Supplementary Information for "Automated Design of Synthetic Ribosome Binding Sites to Precisely Control Protein Expression"

Howard M. Salis[1], Ethan A. Mirsky[2], and Christopher A. Voigt[1*]

[1]Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA, 94158
[2]Graduate Group in Biophysics, University of California San Francisco, San Francisco, CA, 94158

*Corresponding author: cavoigt@picasso.ucsf.edu

Author email addresses: salish@picasso.ucsf.edu / howard.salis@gmail.com
ethan.mirsky@ucsf.edu

Table of Contents

1. Derivation of the Statistical Thermodynamic Model

Inside a cell, each mRNA transcript competes for ribosomal 30S complex binding and translation initiation. This competition is enhanced because most ribosomes are bound to mRNA and engaged in translation elongation, thus leaving much fewer free 30S complexes. Combined with rapid ribosome turnover, these conditions allow us to assume that the population of mRNA transcripts is in chemical equilibrium with the freely available 30S complexes. Accordingly, we

can derive an equation that relates the translation initiation rate of a transcribed protein coding sequence to its predicted Gibbs free energy change upon ribosome binding (G_{tot}). Equation 1 is a simplification of this equation.

Consider a bacterial cell with m_i copies of an mRNA transcript and R copies of freely available 30S complex. On an mRNA transcript, the i^{th} ribosome binding site sequence controls the translation initiation rate of a protein coding sequence, denoted by $mRNA_i$. The mRNA transcript and 30S complex form a bound complex, $mRNA_i$::30S, with C_i copies. The association reaction and equilibrium condition for each mRNA transcript is

$$\mathrm{mRNA_i} + 30\mathrm{S} \leftrightarrow \mathrm{mRNA_i} :: 30\mathrm{S} \qquad C_i = m_i R \exp\left(-\beta \Delta G_i\right) \tag{S1}$$

where G_i is the total Gibbs free energy change (G_{tot}) when the ribosome binds to the i^{th} ribosome binding site sequence on the mRNA transcript. The total amount of 30S complex R_{tot} is the sum of the free and mRNA-bound ribosomes, which is

$$R_{tot} = R + \sum_j C_j = R\left(1 + \sum_j m_j \exp\left(-\beta \Delta G_j\right)\right) \tag{S2}$$

which can be rearranged to give the free amount of 30S complex:

$$R = \frac{R_{tot}}{\left(1 + \sum_j m_j \exp\left(-\beta \Delta G_j\right)\right)} \tag{S4}$$

Substituting Equation S4 into Equation S1 and rearranging, we obtain a relationship between C_i and the Gibbs free energies of ribosome binding to each ribosome binding site sequence inside the cell. This relationship is

$$C_i = \frac{m_i R_{tot} \exp\left(-\beta \Delta G_i\right)}{1 + \sum_j m_j \exp\left(-\beta \Delta G_j\right)} \tag{S5}$$

We then make two assumptions: (i) the total amount of 30S complex, R_{tot}, is constant; and (ii) the amount of 30S complex bound to ribosome binding site sequence upstream of a protein coding sequence is proportional to its translation initiation rate. Using these two assumptions, we simplify Equation S5 to:

$$r_i \propto \frac{m_i R_{tot} \exp\left(-\beta \Delta G_i\right)}{1 + \sum_j m_j \exp\left(-\beta \Delta G_j\right)} \tag{S6}$$

where the j summation is performed over all transcribed protein coding sequences.

Equation S6 relates the translation initiation rate of the i^{th} ribosome binding site and protein coding sequence to the Gibbs free energy change when the 30S complex binds (G_i). This equation describes the translation initiation rates of "the ribosome ensemble" – the pool of

ribosomes that compete with one another to bind to mRNA transcripts, each according to their free energies of binding.

The summation in the denominator of Equation S6 is very large. We can assume that denominator is a constant when we modify the cell to produce a relatively small amount of additional mRNA transcripts with modest translation initiation rates. Using this assumption, we can simplify Equation S6 to:

$$r_i = K \exp(-\beta \Delta G_i) \tag{S7}$$

with a proportionality constant K that includes m_i and R_{tot}.

When we modify a cell to produce many additional mRNA transcripts with high translation initiation rates, we should expect that the denominator of Equation S6 will increase, invalidating our assumption. Under this scenario, Equation S6 predicts that the translation initiation rate of the additional mRNA transcripts will remain high, but that the translation initiation rates of all other mRNA transcripts will decrease due to ribosome competition. The resulting global slowdown in protein production has many phenotypic effects, including a longer doubling time.

We use the Mfold 3.0 free energy model of RNA–RNA interactions [1-3] and the NuPACK suite of algorithms, developed by Pierce and co-workers[4], to compute the Gibbs free energy change upon ribosome binding ΔG_i for each ribosome binding site and protein coding sequence. The Mfold 3.0 model was created by characterizing the folding and hybridization of RNA molecules in an *in vitro* environment [1-3, 5-13]. We use the free energy model to predict the translation initiation rate inside a bacterial cell.

2. Testing the Independent control of transcription and translation rates

We demonstrate that the transcription rate of a promoter and the translation initiation rate of a ribosome binding site sequence have the potential to be independently and differentially controlled. The combination of three constitutive σ^{70} promoters (BioBrick numbers J23100, J23108, and J23114) and five different synthetic ribosome binding site sequences are inserted into the measurement system (**Methods**). BioBrick parts can be found at http://partsregistry.org/Main_Page. The synthetic ribosome binding site sequences have predicted ΔG_{tot} values of 7.15, 2.66, –2.58, –5.34, and –6.99 kcal/mol. The protein expression levels of the 15 sequences are measured using the previously described protocol of growth and flow cytometry analysis (**Methods**).

Considering Equation 1 from the main text, we show that changes to the transcription rate result in an approximately proportional change in the protein expression level by only altering the proportionality factor K (**Figure S1**). At the same time, choosing a ribosome binding site sequence with a different predicted ΔG_{tot} results in a proportional change in the protein expression level by only altering the relative translation initiation rate, which is $\exp(-\beta \Delta G_{tot})$. Thus, these two quantities have the potential to be controlled independently. However, a method for predicting the promoter sequence that achieves a user-selected transcription rate is not currently available.

