of total fractional changes has been observed from 0 to greater than 3 [10, 11, 14, 21–23]. This incompatibility is a consequence of the algebraic structure of the ABC Score formula and

cannot be resolved in a straightforward way. For example, it cannot be resolved by using different types of epigenomic data to assign values to α_i or γ_i.

What, if anything, about enhancer biology can be concluded from the successes and limitations of the ABC Score? We believe that considering this question requires a formal description of the ABC model. The original description of the ABC model is *informal*, in the sense that the relationships between the mechanisms of activity and contact and the ABC Score formula were not determined by formal mathematical arguments. In consequence, the biological and biophysical assumptions that underlie formulas of this kind have not been clarified.

In the present paper, we present a strategy for the formal mathematical modeling of enhancer-gene regulation. We introduce the *default model*, a set of assumptions for how multiple enhancers independently regulate a gene. We show that a formula with the same algebraic structure as the ABC Score formula in Eqn.2 can be rigorously derived from a special case of the default model. This clarifies the assumptions that underlie the ABC Score formula. However, these assumptions also imply that the total fractional change of a gene is equal to one. We show how changing the assumptions of the default model can lead to total fractional changes which are less than or greater than one. More generally, the framework introduced here offers a rigorous foundation for future studies of enhancer-gene regulation.

Results

An activation-communication model of enhancer function

Our approach to modelling enhancer-gene regulation is based on the linear framework, a method of using graphs to analyse biomolecular systems [24–26] that has been previously introduced to study gene regulation [26]; see [27, 28] for up-to-date reviews. The graphs in question have vertices that are linked by labelled, directed edges. The vertices represent molecular states of DNA, the edges represent transitions between these molecular states and the labels represent the transition rates, which are positive numbers with dimensions of $(\text{time})^{-1}$.

An example linear framework graph is shown in Fig.1a. This graph, which we have called H, represents a single enhancer which can be either activated (filled red circle) or not and in communication with its target gene (curved arrow) or not. It thereby captures the two main notions in the original ABC model, although we prefer to speak here of "communication" rather than of "contact" (see below). These two features of the enhancer are treated in the graph as being independent of each other: the rates for becoming activated or deactivated do not depend on the state of communication, and the rates for making or losing communication do not depend on the state of activation. Independence will be one of the central features of our treatment and will appear both in how an individual enhancer is treated, as in this example in Fig.1a, and in how a gene is regulated by multiple enhancers, as we will explain below.

The graph H in Fig.1a represents a *coarse-graining* of the actual complexity of enhancer-gene regulation (Fig.1b). Activation is intended to capture processes local to the enhancer sequence such as transcription factor binding, chromatin reorganisation, nucleosome remodelling, recruitment of co-regulators or transcription of the enhancer sequence itself to generate enhancer RNA. Communication refers to the processes by which information is transferred from the enhancer to its target gene. Many communication mechanisms have been proposed including physical contact through DNA looping [29], diffusion of regulatory molecules [30] and phase separation [31]. We thus use the word 'communication' instead of 'contact' to reflect that enhancer-gene regulation may not require physical contact. It is, of course, possible that the specific activation or communication mechanisms may differ between enhancers. The value of this coarse-graining lies in not making commitments about the underlying mechanism, at the price of ignoring the potential consequences of how activation and communication are implemented in molecular terms. This particular coarse-graining will facilitate our clarification of the ABC Score formula below.

Having provided an example of a linear framework graph and explained how it describes the biological context that we will be studying, we now go into the details of the linear framework. We will make use of the example in Fig.1a throughout this work.

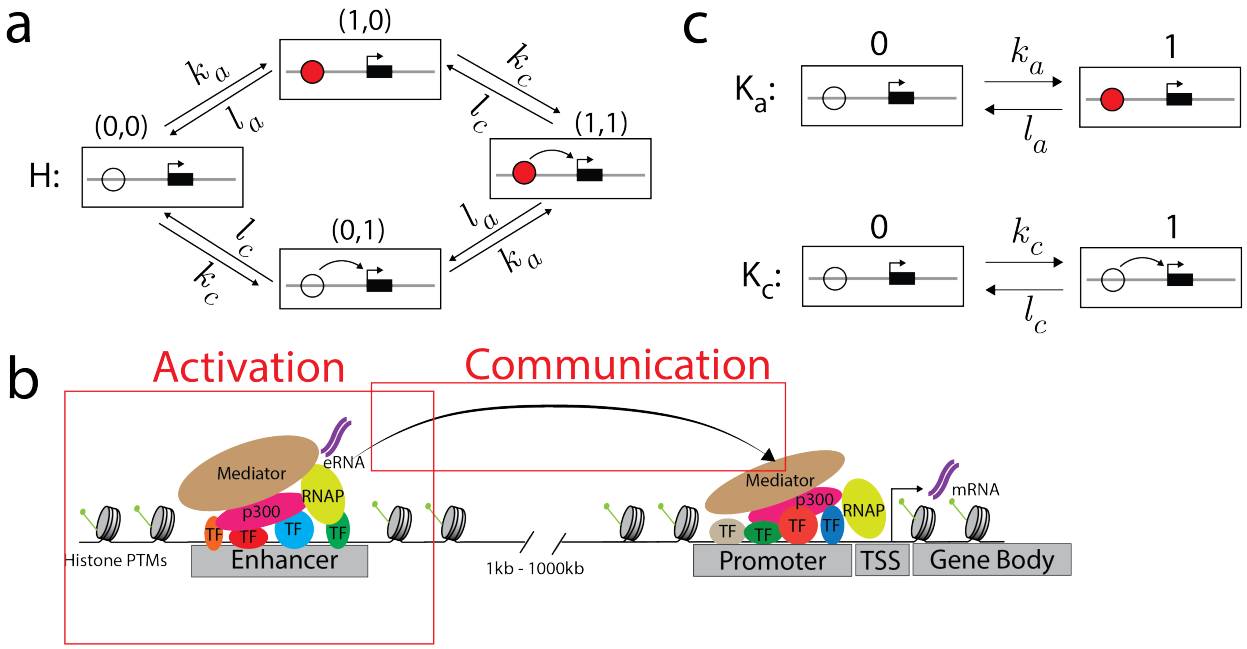

Figure 1: The activation-communication coarse graining. **a)** An example linear framework graph, H, representing a coarse-grained view of an enhancer. Each vertex contains a schematic of the enhancer (circle) and its target gene (black rectangle with the transcription start site marked with an arrow). The enhancer may be activated (filled red circle) or communicating (curved arrow to the target gene), encoded in the notation (i, j) used to denote vertices. The edge labels show that activation and communication take place independently of each other. **b)** A more detailed picture of the molecular complexity that may underlie the coarse-grained graph in panel **a**, as described further in the text. **c)** The example graph H in panel **a** is the graph product of two simpler 2-vertex graphs, K_a, which represents activation, and K_c, which represents communication. The product structure of H is equivalent to the independence of activation and communication.

Preliminaries on the linear framework

Notation and terminology

We will start by introducing some basic ideas about linear framework graphs. We will use a letter like G or H to refer to a graph. Vertices will generally be denoted i, j, etc. We will use the notation $i \in G$ to mean the state i from the graph G. Edges will be denoted $i \to j$ and edge labels will be denoted $\ell(i \to j)$. So, using the notation for the example graph H in Fig.1a, $\ell((0,0) \to (0,1)) = \ell((1,0) \to (1,1)) = k_c$ (the notation for the vertices in this graph arises from its product structure and will be explained later). If some feature X is being discussed for different graphs, we will sometimes use brackets, as in $X(G)$, or a subscript, as in X_G, to specify the graph in question. We will use the word *structure* to refer to just the vertices and edges of a graph, ignoring the edge labels; when we say "graph", we will always be including the labels, even when they are not mentioned explicitly.

The Markov process

A graph G is equivalent to a finite-state, continuous-time, time-homogeneous Markov process [25, 28, 32]. This stochastic behaviour can be understood as follows. If the system is in state i, then for each edge $i \to j$ which leaves i, a "firing" time is randomly chosen from the exponential probability distribution, $\lambda \exp(-\lambda t)$, where λ is the transition rate of that edge, $\lambda = \ell(i \to j)$, and the edge with the lowest firing time is taken,

at that time. This generates a stochastic trajectory of states and transitions. If we follow a trajectory up to time T and measure the proportion of time spent in state i, then that ratio stabilises with increasing T to become the *steady-state probability* of state i [32], which we will denote by $u_i^*(G)$. Provided G is *strongly connected*, this quantity does not depend on the state in which the trajectory starts and the steady-state probability is a property of the graph [25]. A strongly connected graph is one in which any two distinct vertices, i and $j \neq i$, are connected by a directed path, $i = i_1 \to i_2 \to \cdots \to i_k = j$. The example graph H in Fig.1a is strongly connected but ceases to be if the edges $(1,1) \to (1,0)$ and $(1,1) \to (0,1)$ are removed. We will assume from now on that all our graphs are strongly connected.

Thermodynamic equilibrium and steady-state probabilities

One of the advantages of the linear framework is that, provided the graph is finite, its steady-state probabilities can be calculated algebraically in terms of the edge labels. (We will encounter an infinite graph below but, as we will see, we do not have to deal with them directly and can work only with finite graphs.) If the graph can reach *thermodynamic equilibrium* the algebra can be done quite easily but, importantly, it can also be done when the graph is away from thermodynamic equilibrium, although the formulas become more complicated. A graph can reach thermodynamic equilibrium if, and only if, it satisfies two conditions. First, it must be *reversible*, so that if there is an edge $i \to j$, then there is also an edge $j \to i$, which represents the reverse of the process that corresponds to $i \to j$. Second, it must satisfy the *cycle condition*: the product of the label ratios around any cycle of reversible edges must be 1. The graph in Fig.1a is evidently reversible and has only one cycle of reversible edges, $(0,0) \leftrightharpoons (1,0) \leftrightharpoons (1,1) \leftrightharpoons (0,1) \leftrightharpoons (0,0)$, for which the product of label ratios is

$$\left(\frac{k_a}{l_a}\right)\left(\frac{k_c}{l_c}\right)\left(\frac{l_a}{k_a}\right)\left(\frac{l_c}{k_c}\right) = 1\,. \tag{4}$$

For this graph, the independence of activation and communication ensures that the graph can reach thermodynamic equilibrium.

When a graph can reach thermodynamic equilibrium, its steady-state probabilities can be calculated as follows. First, choose any vertex as a reference; let us call it 1. Second, choose any path of reversible edges from 1 to the state in question, say i: $1 \leftrightharpoons i_1 \leftrightharpoons \cdots \leftrightharpoons i_k = i$. The steady-state probability of i is then proportional to the product of the label ratios along this path,

$$u_i^*(G) \propto \left(\frac{\ell(i_1 \to i_2)}{\ell(i_2 \to i_1)}\right) \times \cdots \times \left(\frac{\ell(i_{k-1} \to i_k)}{\ell(i_k \to i_{k-1})}\right)\,. \tag{5}$$

It is a simple consequence of the cycle condition that the quantity on the right-hand side of Eqn.5 does not depend on the choice of path from 1 to i. The proportionality constant in Eqn.5 is readily obtained by exploiting the fact that the sum of all the probabilities must be 1, so that, if the vertices are denoted $1, \cdots, N$, then $u_1^*(G) + \cdots + u_N^*(G) = 1$. If we follow this prescription for the graph in Fig.1a, we find that, for example,

$$u_{(1,1)}^*(G) = \frac{(k_a/l_a)(k_c/l_c)}{1 + (k_a/l_a) + (k_c/l_c) + (k_a/l_a)(k_c/l_c)}\,. \tag{6}$$

$$= \frac{k_a}{k_a + l_a} \cdot \frac{k_c}{k_c + l_c} \tag{7}$$

(We will sometimes use a "$\cdot$" to denote multiplication to make formulas like this look clearer.) The reorganisation of Eqn.6 into Eqn.7 reveals a product structure in the algebra whose significance will emerge below. The formula in Eqn.6 is the same as would arise from equilibrium statistical mechanics. It is one of the features of the linear framework that it reduces to equilibrium statistical mechanics for systems that are at thermodynamic equilibrium but also yields algebraic formulas for systems away from thermodynamic equilibrium.

Product graphs as models of independence

In studying gene regulation, a very helpful construction is that of a *product graph*, because it captures the default situation in which two or more genetic systems operate independently of each other. The example graph in Fig.1a is a case in point. This graph H is the product of the graphs K_a and K_c in Fig.1c. Here, K_a is a two-vertex graph that represents just the activation of the enhancer and K_c is a two-vertex graph that represents just the communication.

