Supplementary Information for How the Avidity of Polymerase Binding to the -35/-10 Promoter Sites Affects Gene Expression

Tal Einav[1,*], **Rob Phillips**[1,2,3,*]
[1]Department of Physics, California Institute of Technology, Pasadena, CA, 91125, USA
[2]Department of Applied Physics, California Institute of Technology, Pasadena, CA, 91125, USA
[3]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, 91125, USA
*Corresponding authors

www.pnas.org/cgi/doi/10.1073/pnas.1905615116

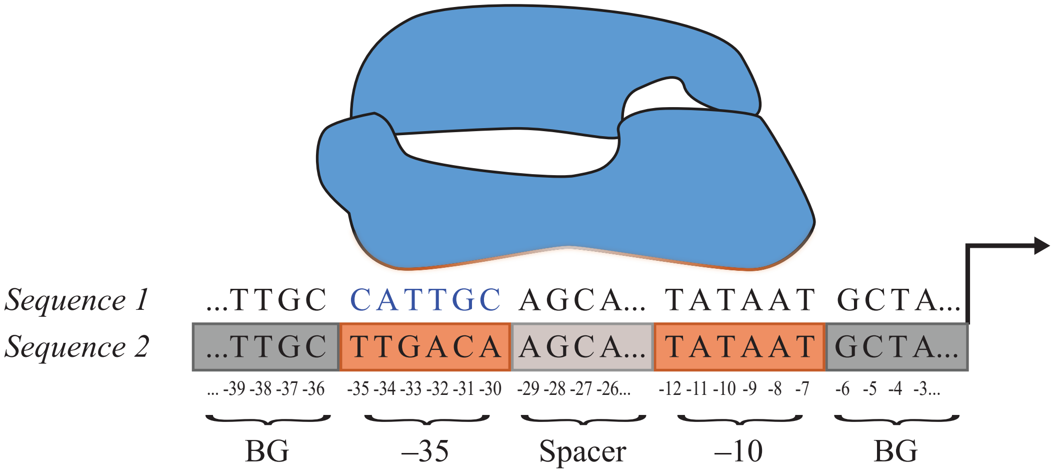
A The Energy Matrix Model

A.1 Translating between an Energy Matrix with Base Pair Resolution and Promoter Element Resolution

In this section, we discuss how an energy matrix model with base-pair resolution can be translated into an equivalent model with the resolution of promoter elements. The former model purports that the RNAP-promoter binding energy is composed of independent and linearly additive contributions from each base pair. More precisely, if at position j the base b_j (either A, T, C, or G) contributes a free energy $E_j^{(b_j)}$ to RNAP binding, then the total free energy of binding is given by $\sum_j E_j^{(b_j)}$ as shown in Fig. S1.

By breaking this sum up over the positions j demarking the -35 ($-35 \leq j \leq -30$), spacer ($-29 \leq j \leq -13$), -10 ($-12 \leq j \leq -7$), UP ($-59 \leq j \leq -38$; where "no UP" used a random sequence that did not enhance gene expression), and background (all the remaining base pairs between $-120 \leq j < 30$) elements, we achieve an energy matrix model where the free energies E_{BG}, $E_{\text{-35}}$, E_{Spacer}, and $E_{\text{-10}}$ represent the sum of all base pair contributions of the particular sequence considered. For simplicity, the UP element is not explicitly drawn in the figure.

As shown for two sample sequences in Fig. S1, modifying the -35 sequence while keeping the rest of the promoter unchanged leads to a different $E_{\text{-35}}$ but keeps E_{BG}, E_{Spacer}, and $E_{\text{-10}}$ unchanged. The expression of the full suite of 12,288 promoters studied in this work can be determined from the free energies of the three UP elements and the eight backgrounds, spacers, -35s, and -10s.



Energy Matrix (*Base Pair Resolution*)

$$\exp\Big(-\beta\{(\ldots + E^T_{-39} + E^T_{-38} + E^G_{-37} + E^C_{-36}) + (E^T_{-35} + E^T_{-34} + E^G_{-33} + E^A_{-32} + E^C_{-31} + E^A_{-30}) + (E^A_{-29} + E^G_{-28} + E^C_{-27} + E^A_{-26} + \ldots) + (E^T_{-12} + E^A_{-11} + E^T_{-10} + E^A_{-9} + E^A_{-8} + E^T_{-7}) + (E^G_{-6} + E^C_{-5} + E^T_{-4} + E^A_{-3})\}\Big)$$

$$= \exp\Big(-\beta\{E^{(1)}_{-35} + E^{(1)}_{\text{Spacer}} + E^{(1)}_{-10} + E^{(1)}_{\text{BG}}\}\Big) \qquad \textit{RNAP Binding, Sequence 1}$$

$$\exp\Big(-\beta\{(\ldots + E^T_{-39} + E^T_{-38} + E^G_{-37} + E^C_{-36}) + (E^C_{-35} + E^A_{-34} + E^T_{-33} + E^T_{-32} + E^G_{-31} + E^C_{-30}) + (E^A_{-29} + E^G_{-28} + E^C_{-27} + E^A_{-26} + \ldots) + (E^T_{-12} + E^A_{-11} + E^T_{-10} + E^A_{-9} + E^A_{-8} + E^T_{-7}) + (E^G_{-6} + E^C_{-5} + E^T_{-4} + E^A_{-3})\}\Big)$$

$$= \exp\Big(-\beta\{E^{(2)}_{-35} + E^{(1)}_{\text{Spacer}} + E^{(1)}_{-10} + E^{(1)}_{\text{BG}}\}\Big) \qquad \textit{RNAP Binding, Sequence 2}$$

Figure S1. An energy matrix model with base pair resolution translates into an energy matrix model with promoter element resolution. Each promoter element contributes to RNAP binding with free energy given by the sum of its free energies from its base pairs. The two sample sequences shown only differ in their -35 sequence (highlighted blue in Sequence 1), resulting in different values of $E^{(1)}_{-35}$ and $E^{(2)}_{-35}$ but the same free energies for the remaining promoter elements.

A.2 Characterizing the Dependence of Gene Expression on RNAP Copy Number

In this section, we explicitly write the dependence of RNAP copy number embedded within the free energies of RNAP binding in Eqs. 1 and 2, thereby making contact with previous models of gene regulation (1). To that end, we consider P RNAP molecules that are free to bind anywhere along a bacterial genome with N_{NS} $\underline{\text{n}}$on-$\underline{\text{s}}$pecific base pairs (i.e., potential RNAP binding sites outside of our promoter of interest). Let $\Delta\epsilon$ be the average energy difference between RNAP bound to the specific promoter versus at any other location along the genome. By definition, the free energy of RNAP binding considered in this work is given by both the entropic and energetic contributions of this binding, namely,

$$e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}}+E_{\text{-35}}+E_{\text{-10}})} \equiv \frac{P}{N_{\text{NS}}} e^{-\beta\Delta\epsilon} = e^{-\beta(\Delta\epsilon - k_B T \log(P/N_{\text{NS}}))}. \tag{S1}$$

Because the gene expression for each promoter generated by Urtecho *et al.* was measured under the same experimental condition, the RNAP copy number is consistent across all constructs, and hence the constant $k_B T \log(P/N_{\text{NS}})$ can be absorbed into the free energies. If these measurements were repeated under experimental conditions where the RNAP copy number is halved ($P \to \frac{P}{2}$), the total free energy of RNAP binding considered in this work would need to be correspondingly modified from $(E_{\text{BG}} + E_{\text{Spacer}} + E_{\text{-35}} + E_{\text{-10}}) \to (E_{\text{BG}} + E_{\text{Spacer}} + E_{\text{-35}} + E_{\text{-10}} + k_B T \log 2)$.

B Model Fitting and Parameter Values

The energy matrix model (Eq. 1) was solved as a least-squares problem that only fit the promoters in Fig. 2A with no UP element. The multivalent model (Eq. 5) was fit using nonlinear regression on promoter sequences with and without an UP element in order to obtain a single, self-consistent set of parameters capable of capturing the data in Fig. 2B and Fig. 3B. The fitting of both models is presented in a supplementary Mathematica notebook available online (doi: 10.22002/D1.1242).