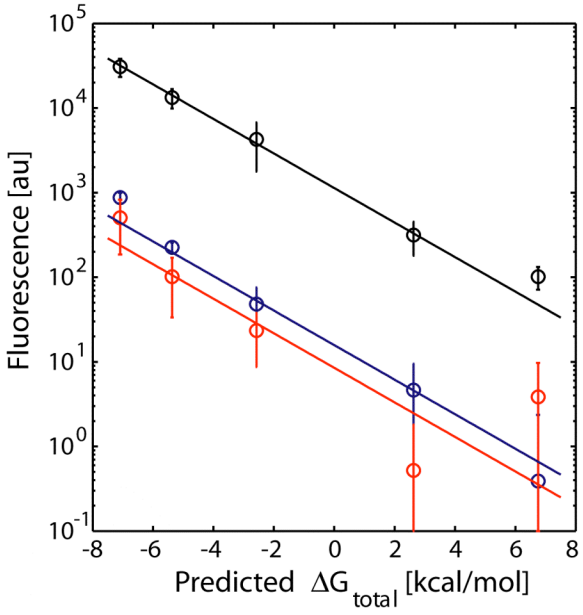

Figure S1: Independent control of transcription and translation rate. Three constitutive σ^{70} promoters (BioBrick numbers J23100, J23108, and J23114) vary the transcription rate of the mRFP1 protein coding sequence. For each promoter, the same five synthetic ribosome binding site sequences vary the translation initiation rates. Protein expression levels were measured in the measurement system. For each promoter, the proportionality factor K that best fits the data is determined, using an experimentally measured $\beta = 0.45$ mol/kcal and each ribosome binding site sequence's predicted G_{tot}. Dashed lines indicate the protein expression levels E for each proportionality constant K while varying the predicted G_{tot}. The best-fit K values are 1300 (black), 16.4 (blue), and 12.7 (red) for the J23100, J23108, and J23114 promoters, respectively.

3. Developing a Model of $G_{spacing}(s)$

We characterized the relationship between the free energy penalty $G_{spacing}$ and the aligned spacing [14] s by designing thirteen synthetic RBSs where the aligned spacing is varied from 0 to 15 nucleotides while verifying that the $G_{mRNA:rRNA}$, G_{mRNA}, G_{start}, and $G_{standby}$ free energies remain constant (**Supplementary Table I**). The protein fluorescences of these sequences are measured in the fluorescent protein measurement system (**Supplementary Figure S2**). From this data, we can infer the $G_{spacing}$ values at each value of the aligned spacing and empirically fit these values to the following formulas. When the 30S complex is stretched (s > 5 nt), the $G_{spacing}$ is fit to a quadratic equation,

$$\Delta G_{spacing} = c_1 (s - s_{opt})^2 + c_2 (s - s_{opt}), \quad (7)$$

where $s_{opt} = 5$ nt, $c_1 = 0.048$ kcal/mol/nt^2, and $c_2 = 0.24$ kcal/mol/nt. When the 30S complex is compressed (s < 5 nt), the $G_{spacing}$ is fit to a sigmoidal function,

$$\Delta G_{spacing} = \frac{c_1}{\left[1 + \exp\left(c_2 (s - s_{opt} + 2)\right)\right]^3}, \quad (8)$$

where c_1 = 12.2 kcal/mol and c_2 = 2.5 nt $^{-1}$. The above parameter values are determined by minimizing the difference between the $G_{spacing}$ values calculated from the experimental measurements (**Supplementary Figure S3**) and the evaluation of Equation 7 or 8. Given any ribosome binding site sequence with an aligned spacing *s*, we calculate the $G_{spacing}$ accordingly.

We tested the model's ability to predict the $G_{spacing}$ of synthetic ribosome binding sites when factors besides the aligned spacing were also changing. We designed 11 additional synthetic ribosome binding site sequences whose aligned spacing varied from s = 0 to s = 15 nt. These sequences formed different mRNA secondary structures and contained different 16S rRNA binding sites. Using the model, we predicted the $G_{spacing}$ for each sequence. The protein expression levels of these sequences were then measured in the measurement system. The apparent $G_{spacing}$ for each sequence was calculated according to $G_{spacing} = -\beta^{-1}\log(E/K) - G_{tot(no\ spacing)}$, where K = 2500 au, β = 0.45 mol/kcal, and $G_{tot(no\ spacing)}$ is result of the free energy model absent the spacing term, which is $G_{tot(no\ spacing)} = (G_{mRNA:rRNA} + G_{start}) - (G_{mRNA:rRNA} - G_{standby})$. The proportionality constant K and Boltzmann factor β were experimentally determined using the data shown in Figures 2 and 3.

We then compared the predicted $G_{spacing}$ and apparent $G_{spacing}$, which is shown in **Figure S2**. The squared correlation coefficient between the predicted and apparent $G_{spacing}$ is R^2 = 0.77. Most of the error arises from spacing values of s = 0 and 1 nt, where the $G_{spacing}$ > 10 kcal/mol, suggesting that the $G_{spacing}$ penalty for these values of s are being slightly underestimated. At extremely non-optimal values of aligned spacing, the protein expression level is very low. Thus, the error in the model parameterization arises from the experimental error in measuring such low expression levels.