We will use K_a and K_c to describe the product graph construction. We will do this in two steps. We will first specify the vertices and edges by building the *product structure*, denoted $K_a \times K_c$, and then we will specify the labels to get the *product graph*, denoted $K_a \otimes K_c$. As we will see below, product structures underlie other constructions in which the independence of the product graph is broken, which is why it is helpful to distinguish structures and graphs. The vertices in $K_a \times K_c$ are ordered pairs, (i, j), of vertices $i \in K_a$ and $j \in K_c$. The edges in $K_a \times K_c$ arise from the edges in either component K_a or K_c, taken independently of the state of the other component. In other words, if $i_1 \to i_2$ is any edge in K_a, then $(i_1, j) \to (i_2, j)$ is an edge in $K_a \times K_c$, for all $j \in K_c$; similarly, if $j_1 \to j_2$ is any edge in K_c, then $(i, j_1) \to (i, j_2)$ is an edge in $K_a \times K_c$, for all $i \in K_a$; these are the only edges in $K_a \times K_c$. This prescription yields the structure of the graph H in Fig.1a.

The labels of the product graph, $K_a \otimes K_c$, are also inherited from those in K_a or K_c, independently of the state of the other component,

$$\ell_{K_a \otimes K_c}((i_1, j) \to (i_2, j)) = \ell_{K_a}(i_1 \to i_2) \quad \text{and} \quad \ell_{K_a \otimes K_c}((i, j_1) \to (i, j_2)) = \ell_{K_c}(j_1 \to j_2)\,.$$

We see that $K_a \otimes K_c$ corresponds exactly to the graph H in Fig.1a. The graph product precisely captures the sense in which the components of the product, here K_a and K_c, operate independently of each other: the transitions in either component are unaffected, as to their occurrence and their rates, by the state of the other component.

In the more general case of a product of m graphs, the vertices are naturally indexed as ordered tuples, $(i_1, \cdots, i_m)$.

One of the consequences of the product graph construction is that its steady-state probabilities are easily calculated. If $K_1, \cdots, K_N$ are any set of N strongly connected graphs, then the steady-state probabilities in the product graph $K_1 \otimes \cdots \otimes K_N$ can be computed by multiplying the steady-state probabilities in the individual graphs,

$$u^*_{(i_1,\cdots,i_N)}(K_1 \otimes \cdots \otimes K_N) = u^*_{i_1}(K_1) \cdots u^*_{i_N}(K_N)\,. \tag{8}$$

Eqn.8, which is proved in [26], again captures the sense in which the components $K_1, \cdots, K_N$ are independent of each other. We note that Eqn.8 holds even for graphs which are unable to reach thermodynamic equilibrium.

We can see Eqn.8 at work for the graph in Fig.1a, which is the product of the graphs K_a and K_c in Fig.1c. If we follow the prescription in Eqn.5, we see that

$$u^*_1(K_a) = \frac{k_a/l_a}{1 + k_a/l_a} \quad \text{and} \quad u^*_1(K_c) = \frac{k_c/l_c}{1 + k_c/l_c}\,. \tag{9}$$

If we apply Eqn.8 to the formulas above, we see that,

$$\begin{aligned} u^*_{(1,1)}(K_a \otimes K_c) &= \left(\frac{k_a/l_a}{1 + k_a/l_a}\right)\left(\frac{k_c/l_c}{1 + k_c/l_c}\right) \\ &= \frac{k_a}{k_a + l_a} \cdot \frac{k_c}{k_c + l_c}\,, \end{aligned} \tag{10}$$

which recovers the expression in Eqn.7, whose algebraic product structure is now seen to reflect the underlying product graph.

Eqn.8 for individual vertices has a straightforward extension to subsets of vertices. To explain this, let K be any graph and let $S \subseteq K$ be any subset of vertices in K. The steady-state probability of being in any

vertex of S, denoted $u^*_S(K)$, is given by $u^*_S(K) = \sum_{i\in S} u^*_i(K)$. Now suppose, as above, that $K_1, \cdots, K_N$ are any strongly connected graphs. Let $S_i \subseteq K_i$ be any subset of vertices of K_i and let $S_1 \times \cdots \times S_N$ be the corresponding *set product* in K. This set product, for which we use, for convenience, the same notation as for the product structure, has the obvious definition that it consists of all those tuples $(i_1, \cdots, i_N)$ where $i_k \in S_k$. It is then a simple consequence of Eqn.8 that,

$$u^*_{S_1\times\cdots\times S_N}(K_1 \otimes \cdots \otimes K_N) = u^*_{S_1}(K_1) \cdots u^*_{S_N}(K_N)\,. \tag{11}$$

One of the implications of Eqn.11 is that if we take i_j to be a coordinate that runs over the vertices of K_j, then the probability that i_j has a particular value, say $i_j = b$, remains the same irrespective of the other factors in the graph product,

$$u^*_{\{i_j=b\}}(K_1 \otimes \cdots \otimes K_N) = u^*_{\{i_j=b\}}(K_j)\,. \tag{12}$$

Eqn.12 follows from Eqn.11 because the subset $\{i_j = b\}$ in $K_1 \otimes \cdots \otimes K_N$ is the product subset,

$$K_1 \times \cdots \times K_{j-1} \times \{i_j = b\} \times K_{j+1} \times \cdots \times K_N\,,$$

and $u^*_{K_i}(K_i) = 1$. We can see an example of Eqn.12 at work in Figure 1. Let $\{a{=}1\} = \{(1,0), (1,1)\}$ be the subset of vertices of H in which the enhancer is activated. Then Eqn.12 shows that $u^*_{\{a=1\}}(H) = u^*_1(K_a)$. We will make further use of Eqn.12 in what follows.

The gene expression response

The graphs we have considered up to now are models of the regulatory state of the gene. We now discuss how to incorporate the production and degradation of mRNA. The standard approach in the literature is known as *kinetic modeling* and uses a Markovian framework based on the chemical master equation [33]. We follow this same approach within the graph-theoretic setting introduced here.

At any given time, the state of gene expression is specified by a certain number of molecules of the corresponding mRNA. This number increases by 1 each time RNA polymerase transcribes the gene and decreases by 1 each time an mRNA molecule is degraded or lost through transport out of the nucleus. We can represent such an expression system by the (semi)-infinite *pipeline* structure, P, in which the state p represents the number p of mRNA molecules, from $p = 0$ onwards, and the edges correspond to mRNA production, $p \to p+1$, and degradation or loss, $p \to p-1$ (Fig.2a).

Given a gene-regulatory graph, G, we represent the overall system of regulation and expression by a *copy-number graph*, $G \ltimes P$, that will be derived from the product structure, $G \times P$ (Fig.2b). The states of $G \ltimes P$ are identical to those of $G \times P$ but $G \ltimes P$ may not have all the edges that are present in $G \times P$. Each state in $G \ltimes P$ keeps track of the regulatory state of the gene and the number of mRNA molecules that are present. We now discuss how to assign labels to this graph (Fig.2b). We assume that each state, $i \in G$, has a corresponding non-negative rate of mRNA production, $r_i(G) \geq 0$. If $r_k(G) = 0$, so that mRNA production is not possible in state k of G, then the edges $(k,p) \to (k,p+1)$ are removed from $G \times P$ for all $p \in P$. (Note that edge labels must always be positive.) This is the only way in which the structure of $G \ltimes P$ differs from that of $G \times P$. If $r_k(G) > 0$, we will assume that the rate of mRNA production does not change with the number of mRNA molecules that have been expressed, so that $\ell((k,p) \to (k,p+1)) = r_k(G)$ for any $p \in P$. As for mRNA degradation or loss, this takes place independently of the regulatory system, so the most parsimonious assumption is that its rate is proportional to the number of mRNAs that are present and is independent of the regulatory state. Accordingly, we may write $\ell((k,p) \to (k,p-1)) = \delta(G) \cdot p$ for any $k \in G$ and any positive $p \in P$, where $\delta(G)$ is the degradation rate constant. Finally, we assume that regulatory transitions do not depend on gene expression, so that $\ell((i,p) \to (j,p)) = \ell_G(i \to j)$ for all $i, j \in G$ and for all $p \in P$. A compact way to visually represent a copy-number graph is shown in Fig.2c.

At steady state, $G \ltimes P$ gives rise to a probability distribution over the mRNA copy number. We will define the response of the gene, which we will denote by $R(G)$, to be given by the average of this number distribution,

$$R(G) := \sum_{(i,p)} p \cdot u^*_{(i,p)}(G \ltimes P)\,. \tag{13}$$

We note that $R(G) \geq 0$.

Because $G \ltimes P$ is not a finite graph, the prescription given in Eqn.5 for calculating steady-state probabilities no longer works. ($G \ltimes P$ is also not at thermodynamic equilibrium, unless every regulatory state has the same rate of mRNA production, as can be checked by following the cycle condition formula in Eqn.4.) However, we can appeal to a very useful theorem, due to Sanchez and Kondev, which tells us that we do not have to operate on $G \ltimes P$ in order to calculate $R(G)$ [34]. Translating their work into the graph-theory language used here, we find that the response of the gene can be calculated in terms of the average of $r_i(G)$ over only the the steady-state probabilities of G, normalized by $\delta(G)$,

$$R(G) = \frac{1}{\delta(G)} \sum_{i \in G} r_i(G) \cdot u_i^*(G) \,. \tag{14}$$

It follows that, although infinite graphs arise to represent the mRNA expression system, we do not need to work with them to calculate the mean steady-state expression $R(G)$, under the assumptions made above. Sanchez and Kondev did not use graph theory in their work, so we provide an independent graph-based proof of Eqn.14 in the Methods. In subsequent work, we will show how the copy-number graphs introduced here lead to generalizations of the results of [34] but we do not need that for the present paper.

This result provides some justification for reducing the notational clutter from multiple instances of P. We will refer to the *regulatory graph* G when we mean G on its own, and to the *copy-number graph* G when we mean $G \ltimes P$, defined for some specified choice of production rates $r_i(G)$ and degradation rate $\delta(G)$. These parameters may not be explicitly mentioned when speaking of a copy-number graph but they should be kept in mind.

As an illustration of Eqn.14, we will assign production rates to the graph in Fig.1a and compute its response. We will make the assumption that mRNA is only produced when the enhancer is both activated and communicating (Fig. 2d). The mRNA production rates of H are therefore given by

$$r_{(0,0)}(H) = r_{(1,0)}(H) = r_{(0,1)}(H) = 0 \quad \text{and} \quad r_{(1,1)}(H) = r \,. \tag{15}$$

We can now use Eqn.14 to calculate the response, $R(H)$, taking advantage of Eqn.10, in which we exploited the product graph decomposition $H = K_a \otimes K_c$. We see that,

$$R(H) = \frac{r}{\delta} \cdot \left(\frac{k_a}{k_a + l_a} \right) \left(\frac{k_c}{k_c + l_c} \right) \,. \tag{16}$$

It follows from Eqns.9 and 12 that we can interpret $k_a/(k_a + l_a)$ as the probability that the enhancer is activated and, similarly, $k_c/(k_c + l_c)$ as the probability that the enhancer is communicating. Eqn.16 tells us that the response of H is the product of the ratio of production to degradation, the probability of activation and the probability of communication.

This concludes our analysis of a gene regulated by a single enhancer using the activation-communication coarse graining. We now turn to considering how multiple enhancers work together to regulate gene expression.