The coefficient of determination R^2 was calculated for $y_{\text{data}} = \log_{10}(\text{gene expression})$ to prevent the largest gene expression values from dominating the result. We trained both the energy matrix and multivalent models on 75% of the data and characterized the predictive power on the remaining 25%, repeating the procedure 10 times. The exact form used was

$$R^2 = 1 - \frac{\sum_{j=1}^{N} \left(y_{\text{data}}^{(j)} - y_{\text{predicted}}^{(j)} \right)^2}{\sum_{j=1}^{N} \left(y_{\text{data}}^{(j)} - \bar{y}_{\text{data}} \right)^2} \tag{S2}$$

where $y_{\text{predicted}}$ is the vector of the N measurements of $\log_{10}(\text{gene expression})$ predicted by the model and $\bar{y}_{\text{data}} = \frac{1}{N} \sum_{j=1}^{N} y_{\text{data}}^{(j)}$ is the average of the logarithmic gene expression data. In this form, the R^2 represents the fraction of variance in the measured gene expression data that arises from the variance in the predicted gene expression data. To test the predictive power of each model, we also trained both models on only 10% of the data and used it to predict the gene expression of the remaining 90% of promoters. We found that the coefficient of determination R^2 only slightly decreased from $0.57 \to 0.54$ for the energy matrix model and from $0.91 \to 0.86$ for the multivalent model when fitting on this much smaller training set, demonstrating that these models require no more than a thousand promoters to reach their full predictive power.

Table S1 shows the parameter values inferred by the energy matrix (Fig. 2A) and multivalent (Figs. 2B and 3B) models. Due to the large number of parameters involved, both models exhibit parameter degeneracy (2) where disparate sets of parameters yield nearly identical results. For example, all of the free energies of the spacer elements can be increased by an arbitrary amount provided that the free energies of all background elements are decreased by this same amount (with similar degeneracy holding between other pairs of promoter elements). To circumvent this degeneracy, one -35, one spacer, one -10, one UP, and one background element (denoted by asterisks in Table S1) were fixed to their corresponding value in the energy matrix model, and as such, the parameters below may not represent the binding energies of the promoter elements, but rather only one possible embodiment of these values.

We point out that our model coarse-grains kinetic details of transcription (e.g., transcription initiation, elongation, transcriptional bursting) into the levels of gene expression r_j shown in Fig. 1B. Modifying the promoter sequence (i.e., considering different spacers or backgrounds) may well change these rates, although our model assumes that such changes only affect the RNAP-promoter binding affinity. If experiments measure the changes in these kinetic rates, they could either be incorporated into the r_j or into an expanded model that explicitly takes these steps of transcription into account (3).

The small but nonzero r_0 term in our model (Fig. 1B) represents the background level of gene expression arising from promoters that lack an RNAP-binding site. Urtecho *et al.* measured 500 negative controls, sequences from the *E. coli* genome that have no promoter or RNA-seq activity, whose expression was nonzero and centered around 0.15 (see Fig. 2E of Ref. (4)), comparable to our inferred $r_0 = 0.18$ value. This nonzero expression may arise from instrumental noise or spurious transcription, and we elected to model it using a nonzero r_0 rather than background subtracting it in order to present the data on a log-scale (upon background subtraction, some gene expression measurements would be negative due to experimental noise which would have precluded log-plots). We note that log-fitting is likely to more accurately portray how gene expression proceeds in the cell, since most endogenous promoters exhibit low gene expression while synergistic effects between promoter elements can play a significant role; in contrast, linear fitting would downplay the importance of all but the strongest promoters.

Lastly, we note from Urtecho *et al.* that the UP elements were named because they increased transcription by 136-fold and 326-fold *in vivo* relative to the physiological *rrnb* P1 UP element. Thus, it follows that the free energy of the 326-fold UP should be smaller than that of the 136-fold UP which should be smaller than the free energy of having no UP element, as seen in Table S1. Additionally, we point out that all spacer elements are 17 bp long; RNAP binding is highly dependent on this length, and promoters with longer or shorter spacers may influence the -35 and -10 binding free energies. Lastly, the sequence composition of all spacers and backgrounds is given in Ref. (4).

Description	Parameter	Energy Matrix Model	Multivalent Model
Level of	$r_{\max}$	0.42	8.6
gene expression	r_0	—	0.18
Interaction energy	E_{int}	—	−6.3
-35 motif	TTGACA	−1.3	−1.3*
($E_{\text{-35}}$)	TTTACA	−0.2	3.3
	ATTACA	0.6	8.3
	TTTACC	0.3	5.7
	TTAAGA	0.6	8.8
	TTGCAA	−0.4	2.5
	CTCAGA	0.7	9.5
	CTTAGA	0.6	9.5
-10 motif	TATAAT	−0.9	−0.9*
($E_{\text{-10}}$)	AATAAT	−0.1	3.6
	GATAAT	−0.1	3.2
	TATAAA	0.0	4.4
	GATAAC	0.6	9.8
	TATGTT	0.1	4.6
	GTTAAA	0.6	>10
	GTTGTA	0.6	>10
Spacer	P1-6	0.0	0.0*
(E_{Spacer})	*lac*	0.4	3.9
	ECK125136938	0.0	1.0
	ECK125137104	0.1	1.5
	ECK125137108	0.1	0.7
	ECK125137405	0.1	1.2
	ECK125137640	0.4	3.8
	ECK125137726	−0.1	−0.8
Background	bg463205:463355	0.0	0.0*
(E_{BG})	bg977040:977190	0.1	2.3
	bg991964:992114	0.3	4.2
	bg1163421:1163571	0.1	1.5
	bg3514590:3514740	0.1	1.9
	bg4323949:4324099	0.1	2.5
	bg4427287:4427437	0.2	3.1
	bg4471352:4471502	−0.1	1.6
UP	No UP	0.0	0.0*
(E_{UP})	136-fold UP	—	−2.6
	326-fold UP	—	−3.2

Table S1. Parameter values for the models of transcriptional regulation considered in this work. The levels of gene expression (r_0 and $r_{\max}$) are in the same arbitrary units as the experimental measurements (4) while the energies are all in $k_B T$ units (energies that are more negative indicate tighter binding). The original nomenclature from Table S1 in Ref. (4) is used for each promoter element. Parameter denoted by an asterisk (*) represent values that were fixed to their corresponding value in the energy matrix model to prevent parameter degeneracy.

C Comparing the Energy Matrix and Multivalent Models of Gene Expression

C.1 An Epistasis-Free Energy Matrix Model with Saturation does not Capture the Trends in Gene Expression Exhibited by the Data

As shown in Fig. 2A, the simplest model where gene expression is proportional to $e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}}+E_{\text{-35}}+E_{\text{-10}})}$ (in the absence of an UP element) fails to characterize the data ($R^2 = 0.57$). In contrast, the multivalent model in Fig. 2B quantitatively matches the behavior of the spectrum of promoters ($R^2 = 0.91$). Thus, it behooves us to examine what properties of the latter model are necessary to achieve this concordance with the data.

To that end, we consider an intermediate model where gene expression is given by

$$\text{GE} = \frac{r_0 + \tilde{r}_{\max} e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}}+E_{\text{-35}}+E_{\text{-10}})}}{1 + e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}}+E_{\text{-35}}+E_{\text{-10}})}} \tag{S3}$$

where r_0 represents the minimum level of gene expression in the absence of RNAP binding, $\tilde{r}_{\max}$ denotes the amount of gene expression when RNAP is fully bound to the promoter, and the E_j represent the free energy contribution of the promoter element j. Note that this model represents the limit of a very strong interaction energy in Eq. 3 ($e^{-\beta E_{\text{int}}} \gg 1$ with $\tilde{r}_{\max} = r_{\max} e^{-\beta E_{\text{int}}}$) where RNAP is either unbound or fully bound to the promoter.

Fig. S2A demonstrates that the data is well characterized by Eq. S3 ($R^2 = 0.91$). Therefore, one key feature missing from the simplest energy matrix model description Eq. 1 was that gene expression will saturate once RNAP binding becomes sufficiently strong (or, mathematically, that the denominator $1 + e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}}+E_{\text{-35}}+E_{\text{-10}})}$ must include the RNAP binding term). Note that the results of this energy matrix model with saturation are nearly identical to the results of the multivalent model in Fig. 2B. Indeed, since the inferred interaction energy $E_{\text{int}} = -6.3\, k_B T$ between the -35 and -10 sites is large and negative (see Table S1), it is not surprising that the two models produce similar results for the majority of promoters.

Intuitively, the difference between these two models will emerge in their predictions for promoters with weak expression. As we will show below, the energy matrix model with saturation (Eq. S3) is epistasis-free: given the gene expression of any initial promoter and two mutants of that promoter, we can predict the expression of the double mutant. If, for example, the initial promoter exhibits weak gene expression and the two mutants exhibit a medium level of gene expression, then the double mutant would be predicted to exhibit a large amount gene expression. As will be explained below, the resulting predictions shown in Fig. S2B are highly damning. On the other hand the multivalent model (Eq. 3) predicts a more complex relationship between these four promoters, and in the Appendix C.2 we examine an analytically tractable limit to show that this model better recapitulates the gene expression measurements.