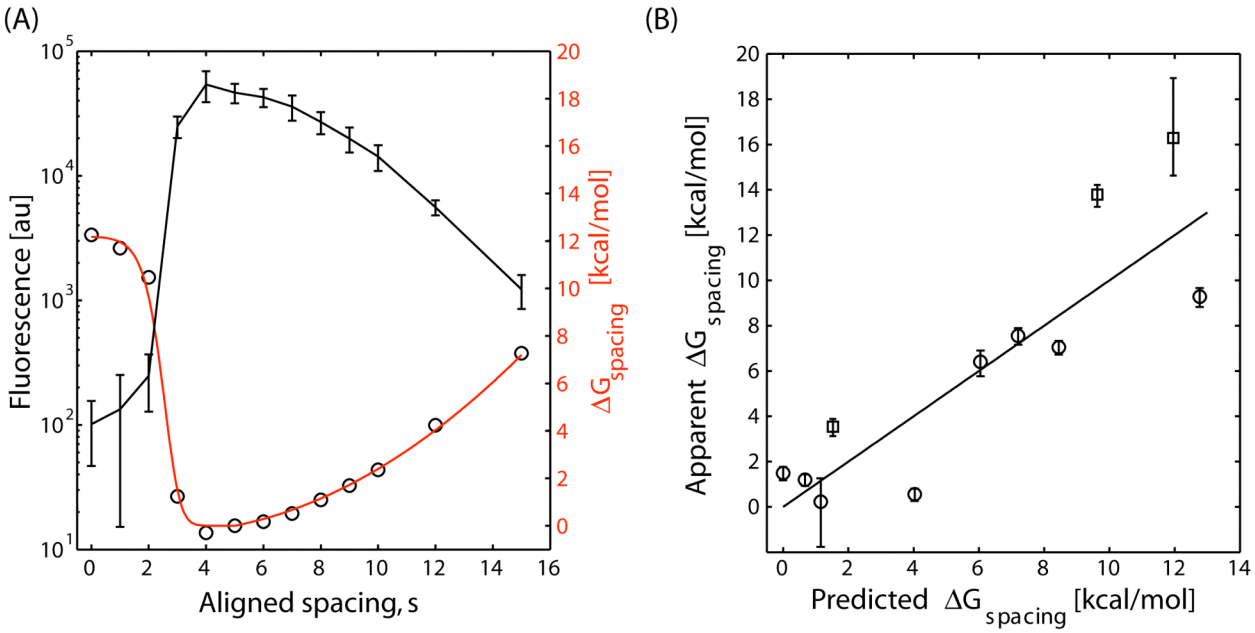

Figure S2: A model of the RBS spacing penalty, $G_{spacing}$. (A) The protein fluorescence levels from 13 synthetic RBSs with varying aligned spacing are measured. The differences in $G_{spacing}$ (black circles, right axis) are then inferred from the fluorescence data (black points with error bars, left axis). A quantitative model is created by fitting the values of $G_{spacing}$ to either a

quadratic or sigmoidal function (red line) for s > 5 nt or s < 5 nt, respectively. The error bars are calculated as the standard deviation of at least 5 measurements performed on two different days. (B) Using an additional set 11 synthetic RBS sequences, a comparison between the predicted $G_{spacing}$ and the apparent $G_{spacing}$ is performed where the aligned spacing is s > 5 nt (circles) or s < 5 nt (squares). Error bars are calculated as the standard deviation of six measurements performed on two different days.

4. Identifying an Accurate RNA Sequence Cutoff

When calculating the G_{tot}, we consider a subsequence of the mRNA transcript surrounding the start codon. The length of the subsequence is determined by a cutoff value. The thermodynamic model uses a cutoff value of 35 nucleotides before and after the start codon, resulting in a 70 nt mRNA subsequence. Below, we recalculate the average error of the thermodynamic model when using different cutoff values, utilizing a dataset that combines the reverse and forward engineered ribosome binding sites shown in **Figure 2**. The cutoff value does not significantly change the model's predictions between 15 and 50 nt (**Figure S3**).

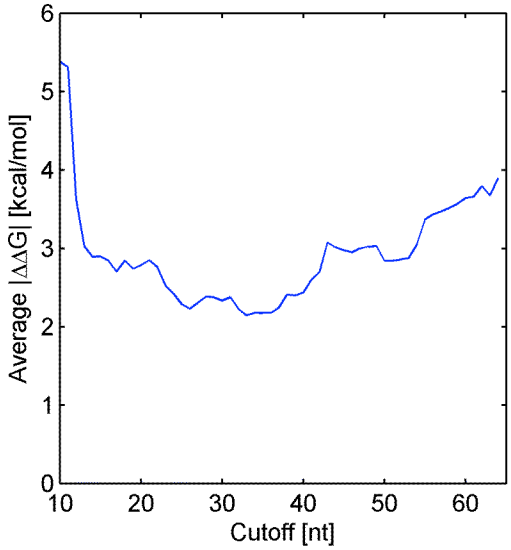

Figure S3: The sensitivity of the model's accuracy to the RNA subsequence cutoff is shown for 81 synthetic ribosome binding site sequences. A cutoff of 35 nucleotides was used to generate these sequences.

5. Growth Rate Measurements of Ribosome Binding Site Sequences

The measurement system is a ColE1 plasmid with chloramphenicol resistance (derived from pProTet, Clontech). The expression cassette contains a σ^{70} constitutive promoter (BioBrick #J23100), a sequence containing a ribosome binding site, followed by the mRFP1 fluorescent protein reporter. XbaI (TCTAGA) and SacI (GAGCTC) restriction sites are located before the ribosome binding site and after the start codon. The annotated vector sequence containing an example ribosome binding site sequence is included in GenBank format as a **Supplementary Data**. Each ribosome binding site sequence is inserted in between the XbaI and SacI sites as described in the main text. A plasmid map of the measurement system is shown in **Figure S9**.

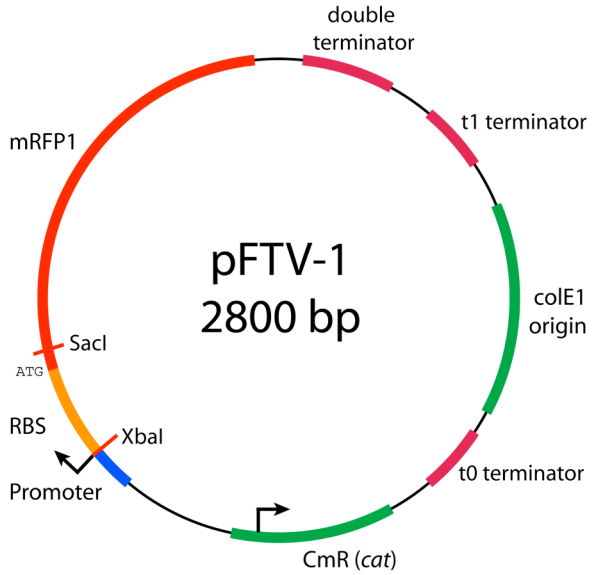

Figure S9: A plasmid map of the vector containing the measurement system.

The doubling times of DH10B *E. coli* cultures were measured while grown in supplemented M9 media (0.4% glucose, 0.05 g/L L-leucine) in a shaking 96-well plate. The average doubling time is about 2 hours. All doubling time data is located in the Supplementary Table. In **Figure S4**, we show the predicted G_{tot} and the log protein fluorescence for the natural and synthetic ribosome binding sites (**Figure 2**) compared to their measured doubling times. For these sequences, there is no observed correlation between the predicted or actual protein expression level and the growth rate of its culture. Consequently, the dilution rate of protein within each culture is approximately constant and the expression level of a protein is proportional to the translation initiation rate of its mRNA transcript.

It is possible to design a very strong ribosome binding site sequence that expresses the mRFP1 protein such that the growth rate does appreciably decrease. We designed and tested 7 such sequences, which are denoted in the **Supplementary Table**. These data points were not used to determine the accuracy of the method. No correlation between ribosome binding site sequence and growth rate is observed when the protein expression level is predicted to be less than about 75,000 au ($G_{tot} > -7.5$ kcal/mol).