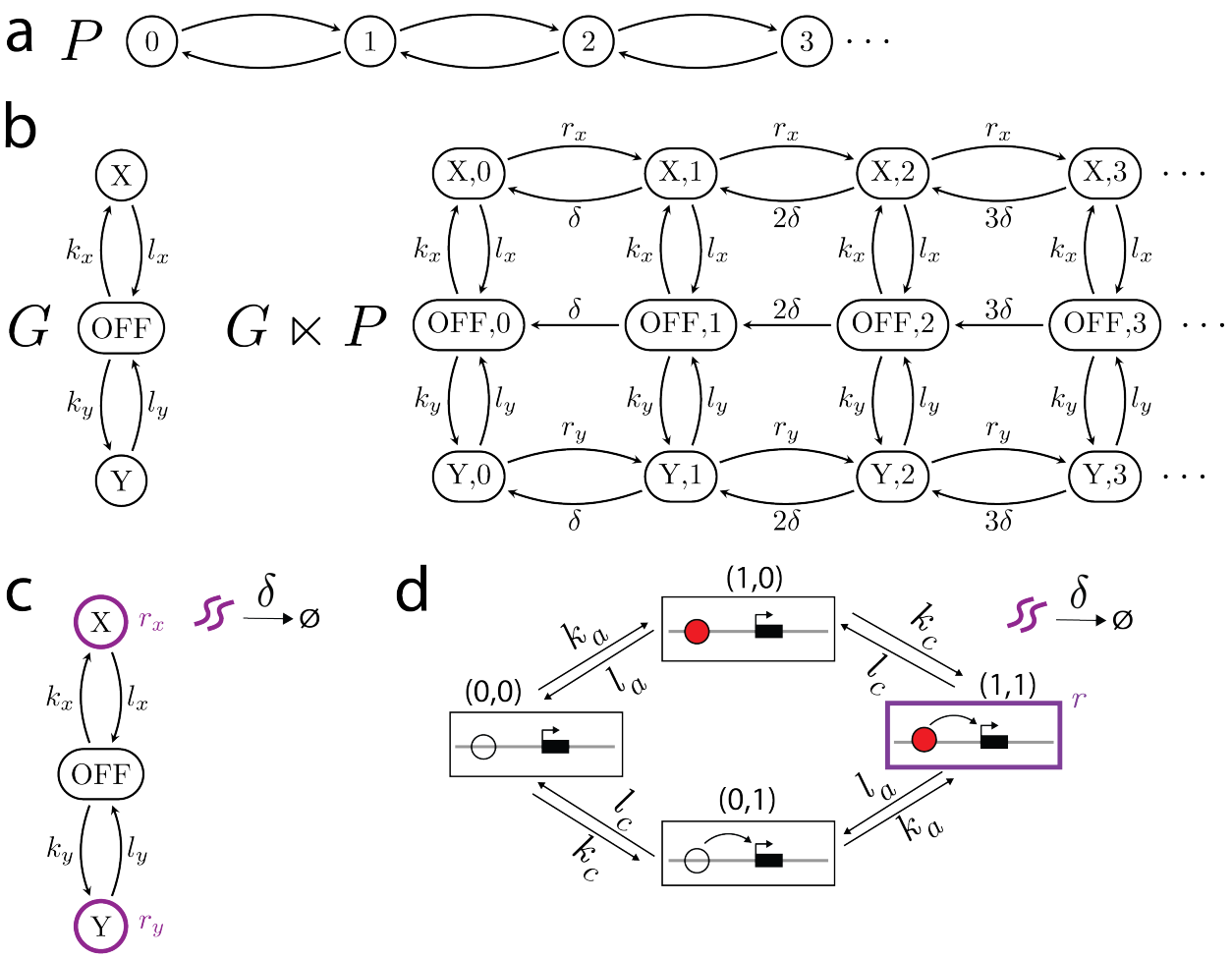

Figure 2: Modeling mRNA production and degradation through copy-number graphs. **a)** The pipeline structure P represents the number of mRNA molecules and their production and loss. **b)** An example regulatory graph, G, and the resulting copy-number graph $G \ltimes P$. In this example G has two production states, X and Y, with corresponding mRNA production rates r_x and r_y respectively. We note that $G \ltimes P$ is a sub-structure of $G \times P$; it has the same vertices but lacks the edges corresponding to a production rate of zero. G and P also do not operate independently in $G \ltimes P$ because the mRNA production rates depend on the regulatory state. **c)** A compact way to represent $G \ltimes P$. The production states are outlined in purple with corresponding mRNA production rates. The degradation rate, δ, is shown above the arrow from purple squiggles (mRNA) to the empty set $\varnothing$. **d)** The compact representation of the graph $H \ltimes P$, where H is given in Fig.1a. The only production state is the state $(1, 1)$, in which the enhancer is both activated and communicating, which has production rate r.

A default model of how multiple enhancers independently regulate a gene

We now introduce the *default model* of enhancer-gene regulation. This is a set of assumptions for how multiple enhancers collectively regulate a gene in an independent manner. We previously introduced the product graph construction which represents independence between regulatory graphs. We now broaden those assumptions to also allow for mRNA production. We expect this default model construction to be of general interest. In the next section we will show how a special case of the default model clarifies the ABC Score formula.

Consider a gene, g, that is regulated by N enhancers, $e_1, \cdots, e_N$. We will assume that enhancer e_l is modelled by the graph G_l. We make no assumptions about G_l other than the prevailing assumption that all our graphs are strongly connected. G_l could be substantially more complicated than the graph in Fig.1a and could incorporate, for example, chromatin organisation, nucleosomes, co-regulators, post-translational modifications, chromosome conformation, etc [26]. In particular, there is no requirement that G_l should be able to reach thermodynamic equilibrium. At this point our assumptions are very general and could apply to essentially any enhancer, when considered from a Markovian perspective.

We denote the graph that models the collective regulation of the enhancers by G and describe how G is defined in terms of the G_l.

The first assumption says that each enhancer has its own individual effect.

1. Individuality. Each enhancer e_l, when acting in the absence of any of the other enhancers, drives gene expression at the rate $r_i(G_l) \geq 0$ for each state $i \in G_l$, and gives rise to the response $R(G_l)$, as defined by Eqn.13. If the enhancer is unable to drive expression on its own, then $r_i(G_l) = 0$ for every state $i \in G_l$.

The next two assumptions specify how the enhancers work together.

2. Regulatory independence. Each enhancer acts independently of all the others, so that the regulatory graph of G is given by the product graph $G_1 \otimes \cdots \otimes G_N$.

3. Production-rate summation. Each enhancer independently influences mRNA production. Accordingly, if $(i_1, \cdots, i_N)$ is a state in G, then its mRNA production rate is a sum of the corresponding production rates in each enhancer graph:

$$r_{(i_1,\cdots,i_N)}(G) = r_{i_1}(G_1) + \cdots + r_{i_N}(G_N)\,. \tag{17}$$

The summation of rates in Assumption 3 arises for the following reason. If each enhancer influences mRNA production independently, then the time at which an mRNA is produced will be the minimum of the times at which each individual enhancer has its effect on production. These individual times are exponentially distributed with rates $r_{i_j}(G_j)$ for state i_j in G_j. The minimum of several exponentially distributed random variables is a random variable that is also exponentially distributed, with rate given by the sum of the individual rates. This leads to Eqn.17. The final assumption specifies the degradation rates.

4. Uniform degradation. Since mRNA degradation is a separate process to gene regulation and gene expression, we consider the characteristic degradation rate to be a property of the gene, not the enhancer. As such, each graph G_l is assumed to have the same degradation rate, $\delta(G_l) = \delta$ for all l, and the mRNA degradation rate of G is also δ: $\delta(G) = \delta$.

For any set of copy-number graphs $G_1, \ldots, G_N$, we denote the copy-number graph which models their collective effect on transcription according to Assumptions 1 to 4 by the graph

$$G_1 \circledast \cdots \circledast G_N\,. \tag{18}$$

Example constructions using the default model are given in Fig.3.

Assumptions 1 to 4 specify our default model of how enhancers collectively regulate a gene. Whether any of the default model assumptions hold for an individual gene is a question that has to be addressed experimentally. In particular, we would expect that Assumption 3 would eventually break down as more enhancers are added to a gene since production rates will be limited by the physical processes involved in transcription. Our goal here is to rigorously work out the consequences of these assumptions, so that we know what to expect when the assumptions do hold and can compare these predictions to what is found experimentally. Of particular significance is that the assumptions above imply that the collective response of the enhancers is always the sum of their individual responses,

$$R(G_1 \circledast \cdots \circledast G_N) = R(G_1) + \cdots + R(G_N)\,. \tag{19}$$

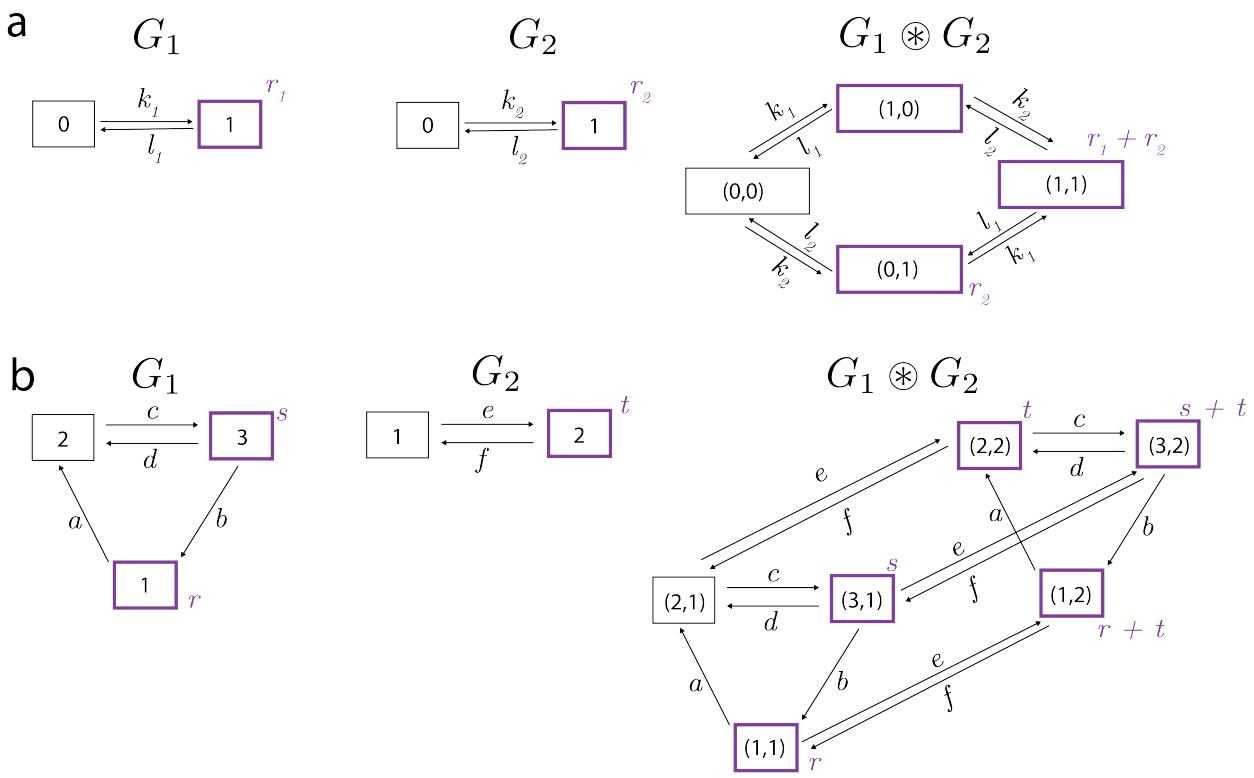

Figure 3: Two examples **a** and **b** of the default model construction. The copy-number graphs are depicted in compact format, as shown in Fig.2c but omitting the degradation symbols for clarity. Production states are outlined in bold purple with corresponding production rates in purple text. The model in **b** is adapted from Figure 9 of [26].

A proof of this fundamental property of the default model is given in the Methods.

A recent commentary has argued that formal definitions and rigorous modeling are necessary to investigate whether a set of enhancers is "greater than the sum of its parts" [35]. We fully agree and suggest that the notion of independence encoded by the default model, which gives rise to Eqn.19, could serve as a definition of what it means for a gene to be the *sum of its parts*.

Transcription in the default model relies on the presence of enhancers. It is well known that the promoter sequences at some eukaryotic genes are sufficient to drive transcription even in the absence of distal enhancers [36, 37]. It is a future area of research to incorporate the role of core promoter elements and promoter proximal regulatory sequences along with their interactions with distal enhancers.

A clarification of the ABC Score formula

Enhancer perturbation and deletion fidelity

To see how formulas similar to the ABC Score can be derived from the default model, we need to consider how to formally model perturbations to enhancers such as genetic deletions or CRISPRi. As previously, we will assume that the target gene g is collectively regulated by enhancers $e_1, \cdots, e_N$. We assume that enhancer e_i is modeled by the graph G_i and that g is modeled by $G_1 \circledast \cdots \circledast G_N$. Let us consider what happens when enhancers $e_{l_1}, \cdots, e_{l_k}$ are perturbed in such a way that they are considered to no longer be working to regulate g. We will use a similar notation to that for the fractional change in the Introduction and denote the graph that arises from this perturbation as $G[\not{e}_{l_1}, \cdots, \not{e}_{l_k}]$. In analogy to the fractional change,

we can define the *deletion effect* of the perturbation, $\Delta(G; e_{l_1}, \cdots, e_{l_k})$, to be the proportional change in response of g,

$$\Delta(G; e_{l_1}, \cdots, e_{l_k}) := \frac{R(G) - R(G|\cancel{e_{l_1}}, \cdots, \cancel{e_{l_k}})}{R(G)} \,. \tag{20}$$

It is important to keep in mind that the deletion effect is defined in terms of a model of gene regulation, whereas the fractional change is defined in terms of experimental data. The definition in Eqn.20 implicitly assumes that the system has returned to steady state following the perturbation. Furthermore, Eqn.20 says nothing about how the enhancer is perturbed or whether a CRISPRi perturbation has the same effect as a genetic deletion.