We proceed by utilizing the epistasis-free nature of Eq. S3. A key feature of the following analysis is that it will not require any model fitting, and hence for the remainder of this Appendix we proceed as if we have no knowledge of the parameter values in Table S1. To begin, we approximate the values of $r_0 \approx 0.2$ and $\tilde{r}_{\max} \approx 10$ from the gene expression data (the minimum and maximum y-values in Fig. S2A, averaging by eye to account for noise). These two values, together with the gene expression measurements for every construct, will be sufficient to make our epistasis-free predictions without explicitly determining any of the E_j.

As in the main text, denote the gene expression $\text{GE}^{(0,0)}$ of a promoter with the consensus -35 and -10 sequences (and any background or spacer sequence). Let $\text{GE}^{(1,0)}$, $\text{GE}^{(0,1)}$, and $\text{GE}^{(1,1)}$ represent promoters (with this same background and spacer) whose -35/-10 sequences are mutated/consensus, consensus/mutated, and mutated/mutated, respectively. Eq. S3 can be inverted to determine

$$e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}}+E_{\text{-35}}+E_{\text{-10}})} = \frac{\text{GE} - r_0}{\tilde{r}_{\max} - \text{GE}} \equiv f(\text{GE}) \tag{S4}$$

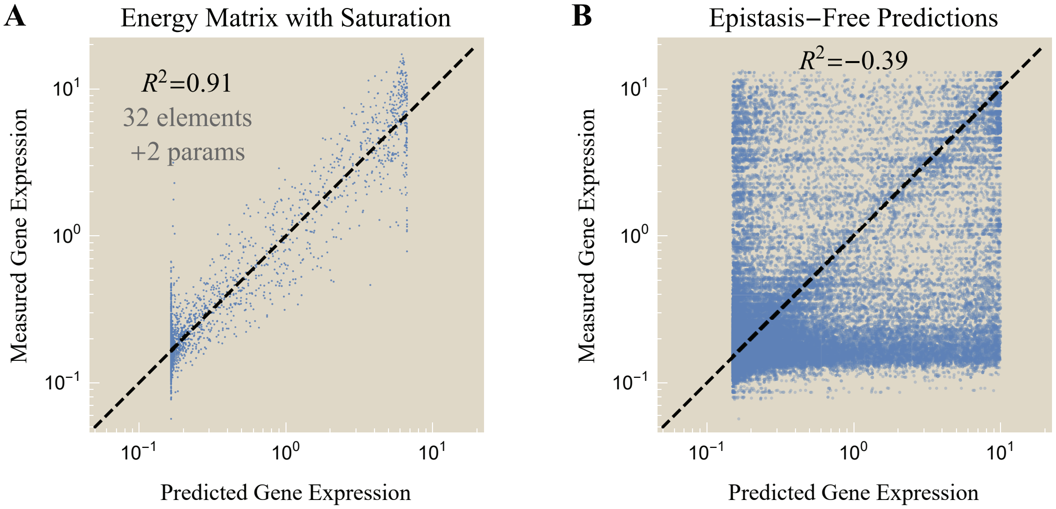


Figure S2. Gene expression represented by an energy matrix model with saturation. (A) Characterization of the same promoters as in Fig. 2 using the energy matrix model with saturation (Eq. S3) with essentially identical fit quality as the multivalent model. (B) Since this model assumes that the RNAP-promoter binding energy is epistasis-free (with the -35 and -10 binding sites contributing additively and independently to the RNAP binding energy), the gene expression of double mutants can be predicted from the expression of single mutants without resorting to fitting (Eqs. S4 and S5). The large deviations demonstrate that the energy matrix with saturation cannot characterize the gene expression of these constructs.

for the double mutant with $\text{GE}^{(1,1)}$ as well as the two singly mutated promoters with $\text{GE}^{(1,0)}$ and $\text{GE}^{(0,1)}$, where we have defined the function f for convenience. Importantly, since the -35 and -10 binding energies additively and independently contribute to the RNAP-promoter free energy, the left-hand side of Eq. S4 for the unmutated construct is given by $f(\text{GE}^{(1,1)})\frac{f(\text{GE}^{(0,1)})}{f(\text{GE}^{(1,1)})}\frac{f(\text{GE}^{(1,0)})}{f(\text{GE}^{(1,1)})}$ (exactly analogous to Eq. 4 for the simple energy matrix model). Therefore, its gene expression is predicted to be

$$\text{GE}^{(0,0)} = \frac{r_0 + \tilde{r}_{\max} f(\text{GE}^{(1,1)})\frac{f(\text{GE}^{(0,1)})}{f(\text{GE}^{(1,1)})}\frac{f(\text{GE}^{(1,0)})}{f(\text{GE}^{(1,1)})}}{1 + f(\text{GE}^{(1,1)})\frac{f(\text{GE}^{(0,1)})}{f(\text{GE}^{(1,1)})}\frac{f(\text{GE}^{(1,0)})}{f(\text{GE}^{(1,1)})}}. \tag{S5}$$

Fig. S2B shows the results of these epistasis-free predictions. Because Eq. S5 applies to *any* pairs of -35 and -10 elements with the same BG and spacer, there is a combinatorial explosion of predictions, providing a solid test for this model. As can be seen, aside from the plethora of data points with correctly-predicted low gene expression in the bottom-left corner of the plot, there are large swathes of data points that do not fall on the expected diagonal line, indicating that the epistasis-free prediction in Eq. S3 cannot accurately capture the gene expression of the constructs considered here. In the next section, we show that the multivalent model is better equipped to characterize these cases. Notably, these results indicate that although a model may fit the majority of data on average (as in Fig. S2A), it may nevertheless make spurious predictions. Such hidden gems may go unnoticed when a pure-fitting mentality is applied to the wealth of data that is becoming increasingly easy to generate.

C.2 A Multivalent Model outperforms the Energy Matrix Model in the Limit of a Weak -35 or Weak -10 RNAP Binding Site

In section C.1, we showed that an energy matrix model with saturation Eq. S3 is epistasis-free and hence makes sharp predictions that are inconsistent with the data (Fig. S2B). In this section, we consider the multivalent model Eq. 3 where binding to the -35 and -10 sites is no longer independent. Because this latter model exhibits epistasis, we will restrict our analysis to the limit of weak promoters with no UP element where we can approximate the multivalent model and compare its results to the energy matrix model with saturation. As before, we proceed without referencing the parameter values in Table S1 to emphasize that this analysis can be done without recourse to fitting.

We define $\text{GE}^{(1,1)}$, $\text{GE}^{(1,0)}$, $\text{GE}^{(0,1)}$, and $\text{GE}^{(0,0)}$ as in section C.1, but we will restrict our attention to promoters where the original sequence exhibits low gene expression ($\text{GE}^{(1,1)} \lesssim 0.25$) and the two mutants exhibit medium gene expression ($0.25 \lesssim \text{GE}^{(1,0)}, \text{GE}^{(0,1)} \lesssim 1.0$). For such cases, we expect that the predicted gene expression $\text{GE}^{(0,0)}$ of the double mutant will be larger in the multivalent model (Eq. 3) than the energy matrix model with saturation (Eq. S3) due to the avidity between the -35 and -10 sites. In other words, the multivalent model acknowledges that the -35 and -10 sites bolster each other and consequently predicts larger gene expression when both sites exhibit even a moderate capability of binding.

As discussed in section C.1, $\text{GE}^{(0,0)}$ is exactly given by Eq. S5 in the energy matrix model with saturation. Applying that result to the present case of weak promoters ($\text{GE}^{(1,1)} \lesssim 0.25$) with medium-strength singly mutants ($0.25 \lesssim \text{GE}^{(1,0)}, \text{GE}^{(0,1)} \lesssim 1.0$), Fig. S3A shows that this model generally underestimates the gene expression of these promoters. This serves as a promising indicator that the avidity of RNAP binding is missing from such an approach.

We next turn to the more complex multivalent model. Because RNAP exhibits epistasis within this framework, the relationship between gene expression is more complex and hence we only roughly approximate $\text{GE}^{(0,0)}$. To that end, it behooves us to generalize the levels of gene expression in Fig. 1(B) so that RNAP bound only at the -35 site leads to a gene expression level of $r_{\text{-35}}$ while RNAP bound only at the -10 site elicits $r_{\text{-10}}$ gene expression (satisfying $r_0 < r_{\text{-35}}, r_{\text{-10}} \leq r_{\max}$), leading to

$$\text{GE} = \frac{r_0 + e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})}\left(r_{\text{-35}} e^{-\beta E_{\text{-35}}} + r_{\text{-10}} e^{-\beta E_{\text{-10}}} + r_{\max} e^{-\beta(E_{\text{-35}}+E_{\text{-10}}+E_{\text{int}})}\right)}{1 + e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})}\left(e^{-\beta E_{\text{-35}}} + e^{-\beta E_{\text{-10}}} + e^{-\beta(E_{\text{-35}}+E_{\text{-10}}+E_{\text{int}})}\right)}. \tag{S6}$$

In the main text, we assumed that $r_{\text{-35}} = r_{\text{-10}} = r_0$ for simplicity (and because relaxing this assumption does not qualitatively change any of our results). Here, we will keep these more general rates, as it will aid in the following analysis.