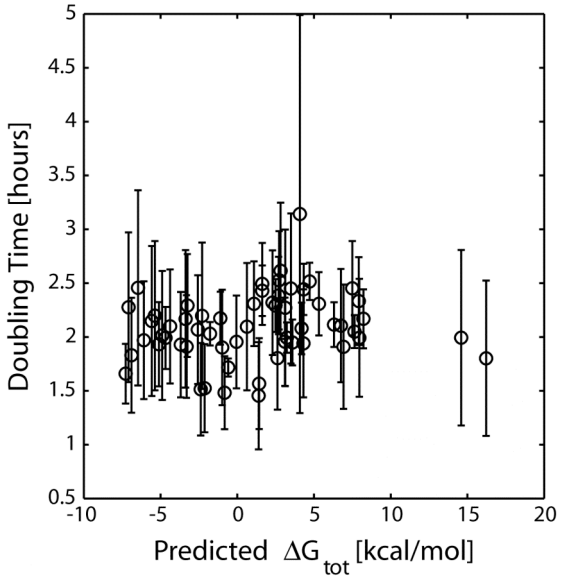

Figure S4: The doubling times of bacterial strains expressing mRFP1 with different ribosome binding site sequences. Their predicted G_{tot} values are shown. Error bars are calculated from the standard deviation of at least two time-courses performed on two different days.

6. A Quantitative Model of the AND Gate Genetic Circuit

The AND gate genetic circuit is a three gene system with two input promoters and one output gene. The first input promoter controls the expression of the amber suppressor tRNA *supD* while the second input promoter controls T7 RNA polymerase expression. The output *gfp* gene is controlled by a T7 promoter. The T7 RNA polymerase contains two amber stop codons. Without the expression of the *supD* tRNA, translation elongation of the T7 RNA polymerase transcript will result in a truncated, non-functional protein. When both the *supD* tRNA and T7 RNA polymerase are expressed, functional T7 RNA polymerase proteins are produced and transcription can initiate from the T7 promoter, turning on the *gfp* output. Thus, the expression of *gfp* requires expression from both input promoters.

For this example, we use the P_{sal} and P_{BAD} sensors as inputs into the AND gate genetic circuit. These promoters are capable of responding to arabinose and salicylate by binding to the AraC and NahR transcription factors, respectively.

We use a previously developed quantitative model of AND gate genetic circuit function [15] to predict the *gfp* expression in response to the expression levels of the input promoters. The quantitative model takes the form of three transfer functions. The first two transfer functions quantify the expression levels of the P_{sal} and P_{BAD} input promoters as a function of their respective inducer concentrations while at steady-state. The form of these transfer functions are:

$$I_1 = g_{1,ref}\varphi_1([\text{Sal}]), \tag{S8}$$

$$I_2 = g_{2,ref}\varphi_2([\text{Ara}]), \tag{S9}$$

where I_1 and I_2 are the expression levels of the P_{sal} and P_{BAD} input promoters when a concentration of salicylate, [Sal], and arabinose, [Ara], are present in the media, respectively.

Here, $g_{1,\mathrm{ref}}$ and $g_{2,\mathrm{ref}}$ are the maximum expression levels (called the *gain*) obtainable using the P_{sal} and P_{BAD} promoters while φ_1 and φ_2 are the fraction of maximal expression at each concentration of inducer (called the normalized transfer function, which varies between zero and one). These functions were experimentally characterized by varying the inducer concentration and measuring fluorescence at steady-state (**Figure S8**) followed by fitting to a generalized polynomial function. Changing the ribosome binding site sequence in front of the P_{BAD} promoter may change the gain g of its transfer function, but not its normalized transfer function, φ.

The third transfer function quantifies the activation of the AND gate genetic circuit, measured by *gfp* expression level χ_{O}, in response to the expression levels of its input promoters, I_1 and I_2. The transfer function for the AND gate genetic circuit was experimentally characterized [15] and fit to the following polynomial:

$$\chi_{\mathrm{O}} = g_{\mathrm{O,ref}} \frac{\mathrm{I}_1 \mathrm{I}_2^2}{a\left(b + \mathrm{I}_2\right)^2 + \mathrm{I}_1 \mathrm{I}_2^2} \tag{S10}$$

where $a = 50 \pm 20$ and $b = 3000 \pm 1000$. Given a concentration of salicylate and arabinose, the input promoters' expression levels I_1 and I_2 are determined according to Equation S8 and S9. These values are then substituted into Equation S10 to predict the output *gfp* expression.

The ability of the AND gate genetic circuit to carry out AND logic is quantified by the use of a fitness function. The fitness of the AND gate genetic circuit is defined as highest when it turns on *gfp* expression only when both I_1 and I_2 are at their maximal values. The fitness of the genetic circuit decreases if *gfp* expression occurs otherwise. In this work, we define the following fitness function to quantify AND logic:

$$F(g_1, g_2) = \max\left(0, \varphi_{\mathrm{O}}(g_1, g_2) - \left(10^{\varphi_{\mathrm{O}}(g_1 \alpha_{1,\min}, g_2)} - 1\right) - \left(10^{\varphi_{\mathrm{O}}(g_1, g_2 \alpha_{2,\min})} - 1\right) - \left(10^{\varphi_{\mathrm{O}}(g_1 \alpha_{1,\min}, g_2 \alpha_{2,\min})} - 1\right)\right) \tag{S11}$$

where g_1 and g_2 are the gains of the P_{sal} and P_{BAD} input promoters, φ_1 and φ_2 are their normalized transfer functions, and $\alpha_{1,\min}$ and $\alpha_{2,\min}$ are the minimums of the normalized transfer functions. Leaky transcription from a promoter results in a positive $\alpha_{1,\min}$ or $\alpha_{2,\min}$. This function uses a power of 10 amplification of any *gfp* expression under undesirable input conditions to greatly reduce the fitness of a circuit that produces OR-like rather than AND-like logic.

In this study, we focus on changing the gain g_2 by manipulating the ribosome binding site sequence downstream of the P_{BAD} promoter. Therefore, the gain of the P_{sal} remains fixed at $g_1 =$ 11500 au throughout. Varying only the gain g_1, there is an optimal region of fitness from $g_2 =$ 320 to 1940 au. The goal of the design method is to generate a synthetic ribosome binding site sequence that modifies the gain of the P_{BAD} promoter to within this optimum region.