To calculate the deletion effect using Eqn.20, we need to know the perturbed graph, $G|\cancel{e_{l_1}}, \cdots, \cancel{e_{l_k}}$. We assume that the graph shows *deletion fidelity*, which implies that the perturbation completely abrogates the function of the targeted enhancers and does not influence other enhancers. Let $m_1, \cdots, m_p$ be the remaining indices in $1, \cdots, N$ after $l_1, \cdots, l_k$ have been removed.

5. Deletion fidelity. The regulatory graph of $G|\cancel{e_{l_1}}, \ldots, \cancel{e_{l_k}}$ is the graph product $G_{m_1} \otimes \cdots \otimes G_{m_p}$, and the production rates in $G|\cancel{e_{l_1}}, \ldots, \cancel{e_{l_k}}$ are directly inherited from G,

$$r_{(i_{m_1}, \cdots, i_{m_p})}(G|\cancel{e_{l_1}}, \ldots, \cancel{e_{l_k}}) = r_{i_{m_1}}(G_{m_1}) + \cdots + r_{i_{m_p}}(G_{m_p}) \,. \tag{21}$$

Deletion fidelity ensures that if G obeys Assumptions 1-4, then $G|\cancel{e_{l_1}}, \cdots, \cancel{e_{l_k}}$ also obeys Assumptions 1-4 for the remaining enhancers $e_{m_1}, \ldots, e_{m_p}$ and that, for the copy-number graphs,

$$G|\cancel{e_{l_1}}, \cdots, \cancel{e_{l_k}} = G_{m_1} \circledast \cdots \circledast G_{m_p} \,. \tag{22}$$

Using Eqn.22, it follows from Eqn.19 that,

$$R(G|\cancel{e_{l_1}}, \cdots, \cancel{e_{l_k}}) = R(G_{m_1}) + \cdots + R(G_{m_p}) \,, \tag{23}$$

and so the formula for the deletion effect in Eqn.20 tells us that,

$$\Delta(G; e_{l_1}, \cdots, e_{l_k}) = \frac{R(G_{l_1}) + \cdots + R(G_{l_k})}{R(G_1) + \cdots + R(G_N)} \,. \tag{24}$$

Eqns.23 and 24 are general properties that hold for the default model whenever Assumption 5 of deletion fidelity also holds. They allow us to formalise the notion of enhancer *additivity*, which we will discuss below, but, first, let us turn to the ABC Score formula.

The Independent-Activation-Communication (IAC) model

In the default model, the graph representing each individual enhancer can be arbitrarily complicated. To show how the ABC Score formula can arise from the default model, we need to impose the further assumption that each enhancer is modeled by the activation-communication coarse graining shown in Figs.1a and 1b.

6. The activation-communication coarse-graining. Enhancer e_i is described by the graph H_i, where H_i is the same graph as H in Fig.2d. Specifically, H_i is the graph product of an activation graph, $K_{a,i}$, with labels $k_{a,i}, l_{a,i}$, and a communication graph, $K_{c,i}$, with labels $k_{c,i}, l_{c,i}$ (Fig.1c), and $H_i = K_{a,i} \otimes K_{c,i}$. The only non-zero production rate of H_i occurs in the state in which the enhancer is both active and communicating, where the rate is r_i.

The overall regulatory system is then described by $G = H_1 \circledast \cdots \circledast H_N$ (Fig. 4, Fig.S1). We call the model obeying Assumptions 1-6 the Independent-Activation-Communication (IAC) model. It follows from Eqn.16 that the response of enhancer i in the IAC model is given by

$$R(H_i) = \frac{r_i}{\delta} \left(\frac{k_{a,i}}{k_{a,i} + l_{a,i}} \right) \left(\frac{k_{c,i}}{k_{c,i} + l_{c,i}} \right) \,. \tag{25}$$

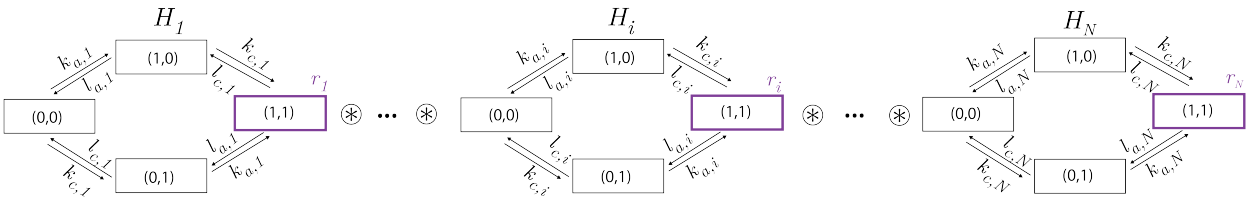

Figure 4: The Independent-Activation-Communication (IAC) model. A gene described by the IAC model follows Assumptions 1-6 for the component graphs $H_1, \ldots, H_N$. The ordered pair of binary digits for the vertices in each H_i represent the activation and communication status, respectively, of each enhancer. Each H_i has the same structure as the graph in Fig.2d, but different labels, and represents the independence of activation and communication within each enhancer. Each H_i is assumed to have the same degradation rate which is omitted for clarity. See also Fig.S1.

Let us define $\tilde{\alpha}_i := k_{a,i}/(k_{a,i}+l_{a,i})$ and $\tilde{\gamma}_i := k_{c,i}/(k_{c,i}+l_{c,i})$ and recall from Eqn.16 that these quantities are the probability of activation and the probability of communication, respectively, of enhancer e_i. According to the fundamental property of the default model in Eqn.19, the response, $R(G)$, of the overall graph, $G = H_1 \circledast \cdots \circledast H_N$, is given by,

$$R(G) = \frac{1}{\delta} \sum_{i=1}^{N} r_i \tilde{\alpha}_i \tilde{\gamma}_i \,. \tag{26}$$

Furthermore, as a consequence of deletion fidelity (Assumption 5), it follows from Eqn.24 that the deletion effect for enhancer e_q is given by,

$$\Delta(G; e_q) = \frac{r_q \tilde{\alpha}_q \tilde{\gamma}_q}{r_1 \tilde{\alpha}_1 \tilde{\gamma}_1 + \cdots + r_N \tilde{\alpha}_N \tilde{\gamma}_N} \,. \tag{27}$$

Eqn.27 shows a striking algebraic similarity to the ABC Score formula in Eqn.2. The quantity $\tilde{\gamma}_i$, which is the probability of communication, is analogous to the 'frequency of contact', γ_i, that was envisaged for the ABC model [11] and appears in Eqn.2. There are different possible interpretations for the other terms. One potential interpretation for the term $r_i \tilde{\alpha}_i$ in Eqn.27, which is the production rate multiplied by the probability of activation, is that it is analogous to the 'strength of the enhancer', α_i, appearing in Eqn.2. Another potential interpretation is that $\tilde{\alpha}_i$ corresponds to α_i and that the production rates r_i are not represented in the ABC model. If the production rates are assumed to be equal, they would cancel out in Eqn.2, which would be consistent with a correspondence between $\tilde{\alpha}_i$ and α_i. Such interpretational ambiguities are to be expected because the ABC model is informal, while the IAC model presented here is formal. Moreover, our formal model separately specifies the regulatory state of the enhancer and its effect on transcription, whereas the ABC model does not make this distinction. An interesting question arises as to how numerical values can be assigned to the terms in Eqn.27, and how this may differ from the strategies used in [11] to give numerical values to the terms in Eqn.2, but this is an area for future work.

Eqn.27 is our clarification of the algebraic structure of the ABC Score formula. Eqn.27 rigorously follows if enhancers collectively regulate a gene according to the IAC model (Assumptions 1 to 6).

Enhancer additivity and departures from it

The default model, satisfying Assumptions 1 to 4, exhibits *response additivity*, as shown by Eqn.19: the response of the gene to all the enhancers acting collectively is just the sum of the responses to each individual enhancer. When the default model also obeys Assumption 5 of deletion fidelity, then response additivity has a counterpart in the deletion effect, as defined in Eqn.20. This allows us to rigorously define the properties of *super-additivity* and *sub-additivity*. These departures from the properties of the default model may be helpful to interpret the effects of experimental perturbations, such as genetic deletions or CRISPRi, in which

subsets of enhancers are prevented from influencing a gene and the effect of these perturbations on the gene expression response is measured.

With Assumptions 1 to 5, if $U_1, \cdots, U_m$ are pairwise disjoint subsets of enhancers, so that $U_i \subseteq \{e_1, \cdots, e_N\}$ and $U_i \cap U_j = \emptyset$ when $i \neq j$, then it follows from Eqn.24 that the effect of deleting all the subsets together is just the sum of the individual deletion effects,

$$\Delta(G; U_1 \cup \cdots \cup U_m) = \Delta(G; U_1) + \cdots \Delta(G; U_m)\,. \tag{28}$$

We refer to this property as *deletion additivity*. Furthermore, it is evident from Eqn.24 that, if all the enhancers are deleted, so that $U_1 \cup \cdots \cup U_m = \{e_1, \cdots, e_N\}$, then the *total deletion effect* must be 1,

$$\Delta(G; U_1) + \cdots + \Delta(G; U_m) = \Delta(G; \{e_1, \cdots, e_N\}) = 1\,. \tag{29}$$

Assuming deletion fidelity, the total deletion effect being 1 is equivalent to the response additivity in Eqn.19. A special case of Eqn.29 arises if all enhancers are deleted individually, when, once again, the total deletion effect is 1,

$$\Delta(G; e_1) + \cdots + \Delta(G; e_N) = 1\,. \tag{30}$$

Now suppose that a gene g is regulated by N enhancers, $e_1, \cdots, e_N$, each enhancer is modeled by the graph G_i and the regulatory graph of g, G, has the product structure, $G_1 \times \cdots \times G_N$. The labels in G need not be related to those of the component graphs G_i, so that G need not be the product graph $G_1 \otimes \cdots \otimes G_N$. We can no longer calculate $R(G)$ in terms of $R(G_i)$. However, we can still define through Eqn.20 the deletion effect $\Delta(G; U)$ for any collection $U \subseteq \{e_1, \cdots, e_N\}$ of enhancers. We say that g exhibits response *super-additivity* if,

$$R(G) > R(G_1) + \cdots + R(G_N)\,. \tag{31}$$

In terms of the deletion effect, this corresponds to when a collective deletion has less effect than the sum of the individual deletions, so that,

$$\Delta(G; U_1 \cup \cdots \cup U_m) < \Delta(G; U_1) + \cdots + \Delta(G; U_m)\,. \tag{32}$$

Similarly, g exhibits response *sub-additivity* if,

$$R(G) < R(G_1) + \cdots + R(G_N)\,. \tag{33}$$

and this corresponds to the collective deletion having more effect that the sum of the individual deletions,

$$\Delta(G; U_1 \cup \cdots \cup U_m) > \Delta(G; U_1) + \cdots + \Delta(G; U_m)\,. \tag{34}$$

Experimentally, response additivity [15, 23, 38–43], super-additivity [15, 21, 39–44] and sub-additivity [38, 40, 41] have all been observed. Because the super-additive and sub-additive findings cannot be accounted for by the default model with deletion fidelity, we next consider some extensions of this model that show how such effects could arise.

Mechanisms beyond the default model

In the following sections, we examine two departures from the default model and consider their impact on whether enhancers act additively (Eqn.19), super-additively (Eqn.31) or sub-additively (Eqn.33). This will also illustrate how our modeling framework can be used to reason about different biological mechanisms.

Non-additivity in mRNA production rates

In the default model, the summation of production rates in Assumption 3 is crucial for the property of enhancer additivity in Eqn.19. The production rate is a convenient abstraction that aggregates over many underlying molecular mechanisms, such as RNA Polymerase recruitment, pausing and elongation. It is

conceivable that, when multiple enhancers jointly influence transcription, the resulting rate is a more complex function than simple addition [38]. Here, we consider the effect of dropping Assumption 3.

Let us assume that we have two enhancers, e_1 and e_2, which are described by the graphs $H_1 = K_{a,1} \otimes K_{c,1}$ and $H_2 = K_{a,2} \otimes K_{c,2}$, respectively, as specified in Assumption 6 in the coarse-grained version of our default model. The overall regulatory graph is given by $H_1 \otimes H_2$, so that e_1 and e_2 remain independent (Assumption 2). Note that $(K_{a,1} \otimes K_{c,1}) \otimes (K_{a,2} \otimes K_{c,2})$ has a product hierarchy and its vertices are therefore indexed by tuples of tuples of the form,

$$((a_1, c_1), (a_2, c_2)) \,. \tag{35}$$

Here, a_i and c_i, for $i = 1, 2$, are coordinates for activation and communication, respectively, which take the values 0 and 1 in all cases. The graphs H_1 and H_2 have mRNA production rates, r_1 and r_2, respectively, as specified in Eqn.15 and mRNA degradation rate δ.