Using Eq. S6, we can approximate gene expression for our four promoters of interest,

$$\text{GE}^{(1,1)} \approx r_0 \tag{S7}$$

$$\text{GE}^{(0,1)} \approx \frac{r_0 + r_{\text{-35}} e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})} e^{-\beta E_{\text{-35}}}}{1 + e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})} e^{-\beta E_{\text{-35}}}} \tag{S8}$$

$$\text{GE}^{(1,0)} \approx \frac{r_0 + r_{\text{-10}} e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})} e^{-\beta E_{\text{-10}}}}{1 + e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})} e^{-\beta E_{\text{-10}}}} \tag{S9}$$

$$\text{GE}^{(0,0)} \approx \frac{r_0 + r_{\max} e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})} e^{-\beta(E_{\text{-35}}+E_{\text{-10}}+E_{\text{int}})}}{1 + e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})} e^{-\beta(E_{\text{-35}}+E_{\text{-10}}+E_{\text{int}})}}. \tag{S10}$$

In Eq. S7, we used the fact that the promoter is very weak ($\text{GE}^{(1,1)} \lesssim 0.25$) to infer that RNAP is unable to bind at either the -35 or -10 sites ($e^{-\beta E_{\text{-35}}}, e^{-\beta E_{\text{-10}}} \ll 1$). Since replacing the -35 site slightly improves gene expression ($0.25 \lesssim \text{GE}^{(0,1)} \lesssim 1.0$), we only keep the -35 binding term in Eq. S8 but continue to neglect the -10 terms (assuming that binding to the -10 is sufficiently unfavored that it overwhelms the avidity term, $e^{-\beta(E_{\text{-10}}+E_{\text{int}})} \ll 1$). Analogous statements hold for $\text{GE}^{(1,0)}$ and the -10 site in Eq. S9. Lastly, when both the -35 and -10 sites are replaced (Eq. S10), the fully bound RNAP state will dominate over the two partially bound states due to avidity.

For every set of four promoters satisfying our criteria, we can use Eq. S7 to infer r_0, Eq. S8 to solve for $e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})}e^{-\beta E_{\text{-35}}}$ (in terms of $r_{\text{-35}}$), and Eq. S9 to solve for $e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})}e^{-\beta E_{\text{-10}}}$ (in terms of $r_{\text{-10}}$). In addition, we can directly estimate $r_{\text{max}} \approx 10$ directly from the maximum gene expression of all promoters. Combining these statements, we can rewrite Eq. S10 as

$$\text{GE}^{(0,0)} \approx \frac{r_0 + r_{\text{max}} A e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})}e^{-\beta E_{\text{-35}}}e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})}e^{-\beta E_{\text{-10}}}}{1 + A e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})}e^{-\beta E_{\text{-35}}}e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})}e^{-\beta E_{\text{-10}}}} \tag{S11}$$

with the unknown quantity $A = e^{\beta(E_{\text{BG}}+E_{\text{Spacer}}-E_{\text{int}})}$. Therefore, the three unknown constants $r_{\text{-35}}$, $r_{\text{-10}}$, and A would permit us to predict $\text{GE}^{(0,0)}$ using single and double mutant data within the multivalent model. To facilitate this, we coarsely approximate that the partially bound RNAP states give rise to intermediate expression levels $r_{\text{-35}} = r_{\text{-10}} \approx \sqrt{r_0 r_{\text{max}}} \approx 1$ and that the average energy of a background and spacer sequence is negligible compared to the favorable interaction energy which is on the order of several $k_B T$ leading to $A \approx e^{-\beta E_{\text{int}}} \approx 100$. Fig. S3B demonstrates that the multivalent model predicts larger gene expression often closer to $r_{\text{max}} \approx 10$. Although this approximate result for the multivalent model exhibits scatter about the predicted diagonal line, it nevertheless show a marked improvement over the energy matrix model, supporting the notion that avidity is a key concept when predicting the gene expression of mutations that greatly weaken or greatly strengthen the -35 and -10 sites.

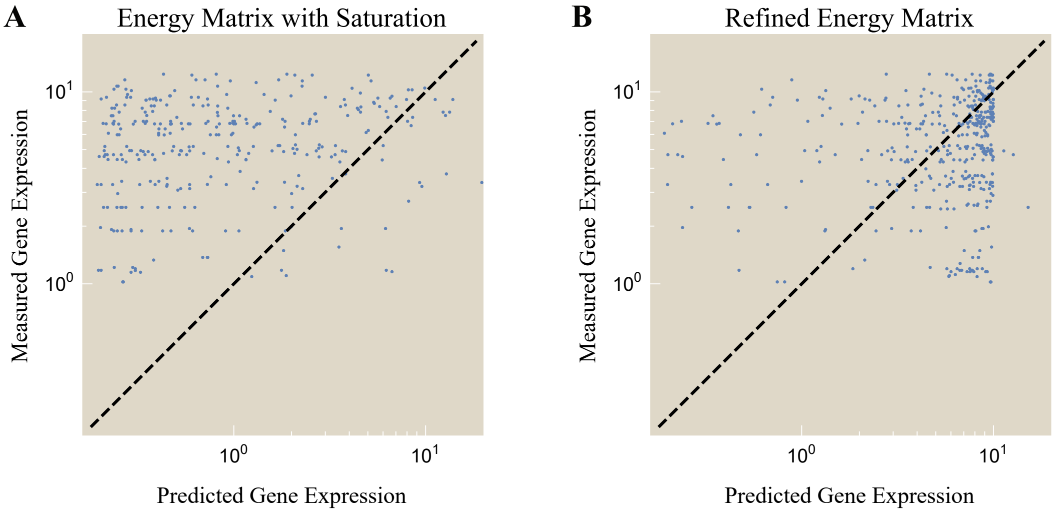


Figure S3. Relating gene expression measurements with minimal fitting. Using gene expression measurements for a weak promoter and two single mutants with higher gene expression, we can predict the expression of the double mutant and compare it to data. (A) The epistasis-free energy matrix model with saturation (Eq. S3) underestimates the gene expression, suggesting that the avidity between the -35 and -10 sites is missing from this analysis. (B) The multivalent model Eq. 3 predicts higher gene expression levels that better characterize the data.

D Interactions Between the Different Promoter Elements

In this section, we extend the analysis shown in the Fig. 2A inset to determine the strength of interactions between every pair of promoter elements as shown in Fig. S5A. As an example, Fig. S4 considers the combinations of a promoter with two possible -35 motifs ($-35^{(1)}$ or $-35^{(2)}$) and two possible spacers (Spacer$^{(1)}$ or Spacer$^{(2)}$) with the same UP, -10, and background sequences.

Suppose that the -35 and spacer elements contribute independently to gene expression (GE) so that we can write $\text{GE} = f_1(E_{-35})f_2(E_{\text{Spacer}})$ as the product of two functions f_1 and f_2 (in the standard energy matrix model, $f_1(E) = f_2(E) = e^{-\beta E}$). This independence implies that the system has no epistasis, namely,

$$\text{GE}^{(0,0)} = \text{GE}^{(1,1)} \frac{\text{GE}^{(0,1)}}{\text{GE}^{(1,1)}} \frac{\text{GE}^{(1,0)}}{\text{GE}^{(1,1)}}. \tag{S12}$$

Thus, for all possible pairs of -35 and spacer elements, we can compare the predicted gene expression given by Eq. S12 with the experimental measurements to discern whether these two segments of the promoter contribute independently to gene expression. In the following analysis, we will also restrict ourselves to promoters where $\text{GE} > 10^{-0.5}$ for all four mutants to ensure that the measurements are within the dynamic range of the experiment (so that we can be certain we are analyzing gene expression measurements and not noise).

D.1 Characterizing Promoters with no UP Element

We first carry out this analysis on the 4,096 promoters with no UP elements as shown in Fig. S5. In each plot, we compare the epistasis-free predicted GE (x-axis) with the measured value (y-axis). If two promoter elements independently contribute to gene expression, their data should fall onto the straight line $y = x$. We can quantify all deviations from such lines using the coefficient of determination R^2, with smaller R^2 values signifying that the promoter elements are not multiplicatively independent.