Here, we derive an equation that relates the predicted $\mathrm{G}_{\mathrm{tot}}$ of a synthetic ribosome binding site sequence to the gain of a promoter. To do this, we first measure the maximum expression level of the promoter with a known ribosome binding site sequence. We call this the reference gain, g_{ref}. We then use the reverse engineering mode of the design method to predict the $\mathrm{G}_{\mathrm{tot}}$ of the ribosome binding site sequence. We call this the reference $\mathrm{G}_{\mathrm{tot}}$, or $\mathrm{G}_{\mathrm{ref}}$. We use these two quantities in the following derivation. First, we use Equation 1 to relate the protein expression level E to the predicted $\mathrm{G}_{\mathrm{tot}}$ with a proportionality factor K and an experimentally measured $\beta =$

0.45 mol/kcal. We then equate the ratio of the variable gain g to the reference gain g_{ref} with the ratio of the variable expression level E to the reference expression level E_{ref}, yielding:

$$\frac{g}{g_{ref}} = \frac{E}{E_{ref}} = \exp\left(-\beta\left(\Delta G_{tot} - \Delta G_{ref}\right)\right) \quad \text{(S12)}$$

Notably, the proportionality factor K is eliminated from this relationship, which is rearranged to become Equation 3 in the main text, or:

$$g = g_{ref} \exp\left(-\beta\left(\Delta G_{tot} - \Delta G_{ref}\right)\right) \quad \text{(S13)}$$

The reference gain g_{ref} is obtained from the transfer function of the promoter. The reference free energy ΔG_{ref} is predicted by inputting the promoter's ribosome binding site sequence into the thermodynamic model. After this initial characterization, this simple equation allows us to relate any changes to the promoter's ribosome binding site sequence to its gain g.

Using Equation S13, we convert the optimal gain range to an optimal ΔG_{tot} range, with $g_{2,ref}$ = 590 au and ΔG_{ref} = –1.05 kcal/mol. The optimal ΔG_{tot} range is –1.7 kcal/mol ± 2.0 kcal/mol. We then used the forward engineering mode of the design method to generate 2 synthetic ribosome binding site sequences within this optimal region (ΔG_{tot} = –1.48 and –1.15 kcal/mol). We also generated 7 additional ribosome binding site sequences with predicted ΔG_{tot} values of 17.17, 12.30, 7.35, 3.45, 1.82, 2.18, and 0.60 kcal/mol. These ribosome binding site sequences were each inserted into the pBACr-AraT7940 plasmid upstream of the T7 RNA polymerase coding sequence. DH10B *Escherichia coli* cells were then transformed with pBR939b, pAC-SalSer914, and this modified pBACr-AraT7940 plasmid to assemble variants of the AND gate genetic circuit.

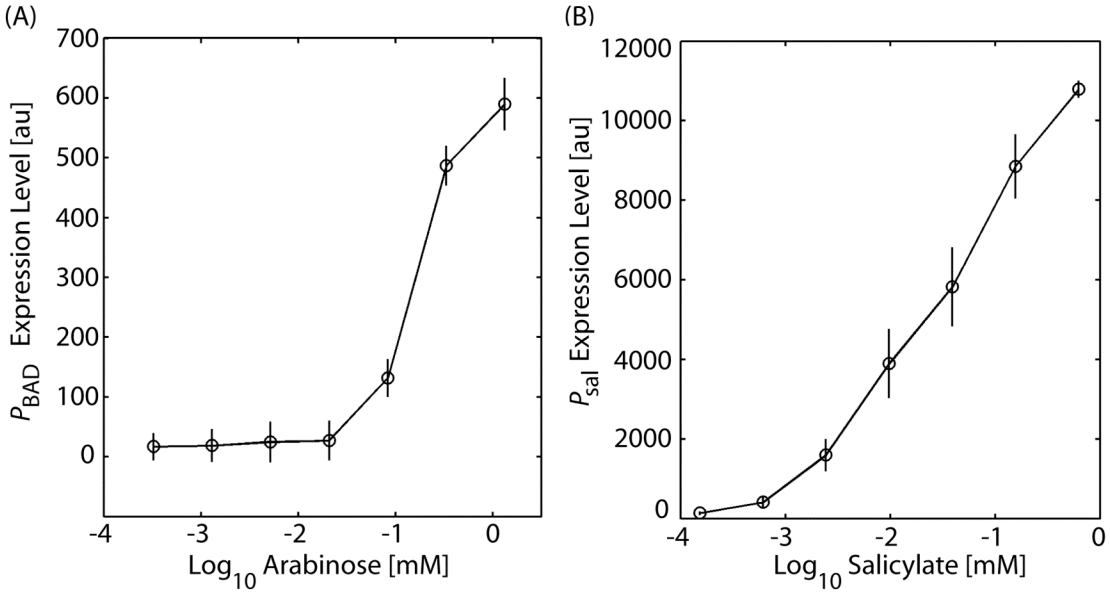

Figure S8 The transfer functions for the (A) P_{BAD} and (B) P_{sal} promoters in response to arabinose and salicylate, respectively. The transfer functions were experimentally characterized in a previous study[5].

7. Probabilities of Design Success and Failure

In the forward engineering mode, the error in the thermodynamic calculation of ΔG_{tot}, or ΔG, is fit by a Gaussian distribution with a standard deviation σ = 2.44 kcal/mol (**Figure 2F**). Consequently, the distribution of the protein expression level is a log-normal distribution with median $\exp(-\beta \Delta G_{tot})$. According to its cumulative distribution function, the probability that a synthetic ribosome binding site achieves the targeted protein expression to within a factor δ is

$$p_s(E/\delta < E < E\delta) = erf\left(\frac{\log(\delta)}{\sqrt{2}\beta\sigma}\right) \tag{S14}$$

where p_s is the probability of success, E is the target expression level, δ is the fold error, σ is 2.44 kcal/mol, β is 0.45 mol/kcal, and *erf* is the error function. To increase the probability of success, multiple synthetic ribosome binding site sequences can be designed with identical target ΔG_{tot}'s. The probability that *at least* one of N synthetic ribosome binding site sequences targets the protein expression level to within a factor of δ is calculated according to a Binomial distribution,

$$P_N = \sum_{k=1}^{N} \binom{N}{k} p_s^k (1-p_s)^{N-k} \tag{S15}$$

The probability of success P_N with a two-fold error ($\delta = 2$) is 0.47 for one ribosome binding site sequence (N = 1). The probability increases to 0.72, 0.85, or 0.92 when two, three, or four ribosome binding site sequences are generated (N = 2, 3, or 4).