We now consider a copy-number graph, $G^\diamond$, whose regulatory graph is given by $(K_{a,1} \otimes K_{c,1}) \otimes (K_{a,2} \otimes K_{c,2})$ but whose production rates do not obey Assumption 3. Note that we use the same symbol, $G^\diamond$, for the regulatory graph and the copy-number graph and rely on the context to clarify which is meant. There are many ways to assign production rates to the vertices of $G^\diamond$ which do not obey Assumption 3; here we consider one of the simplest possible ways. We define 3 subsets of vertices of $G^\diamond$ in terms of the coordinates in Eqn.35: $W := \{a_1 = 1, c_1 = 1, a_2 = 1, c_2 = 1\}$, $U := \{a_1 = 1, c_1 = 1\} \setminus W$ and $V := \{a_2 = 1, c_2 = 1\} \setminus W$. We assign the production rate of vertices in U to be r_1, of vertices in V to be r_2 and of the vertex in W to be $(1 + \mu)(r_1 + r_2)$; all other vertices have production rate 0. We can summarise these assumptions in the following table, which gives the production rate for each of the 16 states in $G^\diamond$ in the coordinate system described by Eqn.35.

state	rate	state	rate
$((0,0),(0,0))$	0	$((1,0),(0,0))$	0
$((0,0),(0,1))$	0	$((1,0),(0,1))$	0
$((0,0),(1,0))$	0	$((1,0),(1,0))$	0
$((0,0),(1,1))$	r_2	$((1,0),(1,1))$	r_2
$((0,1),(0,0))$	0	$((1,1),(0,0))$	r_1
$((0,1),(0,1))$	0	$((1,1),(0,1))$	r_1
$((0,1),(1,0))$	0	$((1,1),(1,0))$	r_1
$((0,1),(1,1))$	r_2	$((1,1),(1,1))$	$(1+\mu)(r_1+r_2)$

We further assume that $\mu \geq -1$ to ensure that the production rate of W does not become negative. If $\mu = 0$, then Assumption 3 holds for $G^\diamond$ but not otherwise. We also assume that $G^\diamond$ has degradation rate δ.

We now calculate $R(G^\diamond)$. Using the Sanchez-Kondev theorem in Eqn.14 we have,

$$R(G^\diamond) = \frac{r_1 u^*_U(G^\diamond) + r_2 u^*_V(G^\diamond) + (1+\mu)(r_1+r_2) u^*_W(G^\diamond)}{\delta} \,. \tag{36}$$

Expanding the term on the right hand side of Eqn.36 and rearranging terms results in,

$$R(G^\diamond) = \frac{r_1 \left[u^*_U(G^\diamond) + u^*_W(G^\diamond)\right] + r_2 \left[u^*_V(G^\diamond) + u^*_W(G^\diamond)\right] + \mu(r_1+r_2) u^*_W(G^\diamond)}{\delta} \,. \tag{37}$$

Using the fact that the pairs of sets U and W, and, V and W are disjoint gives,

$$R(G^\diamond) = \frac{r_1 u^*_{U\cup W}(G^\diamond) + r_2 u^*_{V\cup W}(G^\diamond) + \mu(r_1+r_2) u^*_W(G^\diamond)}{\delta} \,. \tag{38}$$

We now note that, by definition, $U \cup W = \{a_1 = 1, c_1 = 1\}$ and $V \cup W = \{a_2 = 1, c_2 = 1\}$. Given the independence assumption on $G^\diamond$, we can apply Eqn.11 and have that $u^*_{\{a_1=1,c_1=1\}}(G^\diamond) = \tilde{\alpha}_1 \tilde{\gamma}_1$, $u^*_{\{a_2=1,c_2=1\}}(G^\diamond) = \tilde{\alpha}_2 \tilde{\gamma}_2$ and $u^*_W(G^\diamond) = \tilde{\alpha}_1 \tilde{\gamma}_1 \tilde{\alpha}_2 \tilde{\gamma}_2$. Substituting into Eqn.38, we have,

$$R(G^\diamond) = \frac{r_1 \tilde{\alpha}_1 \tilde{\gamma}_1 + r_2 \tilde{\alpha}_2 \tilde{\gamma}_2 + \mu(r_1+r_2) \tilde{\alpha}_1 \tilde{\gamma}_1 \tilde{\alpha}_2 \tilde{\gamma}_2}{\delta} \,. \tag{39}$$

Given that

$$R(H_1) + R(H_2) = \frac{r_1\tilde{\alpha}_1\tilde{\gamma}_1 + r_2\tilde{\alpha}_2\tilde{\gamma}_2}{\delta}\,, \tag{40}$$

we see that e_1 and e_2 act additively for $\mu = 0$ (Eqn.19), act super-additively for $\mu > 0$ (Eqn.31) and act sub-additively for $\mu \in [-1, 0)$ (Eqn.33).

Non-independence in regulatory transitions between enhancers

So far all the graphs we have considered obey regulatory independence as defined by Assumption 2: the state of one enhancer does not affect the transitions or the rates of any other enhancer. Let us examine this more closely for the IAC model, with just two enhancers, e_1 and e_2, described by graphs H_1 and H_2, respectively, as in the previous subsection. According to Assumption 6, $H_1 = K_{a,1} \otimes K_{c,1}$ and $H_2 = K_{a,2} \otimes K_{c,2}$ and the overall regulatory system is therefore described by the graph, $G = H_1 \otimes H_2 = (K_{a,1} \otimes K_{c,1}) \otimes (K_{a,2} \otimes K_{c,2})$. Using Eqn.12 and the coordinate system in Eqn.35, we have that $u^*_{\{a_1=1\}}(G) = u^*_{\{a_1=1\}}(H_1)$ and $u^*_{\{a_2=1\}}(G) = u^*_{\{a_2=1\}}(H_2)$. That is, the probability of activation of an enhancer does not depend on the presence of the other enhancer. However, if probability of activation is measured by H3K27ac ChIP-Seq, there is evidence that perturbation of a single enhancer can result in altered H3K27ac signal at distal enhancers [15, 22, 45, 46]. If such changes at the distal enhancer are not caused by the perturbation method itself, so that the perturbation obeys the fidelity conditions in Assumption 5, then such experiments suggest that there may be non-independence between enhancers at the level of activation.

In order to model non-independence between enhancers, we will consider a gene to be modeled by the graph $G^\sharp$, where $G^\sharp$ is the product between an activation graph $A^\sharp$ and a communication graph C, so that $G^\sharp = A^\sharp \otimes C$ (Fig.5). $A^\sharp$ has the structure $K_{a,1} \times K_{a,2}$, in which $K_{a,1}$ and $K_{a,2}$ are both present as the subgraphs,

$$(0,0) \underset{l_{a,1}}{\overset{k_{a,1}}{\rightleftharpoons}} (1,0) \quad \text{and} \quad (0,0) \underset{l_{a,2}}{\overset{k_{a,2}}{\rightleftharpoons}} (0,1)\,,$$

respectively (Fig.5a). The remaining labels, on the edges $(1,0) \rightleftharpoons (1,1)$, which specify the rates of activation and deactivation of e_2 when e_1 is activated, and on the edges $(0,1) \rightleftharpoons (1,1)$, which specify the rates of activation and deactivation of e_1 when e_2 is activated, can be arbitrary. For simplicity, we assume that $C = K_{c,1} \otimes K_{c,2}$, (Fig.5b), but note that non-independence in communication could be considered similarly. $G^\sharp$ has the same structure as $H_1 \times H_2$, and thus still models the activation and communication statuses of e_1 and e_2, but using the form $G^\sharp = A^\sharp \otimes C$ allows us to clarify the independence relationships in $G^\sharp$. Under this reorganization, the vertex in Eqn.35 is now described in a new coordinate system as,

$$((a_1, a_2), (c_1, c_2))\,. \tag{41}$$

We assume that $G^\sharp$ has the same mRNA production rates as for the IAC model. In terms of the vertex subsets $W = \{a_1 = 1, c_1 = 1, a_2 = 1, c_2 = 1\}$, $U = \{a_1 = 1, c_1 = 1\} \setminus W$ and $V = \{a_2 = 1, c_2 = 1\} \setminus W$, the vertices in U have production rate r_1, those in V have production r_2 and those in W have production rate $r_1 + r_2$; all other vertices have production rate 0. We can summarise this in the following table, in which the states are described by the coordinate system in Eqn.41.

state	rate	state	rate
$((0,0),(0,0))$	0	$((1,0),(0,0))$	0
$((0,0),(0,1))$	0	$((1,0),(0,1))$	0
$((0,0),(1,0))$	0	$((1,0),(1,0))$	r_1
$((0,0),(1,1))$	0	$((1,0),(1,1))$	r_1
$((0,1),(0,0))$	0	$((1,1),(0,0))$	0
$((0,1),(0,1))$	r_2	$((1,1),(0,1))$	r_2
$((0,1),(1,0))$	0	$((1,1),(1,0))$	r_1
$((0,1),(1,1))$	r_2	$((1,1),(1,1))$	$r_1 + r_2$

As in the previous section, we can use the Sanchez-Kondev theorem in Eqn.14 to calculate,

$$R(G^\sharp) = \frac{r_1 u^*_U(G^\sharp) + r_2 u^*_V(G^\sharp) + (r_1 + r_2) u^*_W(G^\sharp)}{\delta} \tag{42}$$

$$= \frac{r_1 u^*_{U\cup W}(G^\sharp) + r_2 u^*_{V\cup W}(G^\sharp)}{\delta} \tag{43}$$

$$= \frac{r_1 u^*_{\{a_1=1,c_1=1\}}(G^\sharp) + r_2 u^*_{\{a_2=1,c_2=1\}}(G^\sharp)}{\delta}\,. \tag{44}$$

Given that $G^\sharp = A^\sharp \otimes C$, the probabilities of the sets $\{a_1 = 1, c_1 = 1\}$ and $\{a_2 = 1, c_2 = 1\}$ factor according to Eqn.11. Continuing from Eqn.44 we have,

$$R(G^\sharp) = \frac{r_1 u^*_{\{a_1=1\}}(A^\sharp) u^*_{\{c_1=1\}}(C) + r_2 u^*_{\{a_2=1\}}(A^\sharp) u^*_{\{c_2=1\}}(C)}{\delta} \tag{45}$$

$$= \frac{r_1 u^*_{\{a_1=1\}}(A^\sharp)\tilde{\gamma}_1 + r_2 u^*_{\{a_2=1\}}(A^\sharp)\tilde{\gamma}_2}{\delta}\,. \tag{46}$$

Eqn.46 shows that the labels of $A^\sharp$ do not directly appear in $R(G^\sharp)$; they only affect $R(G^\sharp)$ through the enhancer activation probabilities $u^*_{\{a_1=1\}}(A^\sharp)$ and $u^*_{\{a_2=1\}}(A^\sharp)$. As in the previous section, we note that the sum of the individual enhancer responses is,

$$R(H_1) + R(H_2) = \frac{r_1\tilde{\alpha}_1\tilde{\gamma}_1 + r_2\tilde{\alpha}_2\tilde{\gamma}_2}{\delta}\,. \tag{47}$$

Comparing Eqns.46 and 47, we see that whether the enhancers act additively (Eqn.19), sub-additively (Eqn.33) or super-additively (Eqn.31) depends on the terms

$$\tilde{\alpha}_1^+ := \left[u^*_{\{a_1=1\}}(A^\sharp) - \tilde{\alpha}_1\right] \quad \text{and} \quad \tilde{\alpha}_2^+ := \left[u^*_{\{a_2=1\}}(A^\sharp) - \tilde{\alpha}_2\right]\,. \tag{48}$$

$\tilde{\alpha}_1^+$ represents the change in the probability of activation of enhancer 1 due to the presence of enhancer 2, and $\tilde{\alpha}_2^+$ represents the change in the probability of activation of enhancer 2 due to the presence of enhancer 1. If both $\tilde{\alpha}_1^+$ and $\tilde{\alpha}_2^+$ are positive, then e_1 and e_2 act super-additively; if they are both negative, then e_1 and e_2 act sub-additively. If $\tilde{\alpha}_1^+$ and $\tilde{\alpha}_2^+$ are of different signs, then the enhancers may act super-additively or sub-additively depending on the relative magnitude of these terms compared to $\tilde{\gamma}_1, \tilde{\gamma}_2, r_1$ and r_2. Experimental data in which both $\tilde{\alpha}_1^+$ and $\tilde{\alpha}_2^+$ have been measured is limited. There are experimentally observed instances in which both of these terms are positive [15, 45] but the precise form that the graph $A^\sharp$ takes in these cases is unknown. We are unaware of experiments that have observed $\tilde{\alpha}_1^+$ and $\tilde{\alpha}_2^+$ of differing signs. Whether the experimentally observed non-additivity between enhancers can be explained by the non-independence between enhancers as described in this section is a future area of research.