This analysis shows that while the -35 and -10 elements interact in a fashion discordant with an energy matrix formulation (leading to a negative R^2), the remaining promoter elements interact approximately independently of each other and can be approximated using an energy matrix model. This rigorously justified our sole consideration of the -35 and -10 binding sites in Fig. 1B, allowing us to avoid, for example, enumerating states where the RNAP is solely bound to the spacer or the background. Instead, the promoter is well approximated by treating the -35 and -10 motifs as cooperative binding sites while the spacer and background contribute independently to RNAP binding (as per Eq. 3).

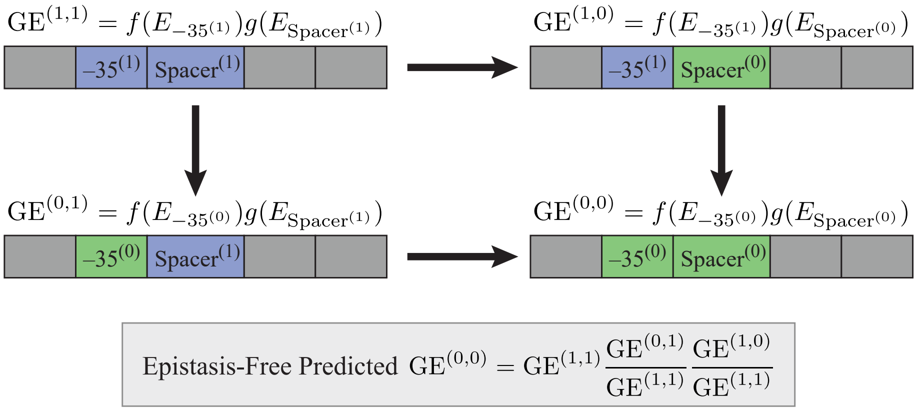

$\text{GE}^{(1,1)} = f(E_{-35^{(1)}})g(E_{\text{Spacer}^{(1)}})$
$-35^{(1)}$
Spacer$^{(1)}$
$\text{GE}^{(1,0)} = f(E_{-35^{(1)}})g(E_{\text{Spacer}^{(0)}})$
$-35^{(1)}$
Spacer$^{(0)}$
$\text{GE}^{(0,1)} = f(E_{-35^{(0)}})g(E_{\text{Spacer}^{(1)}})$
$-35^{(0)}$
Spacer$^{(1)}$
$\text{GE}^{(0,0)} = f(E_{-35^{(0)}})g(E_{\text{Spacer}^{(0)}})$
$-35^{(0)}$
Spacer$^{(0)}$
Epistasis-Free Predicted $\text{GE}^{(0,0)} = \text{GE}^{(1,1)} \frac{\text{GE}^{(0,1)}}{\text{GE}^{(1,1)}} \frac{\text{GE}^{(1,0)}}{\text{GE}^{(1,1)}}$

Figure S4. Quantifying the interactions between promoter elements. If the -35 and spacer promoter elements independently contribute to gene expression, then an epistasis-free prediction of gene expression for the double mutant (bottom right) can be predicted using the gene expression of the other three promoters.

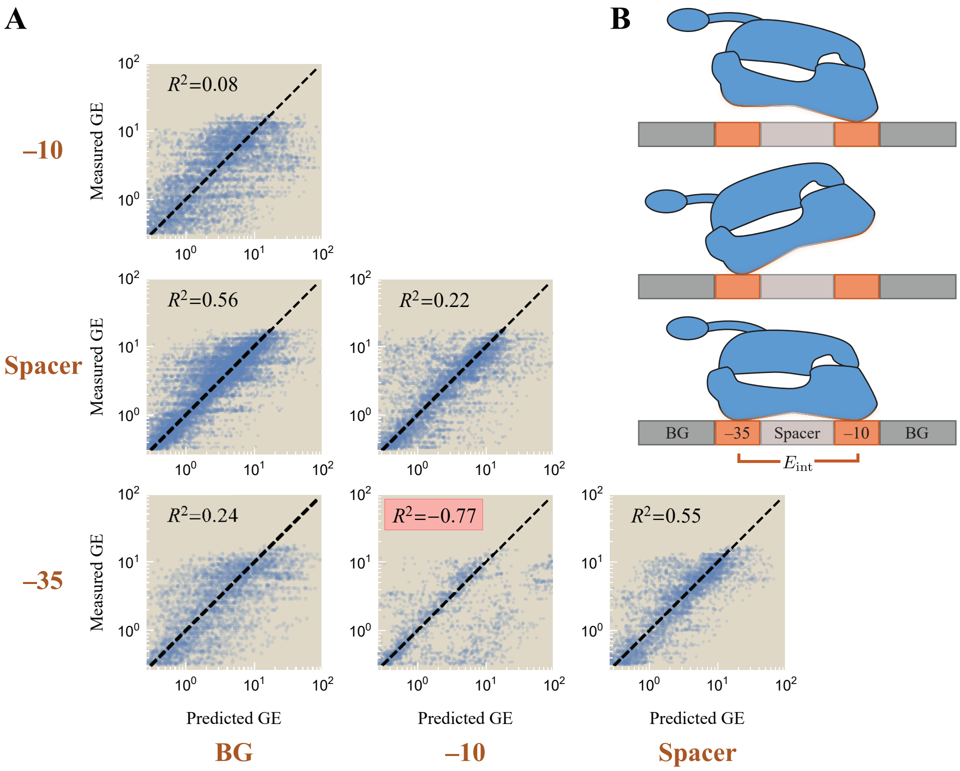


Figure S5. Interactions between the promoter elements with no UP binding site. (A) For every pair of elements (brown labels on the left and bottom), the measured gene expression (y-axis) is compared to the epistasis-free prediction (x-axis, Eq. S12) assuming that the two promoter elements are independent. Deviations between the predictions and measurements indicate that the two promoter elements interact. Data is plotted with low opacity to better show the general trend across the promoters. (B) The resulting schematic of a promoter with no UP element is that RNAP can bind to either the -35 or -10 sites independently with an avidity interaction when both are bound; the spacer and background (BG) contribute independently to the RNAP binding energy provided RNAP is bound to either the -35 or -10 element.

Lastly, we note that the multivalent model (Eq. 3) does not strictly exhibit the multiplicative independence between the -35 and spacer elements (or any of the other weakly interacting promoter elements) that would lead to an $R^2 = 1$ expectation, but as we now show it closely approximates multiplicative independence. First, note that using the parameter values from Table S1, the denominator in the multivalent model is approximately 1 because $e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})}\left(e^{-\beta E_{\text{-35}}}+e^{-\beta E_{\text{-10}}}+e^{-\beta(E_{\text{-35}}+E_{\text{-10}}+E_{\text{int}})}\right)$ is $\lesssim 1$ for approximately 90% of the promoters. Additionally, the numerator in the model may be dominated by either of its terms: For weak promoters that exhibit low levels of gene expression, $r_0 \gg r_0 e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})}\left(e^{-\beta E_{\text{-35}}}+e^{-\beta E_{\text{-10}}}+\frac{r_{\max}}{r_0}e^{-\beta(E_{\text{-35}}+E_{\text{-10}}+E_{\text{int}})}\right)$ and GE $\approx r_0$. In the opposite limit where expression is large, $r_0 \ll r_0 e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})}\left(e^{-\beta E_{\text{-35}}}+e^{-\beta E_{\text{-10}}}+\frac{r_{\max}}{r_0}e^{-\beta(E_{\text{-35}}+E_{\text{-10}}+E_{\text{int}})}\right)$ and we can approximate Eq. 5 as

$$\text{GE} \approx r_0 e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})}\left(e^{-\beta E_{\text{-35}}}+e^{-\beta E_{\text{-10}}}+\frac{r_{\max}}{r_0}e^{-\beta(E_{\text{-35}}+E_{\text{-10}}+E_{\text{int}})}\right), \tag{S13}$$

which exhibits the multiplicative independence implied by Eq. S12 between the weakly interacting promoter elements. Practically speaking, this means that in creating Fig. S5, we only considered data points where gene expression was above the background level that we inferred to be $10^{-0.5}$ based on the gene expression measurements in Fig. 2. In summation, Eq. 5 exhibits approximate independence between the weakly interacting promoter elements which can be identified as the plots for which $R^2 > 0$ in Fig. S5.

D.2 Characterizing Promoters with an UP Element

Here, we extend the analysis in the previous section to a promoter that includes an UP element. As before, we seek to understand whether the UP, -35, spacer, -10, and background elements act independently of each other or whether they interact with avidity to facilitate RNAP binding.