We can use Equations S14 and S15 to calculate the probability of successfully connecting two genetic circuits together. As an example, we calculate the probability of successfully connecting the P_{sal} and P_{BAD} sensors to the AND gate genetic circuit. In section 5, a quantitative model of the genetic system was used to create a fitness curve $F(\Delta G_{tot})$ that relates the P_{BAD} promoter's ribosome binding site sequence to the ability of the genetic system to carry out AND logic. Using the fitness curve, we define "success" as creating a genetic system with a fitness value above a minimum threshold, F_{crit} (**Figure S11A**). We can then determine the optimal ΔG_{tot} that maximizes the fitness and the maximum δ-fold error in the predicted protein expression level that results in $F(\Delta G_{tot}) \geq F_{crit}$. By substituting δ into Equation S15, we obtain the probability that a synthetic ribosome binding site sequence designed with the optimal ΔG_{tot} will successfully connect the P_{BAD} and AND gate genetic circuits. To obtain a 90% confidence level, we use Equation S15 to calculate how many synthetic ribosome binding site sequences are needed to obtain a probability of at least 0.90.

According to the fitness function (Equation S11), the optimal ΔG_{tot} is −1.17 kcal/mol. Selecting an F_{crit} of 0.60, the maximum fold error δ is 2.21 and three synthetic ribosome binding site sequences are needed to successfully connect the genetic circuits with a 90% confidence level (**Figure S11**). As the fitness threshold increases, the maximum fold error decreases and more synthetic ribosome binding site sequences must be attempted to guarantee the same probability of success.

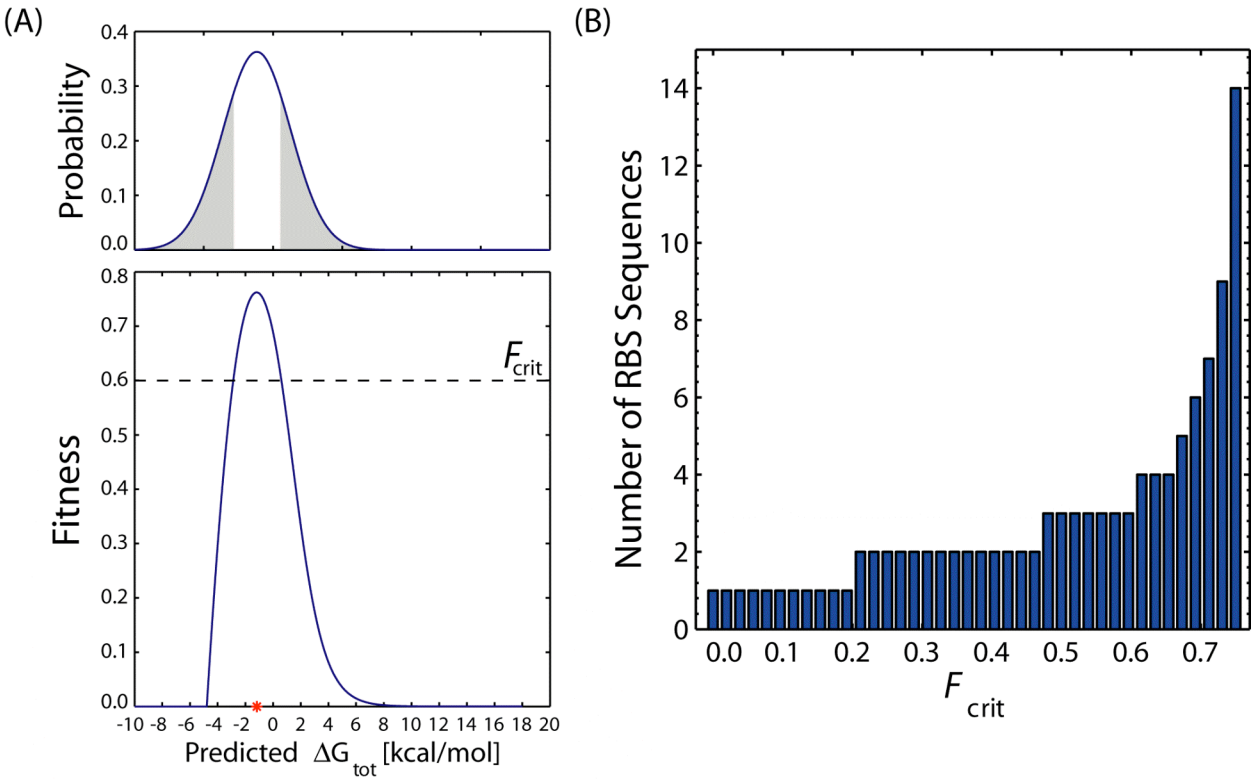

Figure S11: Calculation of the design probabilities for successfully connecting the AND gate genetic circuit to the P_{BAD} sensor. (A) The optimal G_{tot} (red star) and probability of success is calculated according to a fitness curve $F(G_{tot})$ and a fitness threshold F_{crit}. Equation S14 integrates the design method's error distribution (top) between the lower and upper bound defined by F_{crit}. (B) The number of synthetic ribosome binding site sequences that are needed to ensure a 90% confidence level will increase with a higher fitness threshold.

8. Analysis of the Thermodynamic Model

In **Figure S5** we show the results of combining all 81 synthetic ribosome binding site sequences into a single dataset in order to compare the predicted G_{tot} values to their measured protein expression levels. According to the statistical thermodynamic theory, we expect a linear relationship between the log protein fluorescence and the predicted G_{tot}. We obtain a linear relationship ($R^2 = 0.71$) with a slope $\beta = 0.45 \pm 0.05$.

Importantly, the five free energy terms must be summed together (G_{tot}) in order to correctly predict the translation initiation rate of a protein coding sequence. By themselves, each free energy term is a poor predictor of the translation initiation rate ($G_{mRNA:RNA}$: $R^2 = 0.32$; G_{mRNA}: $R^2 = 0.19$; $G_{spacing}$: $R^2 = 0.08$; $G_{standby}$: $R^2 = 0.13$) (**Figures S6**). In addition, the inclusion of each free energy term improves the accuracy of the predictions; leaving one out of the calculation results in a poorer correlation ($G_{mRNA:RNA}$: $R^2 = 0.37$; G_{mRNA}: $R^2 = 0.40$; $G_{spacing}$: $R^2 = 0.64$; $G_{standby}$: $R^2 = 0.66$) (**Figure S7**).