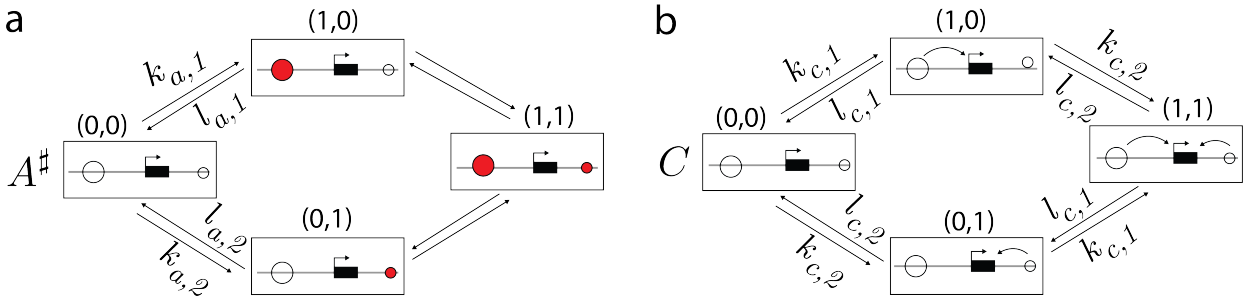

Figure 5: A model of non-independence in activation between enhancers. **(a)** The graph $A^\sharp$ represents the activation components of each enhancer. $A^\sharp$ has the structure of $K_{a,1} \times K_{a,2}$. Labels on the unmarked edges can be arbitrary. **(b)** The graph $C = K_{c,1} \otimes K_{c,2}$ represents the communication components of each enhancer.

Discussion

In this paper, we have introduced mathematical formulations of a *default model* and an *Independent-Activation-Communication model* (IAC model) for how multiple enhancers collectively regulate a gene. The default model encodes the notion that enhancers operate independently of each other (Assumptions 2 and 3). At the same time, the default model imposes no assumptions on how the individual enhancers themselves are working, at the level of transcription factors, co-regulators, chromatin, etc. They can be arbitrarily complicated, so long as they operate within the Markovian setting that is commonly assumed for analysing gene regulation. The default model assumptions imply that the collective response of a gene, as measured by the mean mRNA level, is the sum of the responses coming from each enhancer individually, which we have called *response additivity* (Eqn.19). The default model explains the mechanistic requirements for a gene to exhibit this property and clarifies how 'independence implies additivity'. We emphasize that independence refers here to assumptions about gene regulatory mechanisms whereas additivity refers to the consequences of those assumptions on steady-state gene expression. The default model supports the view that response additivity is a reasonable baseline against which to assess the collective action of enhancers in regulating a gene.

One of the advantages of the default model is that, because it is mathematically formulated, it allows mechanistic departures from its assumptions to be systematically analysed. We have shown how departures from Assumptions 2 and 3 can give rise to response super-additivity (Eqn.32) as well as sub-additivity (Eqn.34). As we have noted, response additivity [15, 23, 38–43], super-additivity [15, 21, 39–44] and sub-additivity [38, 40, 41] have all been observed experimentally. The default model suggests the mechanistic assumptions that could be experimentally tested to determine what underlies the observed response behaviour.

The IAC model is a special case of the default model that further assumes deletion fidelity, which allows enhancers to be removed from the collective without influencing the remaining enhancers (Assumption 5), and also assumes that individual enhancers can be described at a coarse-grained level in which they are independently becoming activated and communicating their state to the gene (Assumption 6). Under Assumptions 1 to 6 for the IAC model, we derive a formula for the deletion effect of an individual enhancer (Eqn.27) that shows a striking algebraic relationship to the ABC Score formula in Eqn.2. This relationship suggests that the IAC model has accurately captured in mathematical terms the core intuitions behind the ABC model from which the Score formula emerged [11].

A persistent conceptual theme that underlies the results reported here is that of independence. The default model assumes that enhancers act independently, both in their regulatory state (Assumption 2) and in their effect on mRNA production (Assumption 3). Furthermore the IAC model assumes that individual enhancers become activated and communicating independently (Assumption 6). Our clarification of the ABC Score formula thus arises from assuming independence *between* enhancers, along with independence of activation and communication *within* each enhancer. The product construction on graphs and on graph structures has been the key mathematical tool for rigorously defining independence, illustrating the value of the graph-based linear framework for analysing gene regulation. We note that the concept of deletion fidelity (Assumption 5) is also easily defined in the context of graphs.

Previous work used finite linear framework graphs to describe gene regulation [26]. Here, we have introduced *copy-number graphs*, which have infinitely many vertices that keep track of both regulatory states as well as the numbers of expressed mRNAs (Fig.2). Copy-number graphs allowed us to exploit the Sanchez-Kondev theorem and calculate the mean mRNA number at steady state in terms solely of the finite regulatory graph (Eqn.14). We therefore avoided dealing with infinite graphs despite relying on them. Importantly, the linear framework also allows the unknown parameters within graphs to be treated symbolically, so that conclusions may be drawn, as we saw above, without the need for assigning numerical values to any of the parameters.

Beyond its utility in clarifying the ABC Score formula, the activation-communication coarse graining in Assumption 6 provides an interesting lens through which to investigate enhancers. Many new experimental technologies have emerged which allow perturbing entire enhancer sequences as a whole (as opposed to small changes to DNA within an enhancer sequence). Such technologies include synthesizing and integrating long

DNA sequences [43, 44], modulating the genomic position of an enhancer [47–51], high throughput enhancer perturbations with CRISPRi [11–13] and combining CRISPRi screens with rapid protein degradation [52]. By considering which perturbations affect, and do not affect, activation and communication, it may be possible to probe the validity of the activation-communication coarse graining itself.

As noted in the Introduction, the ABC Score formula has been widely adopted for predicting enhancer-gene connections. It has also been suggested that it could be combined with other predictive methods [53] and that the ABC model could be used as a guiding principle in formulating other quantitative models [54]. We believe the mathematical formulations that we have introduced here provide a foundation for such efforts.

The ABC Score formula is quantitative (Eq.2) but the ABC model that gave rise to it is not a formal mathematical model but, rather, an informal statement about the features, of enhancer activation and contact, that are believed to be important in determining the response of a gene. Such informal models play a critical role in biology but have the disadvantage that the underlying mechanistic requirements are not clear. It is therefore difficult to know when the model can be applied and what can be deduced from it when it does apply. In contrast, the mechanistic assumptions underlying our formal mathematical models are precisely stated—Assumptions 1 to 4 for the default model and Assumptions 1 to 6 for the IAC model—making it clear when the model can be applied and suggesting experimental tests to check the assumptions. Moreover, if those assumptions are met, then the conclusions we have drawn, such as the response additivity of the default model (Eqn.19) and the enhancer deletion formula for the IAC model (Eqn.27), are guaranteed to hold as a matter of mathematical logic [55]. If those conclusions are not found experimentally, for example, if response additivity is not found, then we know, as a matter of logic, that at least one of the assumptions underlying the corresponding model does not hold. This understanding can, in turn, inform experiments to determine where the departures from the assumptions occur. Such an approach allows a level of rigorous reasoning about enhancer behaviour in gene regulation that is significantly harder to undertake with only an informal quantitative model.

Mathematical theory has typically been introduced to analyse data, but the conceptual issues underlying gene regulation are sufficiently intricate that theory may be necessary to understand the kinds of experiments that are needed and how the data from them should best be interpreted [56]. Studies of the simple repression motif in bacterial gene regulation may have already reached that point [57–60], as reviewed in [61]. The foundation provided here, based on the linear framework, may offer similar opportunities in the eukaryotic context. We believe our rigorous mathematical approach can play a significant role in investigating the intricate interplay of enhancers in regulating gene expression.

Methods

A graph theory interpretation of the Sanchez and Kondev theorem

In this section we provide a proof of Eqn.14. We follow the Sanchez and Kondev approach described in [34] but present it using the graph theory notation and language used in this paper. Sanchez and Kondev provide in [34] a recurrence relation for all the moments of the mRNA probability distribution. A graph theory interpretation of these results, together with generalisations, will be presented in a separate paper; here we focus on the first moment only. We use bold face to denote matrices and vectors.

The steady-state distribution of a finite graph

Let G be a finite regulatory graph on the vertex set $V(G) = \{1, \ldots, N\}$. We define the *Laplacian matrix* of G, $\mathcal{L} = \mathcal{L}(G)$, to be the $N \times N$ matrix,

$$\mathcal{L}(G)_{i,j} := \begin{cases} 0 & \text{if } j \neq i \text{ and } j \not\to i \\ \ell(j \to i) & \text{if } j \neq i \text{ and } j \to i \\ -\sum_{k \,|\, i \to k} \ell(i \to k) & \text{if } i = j. \end{cases} \tag{49}$$

As mentioned in the main text, G is equivalent to a continuous-time Markov process on the state space $\{1, \ldots, N\}$ [25, 28, 32]. Let $u_i(t)$ be the probability that the process occupies state i at time t. Then the time evolution of the probability vector,

$$\mathbf{u}(t) := (u_1(t), \ldots, u_N(t))^T ,$$

is given by the master equation

$$\frac{d\mathbf{u}(t)}{dt} = \mathcal{L}(G)\mathbf{u}(t) . \tag{50}$$

If G is strongly connected, then the kernel of $\mathcal{L}(G)$ is one dimensional, so there is a unique vector, $\mathbf{u}^*(G)$, such that $\mathcal{L}(G)\mathbf{u}^*(G) = 0$ and $u_1(G) + \cdots + u_N(G) = 1$. $\mathbf{u}^*(G)$ is the steady-state probability distribution on G.

The master equation for a copy-number graph

Let $G \ltimes P$ denote a copy-number graph with regulatory graph G, production rate vector $\mathbf{r} \in \mathbb{R}^N$ and degradation rate δ. Let $\mathbf{\Pi}$ be the diagonal matrix of production rates, $\mathbf{\Pi}_{i,i} = r_i$ and $\mathbf{\Pi}_{i,j} = 0$ when $i \neq j$, and let $\mathbf{I}$ be the $N \times N$ identity matrix. Let

$$\mathbf{u}(p,t) := \big(u_{(1,p)}(G \ltimes P; t), \ldots, u_{(N,p)}(G \ltimes P; t)\big)^T ,$$

be the vector of probabilities over the regulatory states with mRNA copy number p. It follows from the definition of the copy-number graph in the main text that $\mathbf{u}(p,t)$ satisfies the master equation,

$$\frac{d}{dt}\mathbf{u}(p,t) = \mathbf{\Pi}[\mathbf{u}(p-1,t) - \mathbf{u}(p,t)] + \delta\mathbf{I}[(p+1)\mathbf{u}(p+1,t) - p\mathbf{u}(p,t)] + \mathcal{L}(G)\mathbf{u}(p,t) , \tag{51}$$

in which terms with arguments of $p-1$ are appropriately omitted when $p = 0$. The first term of Eqn.51 arises from mRNA production, the second term from mRNA degradation and the third term from transitions in the regulatory graph. Eqn.51 can be rewritten as,

$$\frac{d}{dt}\mathbf{u}(p,t) = \mathbf{\Pi}\mathbf{u}(p-1,t) + \delta(p+1)\mathbf{u}(p+1,t) - [\mathbf{\Pi} + p\delta\mathbf{I} - \mathcal{L}(G)]\mathbf{u}(p,t) . \tag{52}$$

We now let

$$\mathbf{q}(t) := \sum_{p=0}^{\infty} \mathbf{u}(p,t)$$

be the vector of marginal probabilities for the regulatory states. Proceeding from Eqn.52 we have,

$$\begin{aligned}\frac{d}{dt}\mathbf{q}(t) &= \sum_{p=0}^{\infty}\frac{d}{dt}\mathbf{u}(p,t)\\ &= \underbrace{\delta\mathbf{u}(1,t) - \mathbf{\Pi}\,u(0,t) + \mathcal{L}(G)\,\mathbf{u}(0,t)}_{p=0} +\\ &\quad \underbrace{\mathbf{\Pi}\,\mathbf{u}(0,t) + 2\delta\mathbf{u}(2,t) - \mathbf{\Pi}\,\mathbf{u}(1,t) - \delta\mathbf{u}(1,t) + \mathcal{L}(G)\,\mathbf{u}(1,t)}_{p=1} +\\ &\quad \underbrace{\mathbf{\Pi}\,\mathbf{u}(1,t) + 3\delta\mathbf{u}(3,t) - \mathbf{\Pi}\,\mathbf{u}(2,t) - 2\delta\mathbf{u}(2,t) + \mathcal{L}(G)\,\mathbf{u}(2,t)}_{p=2} + \dots\,.\end{aligned}$$

This is a telescoping sum which simplifies to

$$\frac{d}{dt}\mathbf{q}(t) = \mathcal{L}(G)(\mathbf{u}(0,t) + \mathbf{u}(1,t) + \mathbf{u}(2,t) + \dots) = \mathcal{L}(G)\,\mathbf{q}(t).$$

It follows that the steady-state marginal probability vector, $\mathbf{q}^*$, lies in the kernel of $\mathcal{L}(G)$ and must therefore be equal to $\mathbf{u}^*(G)$,

$$\mathbf{q}^* = \mathbf{u}^*(G)\,. \tag{53}$$

In other words, the steady-state marginal distribution of regulatory states in a copy-number graph is identical to the steady-state distribution of regulatory states in a finite regulatory graph.