Fig. S6 carries out this analysis using all 12,288 sequences from Urtecho *et al.* for every pair of promoter elements (4). As in the previous section, we find that the -35 and -10 sites do not interact independently (as shown by a negative R^2). We acknowledge that several additional pairs of elements (i.e., -10/BG, -35/BG, and -10/Spacer) exhibit low R^2 values which may arise because: (*i*) Our model only approximately obeys multiplicative independence as discussed in Appendix D.1 (so that $R^2 \approx 1$ even in the absence of experimental noise) or (*ii*) there may be additional interactions between promoter elements that we neglect, such as the importance of the discriminator (5) or weak RNAP binding sites in the background sequences (6). We proceed by only considering interactions sufficiently strong to induce a negative R^2 value, namely, the avidity between the -35 and -10 motifs, with our eyes wide open to the possibility that more complex models could attempt to capture the full suite of higher-order interactions.

We end this section by analyzing which of the three schematics shown in Fig. 3A best characterizes the binding of the UP element. We note that the UP element appears to be particularly independent ($0.6 \lesssim R^2$) compared all other pairings of elements ($0.1 \lesssim R^2 \lesssim 0.6$), suggesting that the RNAP C-terminal binds weakly provided that either the -35 or -10 motifs are bound (Fig. S6B). This supports the bottom schematic in Fig. 3 and gives rise to the form of gene expression Eq. 5 used in the main text.

To complete this argument, we further note that the middle schematic in Fig. 3A would imply that the UP element only binds when the -35 element is bound, which would result in gene expression of the form

$$\mathrm{GE} = r_0 \frac{1 + e^{-\beta(E_{\mathrm{BG}}+E_{\mathrm{Spacer}})}\left(e^{-\beta(E_{\text{-35}}+E_{\mathrm{UP}})} + e^{-\beta E_{\text{-10}}} + \frac{r_{\mathrm{max}}}{r_0} e^{-\beta(E_{\text{-35}}+E_{\mathrm{UP}}+E_{\text{-10}}+E_{\mathrm{int}})}\right)}{1 + e^{-\beta(E_{\mathrm{BG}}+E_{\mathrm{Spacer}})}\left(e^{-\beta(E_{\text{-35}}+E_{\mathrm{UP}})} + e^{-\beta E_{\text{-10}}} + e^{-\beta(E_{\text{-35}}+E_{\mathrm{UP}}+E_{\text{-10}}+E_{\mathrm{int}})}\right)}. \qquad \text{(S14)}$$

In this case, we would expect a low R^2 between the -35 and -10 elements as well as between the UP and -10 elements, but we would have an $R^2 \approx 1$ between UP and -35 since binding of one forces the binding of the other in this model. Given the larger-than-expected $R^2 = 0.56$ value between the UP and -10 elements and the smaller-than-expected $R^2 = 0.62$ value between the UP and -35 elements, this model is unlikely to be correct.

Finally, the top schematic in Fig. 3A implies a low R^2 value between the -35 and -10, the UP and -35, and the UP and -10 elements. In this case, all three elements bind strongly and in a highly dependent manner, so that eight RNAP states would need to be considered (with avidity terms between every pair of elements). Because the R^2 values between the UP/-35 and UP/-10 are larger than expected, this model does not appear to properly characterize RNAP binding, leading us to favor the bottom schematic in Fig. 3A.

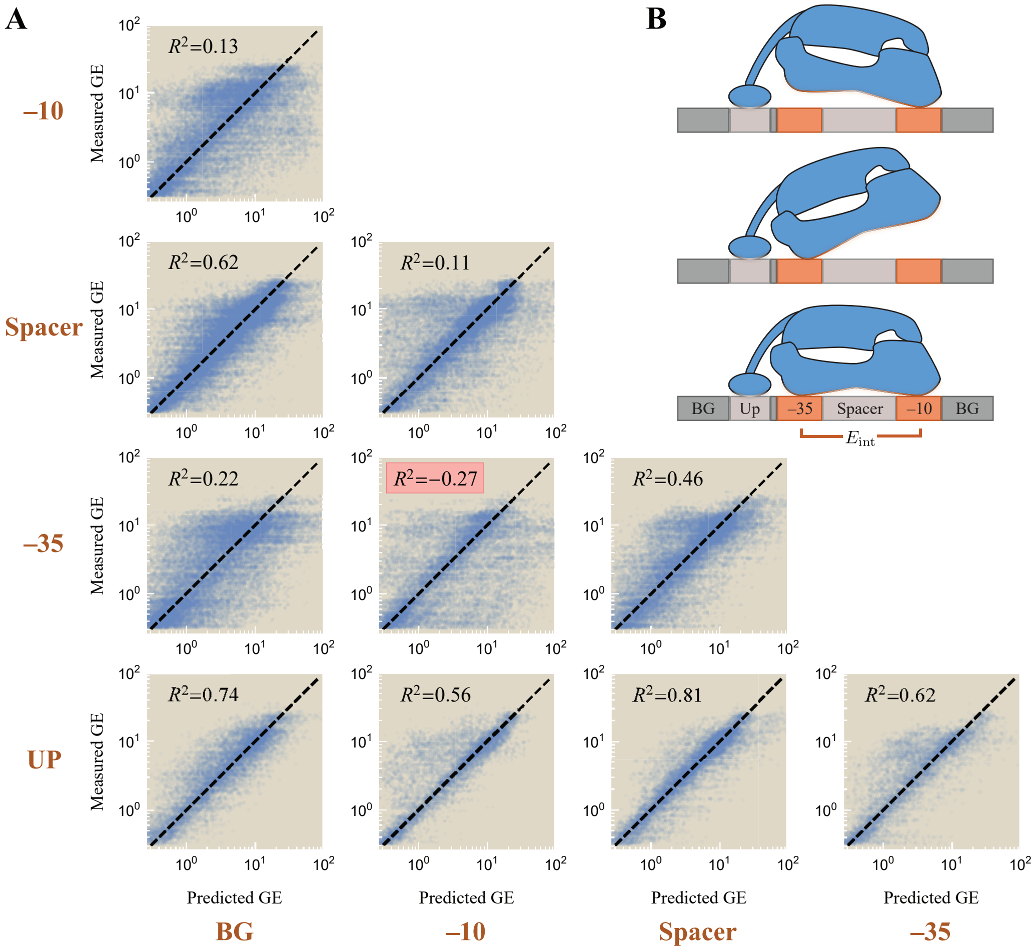


Figure S6. Interactions between the promoter elements with an UP binding site. (A) For every pair of elements (brown labels on the left and bottom), the measured gene expression (y-axis) is compared to the epistasis-free prediction (x-axis, Eq. S12) assuming that the two promoter elements are independent. (B) The resulting schematic of gene expression where RNAP can bind to either the -35 or -10 sites independently with an avidity interaction when both are bound; the UP, spacer, and background (BG) contribute independently to the RNAP binding energy provided RNAP is bound to either the -35 or -10 element.

E RNAP Binding Too Tightly Decreases Gene Expression

All of the gene expression models examined in this work assert that gene expression monotonically increases with the RNAP-promoter binding affinity. In contrast, Urtecho *et al.* found that this monotonic relationship did not hold for the strongest promoters. In other words, gene expression increased as the -35 and -10 motifs approached their consensus sequences (which bind the tightest to RNAP) *except* that promoters with both a consensus -35 and consensus -10 sequence exhibited lower gene expression than the corresponding sequences with one mutation in either motif (4). This suggests that past a certain point, increasing the RNAP-promoter binding energy causes RNA polymerase to bind top tightly, thereby inhibiting gene expression (7).

In this Appendix, we explore this phenomenon and develop a model to account for it. More specifically, our model will posit that the state of transcription initiation can be characterized by a free energy so that the probability of initiating transcription versus remaining bound on the promoter is given by the Boltzmann weight of the two states.

To start our analysis, Fig. S7A shows the predicted versus measured gene expression of the multivalent model (Eq. 5) for promoters with an UP element, where all data is plotted with low opacity except the promoters with the lowest or highest levels of predicted gene expression. The sharp left edge of the data is set by the background level of gene expression $r_0 = 0.18$ in the absence of RNAP, while the right edge represents the maximal expression $r_{\text{max}} = 8.6$ of very strong promoters. Note that if the scatter in data on the left edge is attributable to noise, then the outliers on the right edge cannot simply arise from noise, since there are 10x fewer data points and hence we expect 10x fewer outliers (although there are roughly the same number of outliers 2σ, 3σ, and 4σ away from the predicted values on both edges of the plot). This suggests that there is a mechanistic explanation for why the promoters that our model predicts should bind very tightly to RNAP exhibit low gene expression.