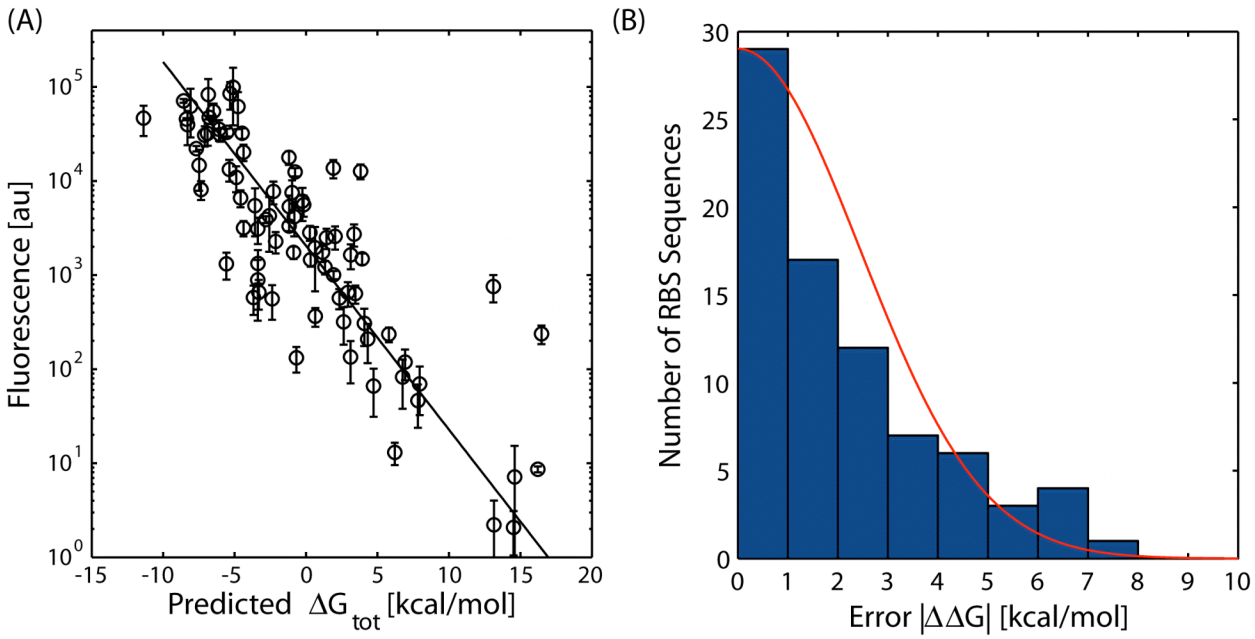

Figure S5: (A) The predicted G_{tot} values for 81 synthetic ribosome binding sites are compared to their experimentally measured protein expression levels (black circles). According to the theory, we expect a linear relationship between the logarithm of the protein expression level and the predicted G_{tot}. (B) A histogram shows that thermodynamic model is capable of predicting the protein expression level of most of the ribosome binding site sequences. Error bars are calculated as the standard deviation of at least five measurements on two different days.

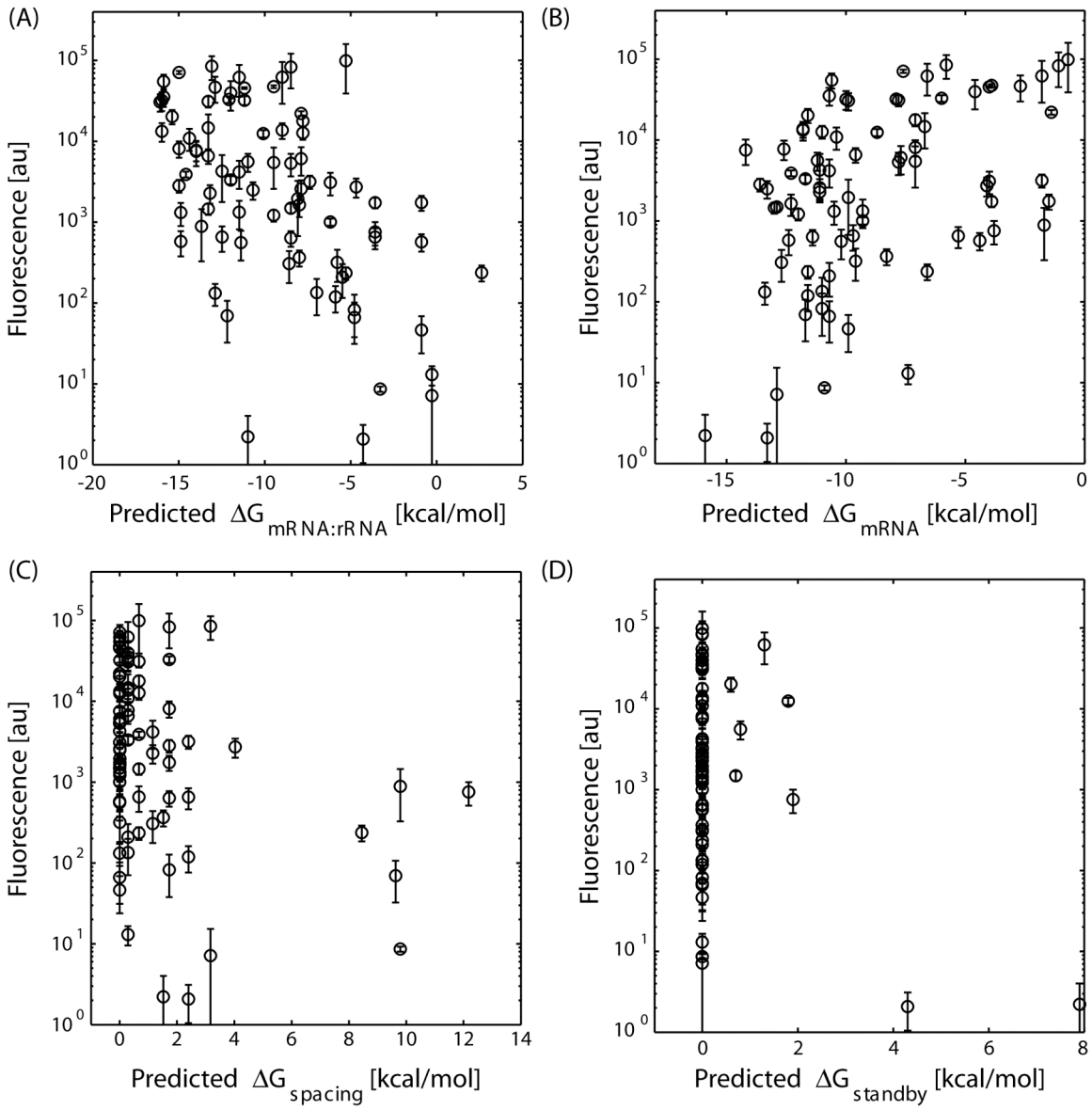

Figure S6: The predicted (A) $G_{mRNA:rRNA}$, (B) G_{mRNA}, (C) $G_{spacing}$, and (D) $G_{standby}$ free energy terms for 81 synthetic ribosome binding site sequences are compared to the experimentally measured protein expression levels. Error bars are calculated as the standard deviation of at least five measurements on two different days.