Proof of Eqn.14

We want to show that,

$$R(G \ltimes P) = \sum_{(i,p)} p \cdot u^*_{(i,p)}(G \ltimes P) = \frac{1}{\delta}\sum_{i\in V(G)} r_i \cdot u^*_i(G)\,.$$

Let $\mathbf{u}^*(p)$ denote the steady-state probability distribution over the copy-number graph. (Note the distinction with the marginal probability distribution over the regulatory states, $u^*(G) = \sum_p u^*(p)$.) Let

$$\boldsymbol{\mu}^* := \sum_{p=1}^{\infty} p\,\mathbf{u}^*(p)$$

be the corresponding steady-state average copy number vector. Evidently,

$$R(G \ltimes P) = \mathbf{1}^{\mathrm{T}}\boldsymbol{\mu}^*\,, \tag{54}$$

where $\mathbf{1}$ is the all-ones column vector of dimension N. Now let

$$\boldsymbol{\mu}(t) := \sum_{p=1}^{\infty} p\mathbf{u}(p,t)$$

be the time-dependent average copy-number vector. It follows from Eqn.52 that,

$$\frac{d}{dt}\boldsymbol{\mu}(t) = \sum_{p=1}^{\infty} p\frac{d}{dt}\mathbf{u}(p,t)$$

$$= \sum_{p=1}^{\infty} p\Big(\mathbf{\Pi}\,\mathbf{u}(p-1,t) + (p+1)\,\delta\mathbf{u}(p+1,t) - (\mathbf{\Pi} + p\delta\mathbf{I} - \mathcal{L}(G))\,\mathbf{u}(p,t)\Big)$$

$$= \sum_{p=1}^{\infty} p\,\mathbf{\Pi}\left(\mathbf{u}(p-1,t)-\mathbf{u}(p,t)\right)+\sum_{p=1}^{\infty} p\left((p+1)\,\delta\mathbf{u}(p+1,t)-p\delta\mathbf{u}(p,t)\right)+\sum_{p=1}^{\infty} p\mathcal{L}(G)\,\mathbf{u}(p,t) \tag{55}$$

The first summand in Eqn.55 can be simplified to,

$$\mathbf{\Pi}\left((\mathbf{u}(0,t)-\mathbf{u}(1,t))+2\left(\mathbf{u}(1,t)-\mathbf{u}(2,t)\right)+\cdots\right)=\mathbf{\Pi}\left(\mathbf{u}(0,t)+\mathbf{u}(1,t)+\cdots\right)=\mathbf{\Pi}\mathbf{q}(t)\,.$$

The second summand can be simplified to,

$$\delta\left((2\mathbf{u}(2,t)-\mathbf{u}(1,t))+2\left(3\mathbf{u}(3,t)-2\mathbf{u}(2,t)\right)+\cdots\right)=-\delta\sum_{p=1}^{\infty} p\mathbf{u}(p,t)=-\delta\boldsymbol{\mu}(t)\,.$$

And the third summand is evidently just $\mathcal{L}(G)\,\boldsymbol{\mu}(t)$. Combining these three simplifications, we see that,

$$\frac{d}{dt}\boldsymbol{\mu}(t)=\mathbf{\Pi}\,\mathbf{q}(t)-\delta\boldsymbol{\mu}(t)+\mathcal{L}(G)\,\boldsymbol{\mu}(t)\,. \tag{56}$$

At steady state this becomes,

$$\delta\boldsymbol{\mu}^*-\mathcal{L}(G)\,\boldsymbol{\mu}^*=\mathbf{\Pi}\,\mathbf{q}^*.$$

Multiplying both sides $\mathbf{1}^{\mathrm{T}}$, and recalling that,

$$\mathbf{1}^{\mathrm{T}}\mathcal{L}(G)=\mathbf{0}^{\mathrm{T}}\quad\text{and}\quad\mathbf{1}^{\mathrm{T}}\mathbf{\Pi}=\mathbf{r}^{\mathrm{T}}\,,$$

we find that,

$$\mathbf{1}^T\boldsymbol{\mu}^*=\frac{\mathbf{r}^T\mathbf{q}^*}{\delta}\,. \tag{57}$$

Using Eqns.53 and 54, we see that Eqn.57 becomes,

$$R(G\ltimes P)=\frac{\mathbf{r}^T\mathbf{u}^*(G)}{\delta} \tag{58}$$

$$=\frac{1}{\delta}\sum_{i\in V(G)} r_i\cdot u_i^*(G)\,. \tag{59}$$

as required. This completes the proof of Eqn.14.

Proof of response summation in the default model

In this section we prove Eqn.19 which shows that, within the default model, the collective response of all the enhancers is the sum of their individual responses. That is, if $G=G_1\circledast\cdots\circledast G_N$, then

$$R(G)=R(G_1)+\cdots+R(G_N)\,. \tag{60}$$

We consider the case with only two enhancers, $N=2$, from which the general case follows easily. Recall from the Sanchez and Kondev formula in Eqn.14 that

$$R(G)=\frac{1}{\delta}\sum_{(i_1,i_2)} r_{(i_1,i_2)}(G)\cdot u^*_{(i_1,i_2)}(G)\,. \tag{61}$$

Assumption 3 on response summation tells us that $r_{(i_1,i_2)}(G)=r_{i_1}(G_1)+r_{i_2}(G_2)$ and Eqn.8 for the product graph tells us that $u^*_{(i_1,i_2)}(G)=u^*_{i_1}(G_1)\cdot u^*_{i_2}(G_2)$. Substituting these expressions into Eqn.61 gives the following formula for $\delta\cdot R(G)$,

$$\sum_{(i_1,i_2)}\left(r_{i_1}(G_1)+r_{i_2}(G_2)\right)\cdot u^*_{i_1}(G_1)\cdot u^*_{i_2}(G_2)\,.$$

We can perform the summation over (i_1, i_2) in any order, for instance by first summing over i_2 and then summing over i_1. This gives,

$$\sum_{i_1} \left(\sum_{i_2} r_{i_1}(G_1) \cdot u^*_{i_1}(G_1) \cdot u^*_{i_2}(G_2) + \sum_{i_2} r_{i_2}(G_2) \cdot u^*_{i_1}(G_1) \cdot u^*_{i_2}(G_2) \right) . \tag{62}$$

In the inner left-hand sum over i_2, the terms indexed by i_1 are constant and may be extracted from that sum to give

$$r_{i_1}(G_1) \cdot u^*_{i_1}(G_1) \cdot \left(\sum_{i_2} u^*_{i_2}(G_2) \right) . \tag{63}$$

Total probability always sums to 1, so that $\sum_{i_2} u^*_{i_2}(G_2) = 1$, and Eqn.63 reduces to

$$r_{i_1}(G_1) \cdot u^*_{i_1}(G_1) . \tag{64}$$

Similarly, the inner right-hand sum in Eqn.62 may be written as

$$u^*_{i_1}(G_1) \cdot \left(\sum_{i_2} r_{i_2}(G_2) \cdot u^*_{i_2}(G_2) \right) . \tag{65}$$

We recognise from Eqn.14 that the sum in brackets is the response of graph G_2, so that Eqn.65 becomes,

$$u^*_{i_1}(G_1) \cdot \delta \cdot R(G_2) . \tag{66}$$

We can now substitute Eqns.64 and 66 back into Eqn.62 to get,

$$\sum_{i_1} r_{i_1}(G_1) \cdot u^*_{i_1}(G_1) + \sum_{i_1} u^*_{i_1}(G_1) \cdot \delta \cdot R(G_2) . \tag{67}$$

We recognise from Eqn.14 that the left-hand sum is δ times the response of graph G_1. In the right-hand sum, we can extract the terms that do not depend on i_1 and use once again that the total probability is 1. This allows us to rewrite Eqn.67 as,

$$\delta \cdot R(G_1) + \delta \cdot R(G_2) ,$$

from which we conclude that, indeed,

$$R(G) = R(G_1) + R(G_2) ,$$

as claimed. This completes the proof of Eqn.19.

Symbolic computations

We have provided mathematical proofs for all of our results. However, many of our results were originally discovered by exploration using computer algebra systems. Specifically, the Sage [62] computer algebra system, and the SymPy [63] and NetworkX [64] Python packages were crucial for the development of this paper.

Acknowledgements

JN, K-MN and JG were funded in part by NIH award R01GM122928; JN was also funded the NSF-Simons Center for Mathematical and Statistical Analysis of Biology at Harvard University. We thank Zeba Wunderlich, Rosa Martinez-Corral, Jané Kondev and members of the Gunawardena lab for discussions and comments on the manuscript.

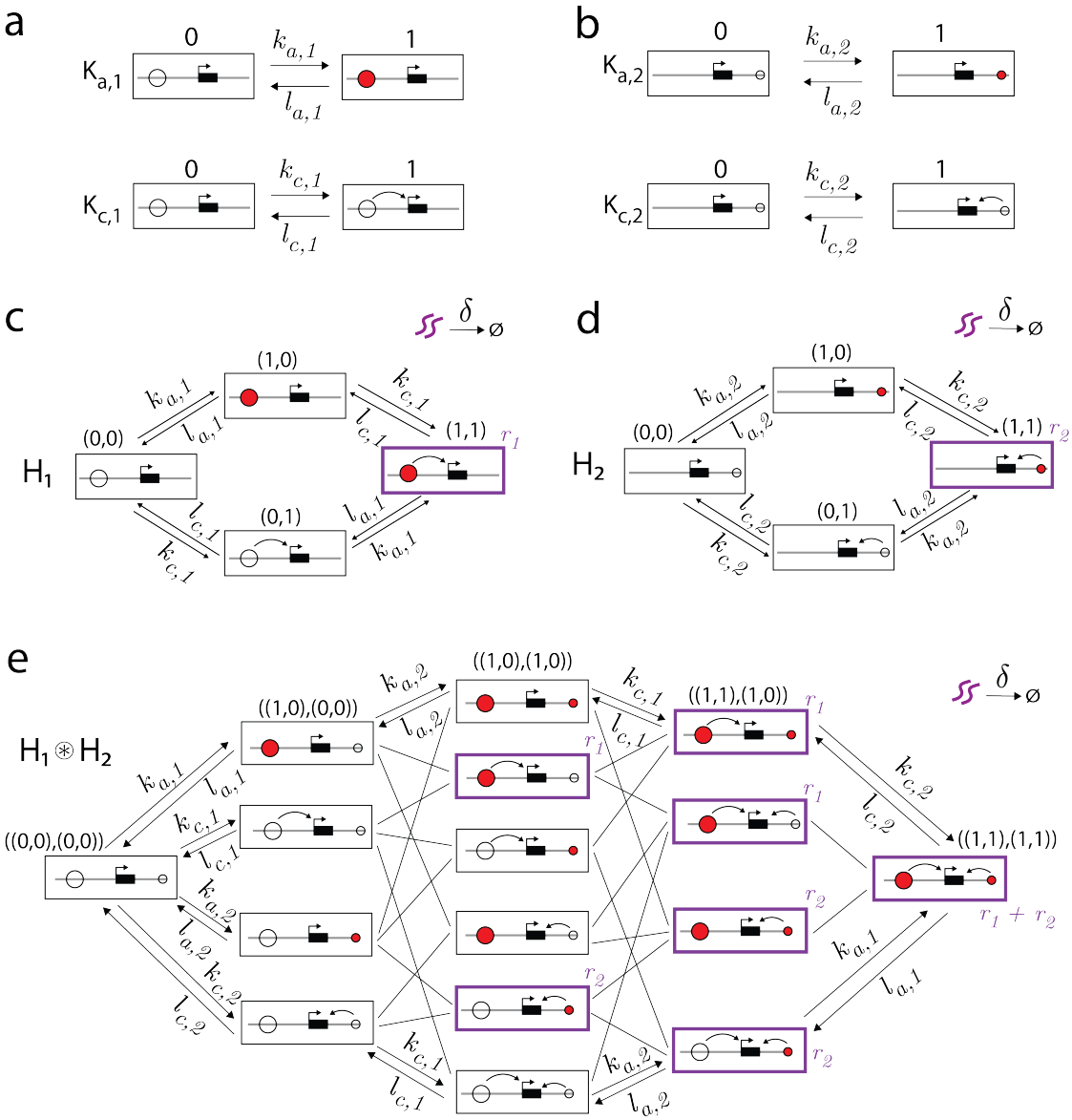

Figure S1: IAC model for $N = 2$ enhancers as a product graph of product graphs. **(a)** Graphs describing the activation and communication status of enhancer 1. **(b)** Graphs describing the activation and communication status of enhancer 2. **(c)** The graph H_1 whose underlying regulatory graph is $K_{a,1} \otimes K_{c,1}$. **(d)** The graph H_2 whose underlying regulatory graph is $K_{a,2} \otimes K_{c,2}$. **(e)** The graph $H_1 \circledast H_2$ satisfying the default model Assumptions 1-4 with components H_1 and H_2. The regulatory graph of $H_1 \circledast H_2$ is given by $(K_{a,1} \otimes K_{c,1}) \otimes (K_{a,2} \otimes K_{c,2})$. Each vertex of this graph corresponds to the activation and communication statuses of both enhancers. For the vertices along the top of the graph, the product-graph binary notation is also provided using the coordinate system $((a_1, c_1), (a_2, c_2))$. Reverse edges and labels of most edges are omitted for clarity. Production states are highlighted in purple with corresponding production rates also in purple font.