We next analyzed whether any promoters increased expression when their -35 or -10 sites were replaced by the consensus sequences, but exhibited decreased gene expression when both the -35 and -10 sites became the consensus sequences. Out of the 12,000 constructs, 850 exhibited this pattern of expression. One possible explanation is that although strong binding helps recruit RNAP to the promoter, overly strong binding could inhibit transcription initiation and decrease gene expression. We note, however, that this is a coarse-grained effective model that neglects molecular details of the transition from the closed to open complex, transcription initiation, and other critical steps of RNAP functioning (8). Nevertheless, in the multivalent model, the effect of overly strong RNAP-promoter binding must be to decrease the single parameter r_{max} (which represents the level of gene expression when RNAP is fully bound to the promoter), since no other parameters should depend upon the total RNAP-promoter binding strength.

To proceed, we assume that fully bound RNAP with free energy

$$\Delta E_{\text{RNAP}} = E_{\text{BG}} + E_{\text{Spacer}} + E_{\text{UP}} + E_{\text{-35}} + E_{\text{-10}} + E_{\text{int}} \tag{S15}$$

relative to the unbound state can initiate transcription by moving into a transcription initiation state with free energy ΔE_{trans} relative to the unbound state as shown schematically in Fig. S7B. Intuitively, bound RNAP will always immediately transcribe when $\Delta E_{\text{trans}} - \Delta E_{\text{RNAP}}$ is large and negative, but when the affinity between the RNAP and promoter becomes sufficiently strong (the case depicted in Fig. S7B), RNAP will prefer to stay bound to the promoter and not transcribe immediately. We posit that the rate of entering the transcribing state (8), and hence the rate of gene expression r_{max} in Fig. 1B, should be modified to

$$\tilde{r}_{\text{max}} \equiv \frac{r_{\text{max}} + r_0 e^{-\beta(\Delta E_{\text{RNAP}} - \Delta E_{\text{trans}})}}{1 + e^{-\beta(\Delta E_{\text{RNAP}} - \Delta E_{\text{trans}})}}, \tag{S16}$$

similar to recently proposed scrunching models of transcription initiation (9). For promoters whose RNAP binding is far weaker than transcription initiation ($e^{-\beta(\Delta E_{\text{RNAP}} - \Delta E_{\text{trans}})} \ll 1$), this rate reduces to the constant value r_{max}. Increasing the RNAP-promoter affinity decreases ΔE_{RNAP} which leads to a decrease in the level of gene expression. In the limit of an infinitely strong promoter ($\Delta E_{\text{RNAP}} \to -\infty$),

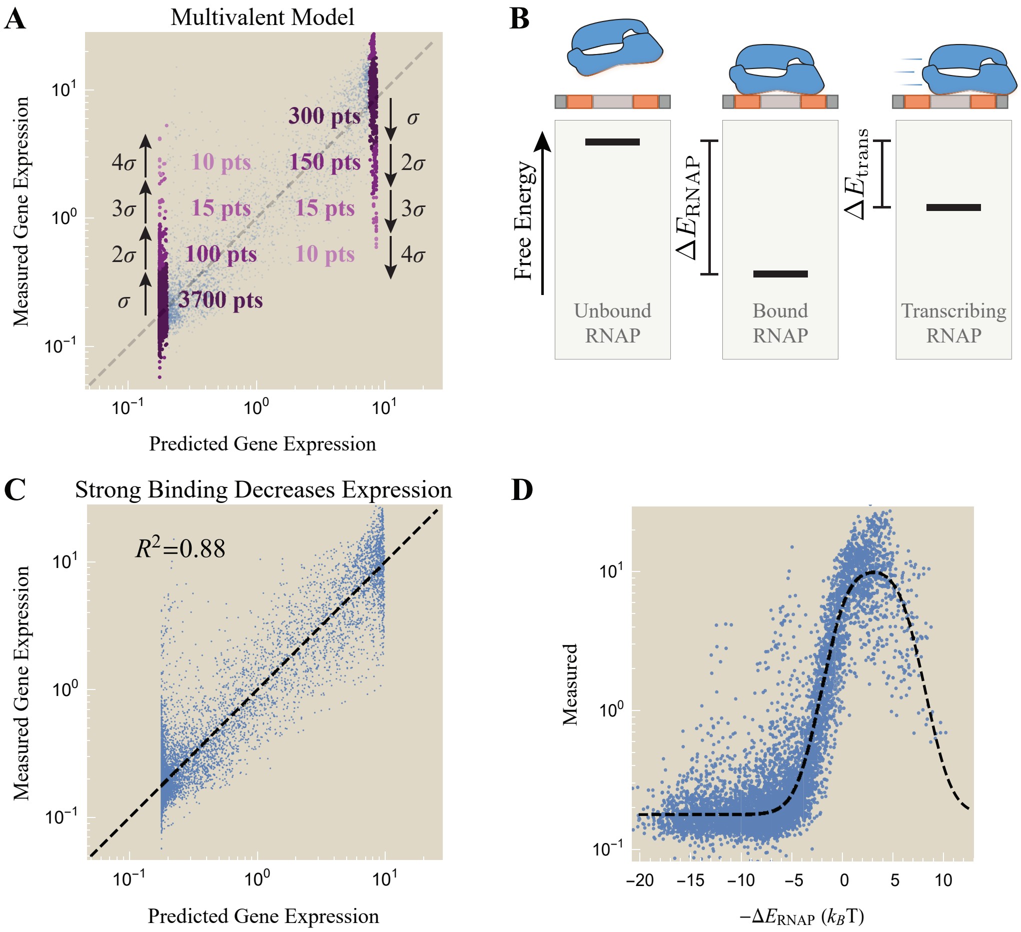


Figure S7. Gene expression is reduced for promoters that bind RNAP too tightly. (A) In the multivalent model (Eq. 5), although there are 10x fewer points on the right edge of the plot than the left edge, there are the same number of outliers, suggesting a biophysical mechanism for the reduction in gene expression of the strongest promoters. (B) The average level of transcription modeled as a two state system where the bound RNAP state (with free energy ΔE_{RNAP} relative to the unbound state) can enter a transcription initiation state with free energy $\Delta E_{\mathrm{trans}}$. (C) Gene expression characterized using the modified maximum level of gene expression using Eq. S16 with $\Delta E_{\mathrm{trans}} = -6.2\, k_B T$. (D) Measured gene expression versus the promoter strength ΔE_{RNAP} (stronger promoters on the right because of the minus sign). The dashed line shows the prediction of the multivalent model modified using Eq. S16.

RNAP is glued in place and unable to transcribe, thereby reducing the level of gene expression to the background level r_0.

Fig. S7C shows the gene expression data refit to the multivalent model with the maximal level of gene expression given by Eq. S16 (using $\Delta E_{\mathrm{trans}} = -6.2\, k_B T$ inferred by nonlinear regression). We note that using this model eliminates the sharp right edge of the data (red ellipse in Panel A), signifying that the promoters with extremely tight RNAP binding have shifted left, moving closer to the level of gene expression predicted by the model. Fig. S7D compares the predicted $-\Delta E_{\mathrm{trans}}$ for each promoter (using the best fit parameter in Table S1) against the measured level of gene expression. To facilitate a comparison with the multivalent model, we overlay this data with the approximate predicted level of gene expression

$$\mathrm{GE} \approx \frac{r_0 + \tilde{r}_{\mathrm{max}} e^{-\beta \Delta E_{\mathrm{RNAP}}}}{1 + e^{-\beta \Delta E_{\mathrm{RNAP}}}}, \tag{S17}$$

where we have ignored the two partially bound RNAP states and used the maximum level of gene expression in Eq. S16. Although only a small number of promoters exhibits sufficiently strong binding that diminishes their gene expression, the data exhibits a clear downwards trend in this limit.

F Dynamics of RNAP with Avidity

F.1 Probability of the RNAP States at Equilibrium

In this section, we derive the probabilities of the four RNAP states shown in Fig. 1C in equilibrium. RNAP may be unbound (concentration U), singly bound at the -35 site ($B_{\text{-35}}$), singly bound at the -10 site ($B_{\text{-10}}$), or bound to both sites ($B_{\text{-35,-10}}$). These concentrations must obey detailed balance,

$$B_{\text{-35}} = \frac{[\text{RNAP}]k_{\text{on}}}{k_{\text{off,-35}}}U \tag{S18}$$

$$B_{\text{-10}} = \frac{[\text{RNAP}]k_{\text{on}}}{k_{\text{off,-10}}}U \tag{S19}$$

$$B_{\text{-35,-10}} = \frac{\tilde{k}_{\text{on}}}{k_{\text{off,-10}}}B_{\text{-35}}, \tag{S20}$$

as well as the normalization condition

$$[\text{RNAP}] = U + B_{\text{-35}} + B_{\text{-10}} + B_{\text{-35,-10}}. \tag{S21}$$

In writing Eqs. S18 and S19, we have assumed a sufficiently large reservoir of RNAP so that binding to the promoter of interest does not appreciably decrease the concentration of free RNAP (a reasonable assumption in *E. coli* where there are ≈ 2000 RNAP molecules (10)).

Eqs. S18-S21 can be solved to obtain the concentration of each RNAP state, namely,

$$U = \frac{K_{\text{-35}}K_{\text{-10}}}{K_{\text{-35}}K_{\text{-10}} + K_{\text{-35}}[\text{RNAP}] + K_{\text{-10}}[\text{RNAP}] + c_0[\text{RNAP}]}[\text{RNAP}] \tag{S22}$$

$$B_{\text{-35}} = \frac{K_{\text{-35}}[\text{RNAP}]}{K_{\text{-35}}K_{\text{-10}} + K_{\text{-35}}[\text{RNAP}] + K_{\text{-10}}[\text{RNAP}] + c_0[\text{RNAP}]}[\text{RNAP}] \tag{S23}$$

$$B_{\text{-10}} = \frac{K_{\text{-10}}[\text{RNAP}]}{K_{\text{-35}}K_{\text{-10}} + K_{\text{-35}}[\text{RNAP}] + K_{\text{-10}}[\text{RNAP}] + c_0[\text{RNAP}]}[\text{RNAP}] \tag{S24}$$

$$B_{\text{-35,-10}} = \frac{c_0[\text{RNAP}]}{K_{\text{-35}}K_{\text{-10}} + K_{\text{-35}}[\text{RNAP}] + K_{\text{-10}}[\text{RNAP}] + c_0[\text{RNAP}]}[\text{RNAP}], \tag{S25}$$

where we have defined the dissociation constants $K_j = \frac{k_{\text{off},j}}{k_{\text{on}}}$ of free RNAP binding to the site j as well as the effective concentration $c_0 = \frac{\tilde{k}_{\text{on}}}{k_{\text{on}}}$ of singly bound RNAP binding to the remaining promoter site. If we further define the effective dissociation constant

$$K_D^{\text{eff}} = \frac{K_{\text{-35}}K_{\text{-10}}}{c_0 + K_{\text{-35}} + K_{\text{-10}}}, \tag{S26}$$

we can rewrite the probability of the unbound state as

$$U = \frac{K_D^{\text{eff}}}{K_D^{\text{eff}} + [\text{RNAP}]}[\text{RNAP}]. \tag{S27}$$

From this equation, we see that the promoter is bound 50% of the time ($U = \frac{[\text{RNAP}]}{2}$) when $[\text{RNAP}] = K_D^{\text{eff}}$, as stated in the main text.

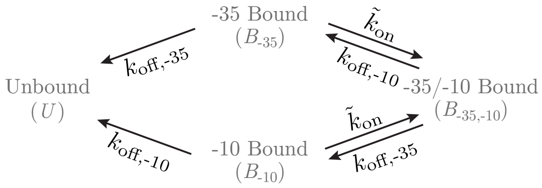


Figure S8. Dynamics of RNAP unbinding from the -35 and -10 sites. The avidity between the -35 and -10 sites will prolong the time before RNAP unbinds from the promoter.

F.2 Dynamics of RNAP Unbinding with Avidity

Here, we rederive the results from the previous section by analyzing the dynamics of RNAP binding rather than its equilibrium configuration. This calculation highlights the intimate connection between the effective dissociation constant in Eq. S26 and the kinetics of RNAP binding.

To that end, we first compute the probability that a bound RNAP will remain bound after a time t. Since we are only interested in the unbinding process, we consider the rates diagram in Fig. S8 where the on-rates from the unbound state have been removed. Following Ref. (11), we assume that the three bound states – RNAP bound to only the -35 site (concentration $B_{\text{-35}}$), only the -10 site ($B_{\text{-10}}$), or to both sites ($B_{\text{-35,-10}}$) – quickly equilibrate and compute the effective off-rate from these bound states to the RNAP unbound state (U). If the three bound states are in equilibrium, then there is no flux between any two states, namely,

$$\tilde{k}_{\text{on}} B_{\text{-35}} = k_{\text{off,-10}} B_{\text{-35,-10}} \tag{S28}$$

and

$$\tilde{k}_{\text{on}} B_{\text{-10}} = k_{\text{off,-35}} B_{\text{-35,-10}}. \tag{S29}$$

The total concentration of bound RNAP is given by

$$[\text{RNAP}]_{\text{bound}} = B_{\text{-35}} + B_{\text{-10}} + B_{\text{-35,-10}} = B_{\text{-35,-10}} \left(1 + \frac{k_{\text{off,-35}}}{\tilde{k}_{\text{on}}} + \frac{k_{\text{off,-10}}}{\tilde{k}_{\text{on}}}\right). \tag{S30}$$

The loss of bound RNAP is caused by unbinding from the two singly bound forms, leading to the effective off-rate

$$\begin{aligned} \frac{d}{dt}[\text{RNAP}]_{\text{bound}} &\equiv -k_{\text{off}}^{\text{eff}}[\text{RNAP}]_{\text{bound}} && \text{(S31)} \\ &= -k_{\text{off,-35}} B_{\text{-35}} - k_{\text{off,-10}} B_{\text{-10}} && \text{(S32)} \\ &= -\frac{2k_{\text{off,-35}} k_{\text{off,-10}}}{\tilde{k}_{\text{on}} + k_{\text{off,-35}} + k_{\text{off,-10}}}[\text{RNAP}]_{\text{bound}}. && \text{(S33)} \end{aligned}$$

Hence, the dynamics of RNAP unbinding are characterized by

$$[\text{RNAP}]_{\text{bound},t} = [\text{RNAP}]_{\text{bound},0} e^{-k_{\text{off}}^{\text{eff}} t} \tag{S34}$$

where the likelihood of remaining bound decreases exponentially according to the timescale $\tau = \frac{1}{k_{\text{off}}^{\text{eff}}}$.

Lastly, to connect this result to the calculations in the previous section, we return to the full model in Fig. 1C where unbound RNAP can associate onto the promoter. As in simple monovalent ligand-receptor systems, the effective dissociation constant Eq. S26 is related to the off-rate from the bound to unbound states ($k_{\text{off}}^{\text{eff}}$) divided by the on-rate from the unbound to bound states ($2k_{\text{on}}$), namely,

$$K_D^{\text{eff}} = \frac{k_{\text{off}}^{\text{eff}}}{2k_{\text{on}}}. \tag{S35}$$

References

[1] Bintu L, et al. (2005) Transcriptional regulation by the numbers: Models. *Curr Opin Genetics Dev* 15:116–124.

[2] Wang F, et al. (2013) The promoter-search mechanism of *Escherichia coli* RNA polymerase is dominated by three-dimensional diffusion. *Nat Struct Mol Biol* 20:174–181.

[3] Scholes C, et al. (2017) Combinatorial gene regulation through kinetic control of the transcription cycle. *Cell Syst* 4:97–108.e9.

[4] Urtecho G, Tripp AD, Insigne KD, Kim H, Kosuri S (2019) Systematic dissection of sequence elements controlling σ70 promoters using a genomically encoded multiplexed reporter assay in *Escherichia coli*. *Biochem* 58:1539–1551.

[5] Feklístov A, Sharon BD, Darst SA, Gross CA (2014) Bacterial sigma factors: A historical, structural, and genomic perspective. *Annu Rev Microbiol* 68:357–376.

[6] Yona AH, Alm EJ, Gore J (2018) Random sequences rapidly evolve into de novo promoters. *Nat Commun* 9:1530.

[7] Ellinger T, Behnke D, Bujard H, Gralla JD (1994) Stalling of *Escherichia coli* RNA polymerase in the +6 to +12 region in vivo is associated with tight binding to consensus promoter elements. *J Mol Biol* 239:455–465.

[8] Ruff EF, Record MT, Artsimovitch I (2015) Initial events in bacterial transcription initiation. *Biomolecules* 5:1035–1062.

[9] Henderson KL, et al. (2017) Mechanism of transcription initiation and promoter escape by *E. coli* RNA polymerase. *Proc Natl Acad Sci USA* 114:E3032–E3040.

[10] Klumpp S, Hwa T (2008) Growth-rate-dependent partitioning of RNA polymerases in bacteria. *Proc Natl Acad Sci USA* 105:20245–20250.

[11] Stone JD, et al. (2011) Interaction of streptavidin-based peptide-MHC oligomers (tetramers) with cell-surface tcrs. *J Immunol* 187:6281–6290.