References

1. Ptashne, M. & Gann, A. *Genes and Signals* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, USA, 2002).
2. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27,** 299–308 (1981).
3. Moreau, P., Hen, R., Wasylyk, B., Everett, R., Gaub, M. & Chambon, P. The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic Acids Research* **9,** 6047–6068 (1981).
4. Levine, M. Transcriptional enhancers in animal development and evolution. *Current biology : CB* **20,** R754–R763 (2010).
5. Lempradl, A., Pospisilik, A. & Penninger, J. M. Exploring the emerging complexity in transcriptional regulation of energy homeostasis. *Nat. Reviews. Genet.* **16,** 665–81 (2015).
6. Indjeian, V. B., Kingman, G. A., Jones, F. C., Guenther, C. A., Grimwood, J., Schmutz, J., Myers, R. M. & Kingsley, D. M. Evolving new skeletal traits by cis-regulatory changes in bone morphogenetic proteins. *Cell* **164,** 45–56 (2016).
7. Corradin, O. & Scacheri, P. C. Enhancer variants: evaluating functions in common disease. *Genome Medicine* **6,** 85 (2014).
8. Noonan, J. P. & McCallion, A. S. Genomics of long-range regulatory elements. *Annual Review of Genomics and Human Genetics* **11,** 1–23 (2010).
9. Zhang, X. & Meyerson, M. Illuminating the noncoding genome in cancer. *Nature Cancer* **1,** 864–872 (2020).
10. Fulco, C. P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S. R., Perez, E. M., Kane, M., Cleary, B., Lander, E. S. & Engreitz, J. M. Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science* **354,** 769–773 (2016).
11. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature Genetics* **51,** 1664–1669 (2019).
12. Schraivogel, D., Gschwind, A. R., Milbank, J. H., Leonce, D. R., Jakob, P., Mathur, L., Korbel, J. O., Merten, C. A., Velten, L. & Steinmetz, L. M. Targeted perturb-seq enables genome-scale genetic screens in single cells. *Nature Methods* **17,** 629–635 (2020).
13. Reilly, S. K. *et al.* Direct characterization of cis-regulatory elements and functional dissection of complex genetic associations using HCR–FlowFISH. *Nature Genetics* **53,** 1166–1176 (2021).
14. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593,** 238–243 (2021).
15. Gschwind, A. R. *et al.* An encyclopedia of enhancer-gene regulatory interactions in the human genome. *bioRxiv* (2023).
16. Bick, A. G. *et al.* Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* **586,** 763–768 (2020).
17. Xu, Z. *et al.* Structural variants drive context-dependent oncogene activation in cancer. *Nature* **612,** 564–572 (2022).
18. Bendl, J. *et al.* The three-dimensional landscape of cortical chromatin accessibility in Alzheimer's disease. *Nature Neuroscience* **25,** 1366–1378 (2022).
19. Chen, Q., Dai, J. & Bian, Q. Integration of 3D genome topology and local chromatin features uncovers enhancers underlying craniofacial-specific cartilage defects. *Science Advances* **8,** eabo3648 (2022).
20. Naqvi, S. *et al.* Precise modulation of transcription factor levels identifies features underlying dosage sensitivity. *Nature Genetics,* 1–11 (2023).

21. Shin, H. Y., Willi, M., Yoo, K. H., Zeng, X., Wang, C., Metser, G. & Hennighausen, L. Hierarchy within the mammary STAT5-driven Wap super-enhancer. *Nature Genetics* **48,** 904–911 (2016).

22. Li, K. *et al.* Interrogation of enhancer function by enhancer-targeting CRISPR epigenetic editing. *Nature Communications* **11,** 485 (2020).

23. Hay, D. *et al.* Genetic dissection of the α-globin super-enhancer in vivo. *Nature Genetics* **48,** 895–903 (2016).

24. Gunawardena, J. A linear framework for time-scale separation in nonlinear biochemical systems. *PLOS ONE* **7,** e36321 (2012).

25. Mirzaev, I. & Gunawardena, J. Laplacian dynamics on general graphs. *Bulletin of Mathematical Biology* **75,** 2118–2149 (2013).

26. Ahsendorf, T., Wong, F., Eils, R. & Gunawardena, J. A framework for modelling gene regulation which accommodates non-equilibrium mechanisms. *BMC Biology* **12,** 102 (2014).

27. Nam, K.-M., Martinez-Corral, R. & Gunawardena, J. The linear framework: using graph theory to reveal the algebra and thermodynamics of biomolecular systems. *Interface Focus* **12,** 20220013 (2022).

28. Nam, K.-M. & Gunawardena, J. The linear framework II: using graph theory to analyse the transient regime of markov processes. *Front. Cell Dev. Biol* **11,** 1233808 (2023).

29. Popay, T. M. & Dixon, J. R. Coming full circle: on the origin and evolution of the looping model for enhancer–promoter communication. *Journal of Biological Chemistry* **298** (2022).

30. Karr, J. P., Ferrie, J. J., Tjian, R. & Darzacq, X. The transcription factor activity gradient (TAG) model: contemplating a contact-independent mechanism for enhancer–promoter communication. *Genes & Development* (2021).

31. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A phase separation model for transcriptional control. *Cell* **169,** 13–23 (2017).

32. Norris, J. R. *Markov Chains* (Cambridge University Press, Cambridge, UK, 1997).

33. Peccoud, J. & Ycart, B. Markovian modeling of gene-product synthesis. *Theoretical Population Biology* **48,** 222–234 (1995).

34. Sánchez, Á. & Kondev, J. Transcriptional control of noise in gene expression. *Proceedings of the National Academy of Sciences* **105,** 5081–5086 (2008).

35. Dukler, N., Gulko, B., Huang, Y.-F. & Siepel, A. Is a super-enhancer greater than the sum of its parts? *Nature Genetics* **49,** 2–3 (2017).

36. Van Arensbergen, J., FitzPatrick, V. D., de Haas, M., Pagie, L., Sluimer, J., Bussemaker, H. J. & van Steensel, B. Genome-wide mapping of autonomous promoter activity in human cells. *Nature Biotechnology* **35,** 145–153 (2017).

37. Weingarten-Gabbay, S., Nir, R., Lubliner, S., Sharon, E., Kalma, Y., Weinberger, A. & Segal, E. Systematic interrogation of human promoters. *Genome Research* **29,** 171–183 (2019).

38. Martinez-Ara, M., Comoglio, F. & van Steensel, B. Large-scale analysis of the integration of enhancer-enhancer signals by promoters. *eLife* **12,** RP91994 (2024).

39. Loubiere, V., Almeida, B. P. d., Pagani, M. & Stark, A. Developmental and housekeeping transcriptional programs display distinct modes of enhancer-enhancer cooperativity in drosophila. *bioRxiv* (2023).

40. Bothma, J. P., Garcia, H. G., Ng, S., Perry, M. W., Gregor, T. & Levine, M. Enhancer additivity and non-additivity are determined by enhancer strength in the Drosophila embryo. *eLife* **4,** e07956 (2015).

41. Scholes, C., Biette, K. M., Harden, T. T. & DePace, A. H. Signal Integration by Shadow Enhancers and Enhancer Duplications Varies across the Drosophila Embryo. *Cell Reports* **26,** 2407–2418.e5 (2019).

42. Lam, D. D. *et al.* Partially redundant enhancers cooperatively maintain mammalian pomc expression above a critical functional threshold. *PLOS Genetics* **11,** e1004935 (2015).

43. Brosh, R. *et al.* Synthetic regulatory genomics uncovers enhancer context dependence at the Sox2 locus. *bioRxiv* (2022).
44. Blayney, J. W. *et al.* Super-enhancers include classical enhancers and facilitators to fully activate gene expression. *Cell* **186,** 5826–5839.e18 (2023).
45. Thomas, H. F. *et al.* Temporal dissection of an enhancer cluster reveals distinct temporal and functional contributions of individual elements. *Molecular Cell* **81,** 969–982.e13 (2021).
46. Mansour, M. R. *et al.* An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346,** 1373–1377 (2014).
47. Zuin, J. *et al.* Nonlinear control of transcription through enhancer–promoter interactions. *Nature,* 1–7 (2022).
48. Rinzema, N. J. *et al.* Building regulatory landscapes reveals that an enhancer can recruit cohesin to create contact domains, engage CTCF sites and activate distant genes. *Nature Structural & Molecular Biology* **29,** 563–574 (2022).
49. Thomas, H., Feng, S., Huber, M., Loubiere, V., Vanina, D., Pitasi, M., Stark, A. & Buecker, C. Enhancer cooperativity can compensate for loss of activity over large genomic distances. *bioRxiv* (2023).
50. Jensen, C. L., Chen, L.-F., Swigut, T., Crocker, O. J., Yao, D., Bassik, M. C., Ferrell, J., Boettiger, A. & Wysocka, J. Long range regulation of transcription scales with genomic distance in a gene specific manner. *bioRxiv* (2024).
51. Brückner, D. B., Chen, H., Barinov, L., Zoller, B. & Gregor, T. Stochastic motion and transcriptional dynamics of pairs of distal DNA loci on a compacted chromosome. *Science* **380,** 1357–1362 (2023).
52. Guckelberger, P. *et al.* Cohesin-mediated 3D contacts tune enhancer-promoter regulation. *bioRxiv* (2024).
53. De Almeida, B. P., Reiter, F., Pagani, M. & Stark, A. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nature Genetics* **54,** 613–624 (2022).
54. Kim, S. & Wysocka, J. Deciphering the multi-scale, quantitative cis-regulatory code. *Molecular Cell. Reimagining the Central Dogma* **83,** 373–392 (2023).
55. Gunawardena, J. Models in biology: 'accurate descriptions of our pathetic thinking'. *BMC Biol.* **12,** 29 (2014).
56. Phillips, R. Theory in Biology: Figure 1 or Figure 7? *Trends in Cell Biology* **25,** 723–729 (2015).
57. Garcia, H. G. & Phillips, R. Quantitative dissection of the simple repression input–output function. *Proceedings of the National Academy of Sciences* **108,** 12173–12178 (2011).
58. Jones, D. L., Brewster, R. C. & Phillips, R. Promoter architecture dictates cell-to-cell variability in gene expression. *Science* **346,** 1533–1536 (2014).
59. Brewster, R. C., Weinert, F. M., Garcia, H. G., Song, D., Rydenfelt, M. & Phillips, R. The transcription factor titration effect dictates level of gene expression. *Cell* **156,** 1312–1323 (2014).
60. Razo-Mejia, M., Barnes, S. L., Belliveau, N. M., Chure, G., Einav, T., Lewis, M. & Phillips, R. Tuning transcriptional regulation through signaling: a predictive theory of allosteric induction. *Cell Systems* **6,** 456–469.e10 (2018).
61. Phillips, R., Belliveau, N. M., Chure, G., Garcia, H. G., Razo-Mejia, M. & Scholes, C. Figure 1 theory meets figure 2 experiments in the study of gene expression. *Annual Review of Biophysics* **48,** 121–163 (2019).
62. The Sage Developers. *SageMath, the Sage Mathematics Software System (Version 9.6)* (2022).
63. Meurer, A. *et al.* SymPy: symbolic computing in Python. *PeerJ Computer Science* **3,** e103 (2017).
64. Hagberg, A. A., Schult, D. A. & Swart, P. J. *Exploring network structure, dynamics, and function using NetworkX* in *Proceedings of the 7th Python in Science Conference* (Pasadena, CA USA, 2008), 11–15.