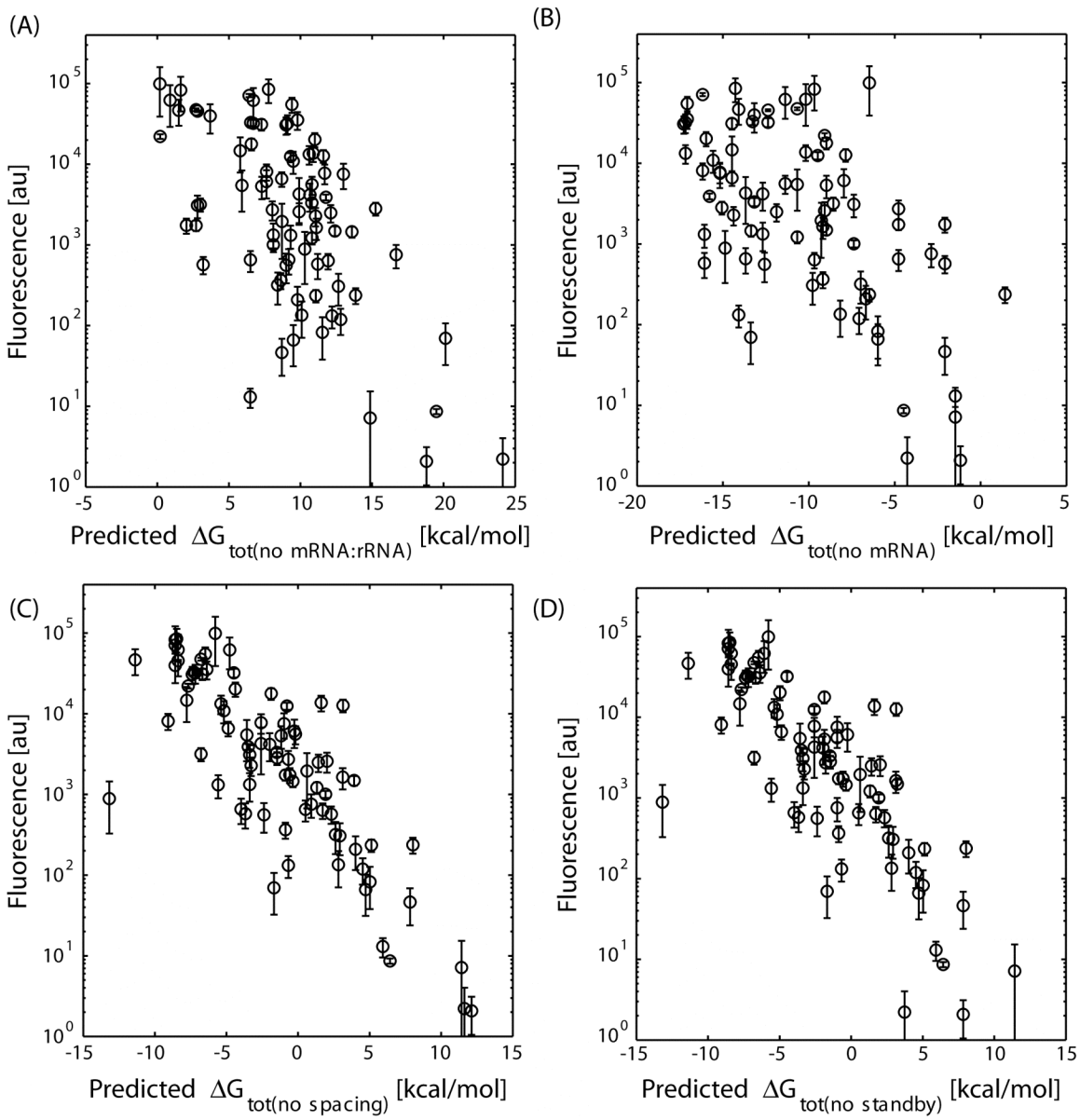

Figure S7: The thermodynamic model is used to calculate the G_{tot} for 81 synthetic ribosome binding sites, while excluding either the (A) $G_{mRNA:rRNA}$, (B) G_{mRNA}, (C) $G_{spacing}$, or (D) $G_{standby}$ free energy term from the model. The resulting prediction is then compared to the experimentally measured protein expression levels as before. Error bars are calculated as the standard deviation of at least five measurements on two different days.

9. Characterization of Ensembles of Synthetic Ribosome Binding Sites

The design method provides a connection between the sequence of a synthetic ribosome binding site, the strengths of the participating molecular interactions, and its translation initiation rate. During evolution, nucleotide conservation appears as a result of a physiological function that improves fitness and selection. We can use the design method to simulate the interplay between user-defined constraints on ribosome binding site function and the nucleotide conservation [16] that appears as a result, thus simulating the effects of evolution.

Using the design method, we create ensembles of synthetic RBS sequences where the target translation initiation rate was systematically varied across a 100,000-fold range; sequences in the same ensemble are predicted to have similar translation initiation rates. To do this, we specified the mRFP1 protein coding sequence and target ΔG_{tot} into the design method and generated a synthetic ribosome binding site sequence. We then repeated this process 1000 times, using the same target ΔG_{tot}, to create an ensemble of 1000 synthetic ribosome binding site sequences with similar predicted translation initiation rates. All ribosome binding site sequences were 35 nucleotides long. The target ΔG_{tot} was then varied from −9 kcal/mol to 20 kcal/mol, in steps of 1 kcal/mol, creating thirty ensembles of synthetic ribosome binding sites. We then calculate the Boltzmann sequence entropy for each ensemble according to:

$$S(\Delta G_{tot})/k_b = -\sum_{i=1}^{35} \sum_{j=\{A,G,C,U\}} p_i(n_j) \log(p_i(n_j)), \quad \text{(S16)}$$

where $p_i(n_j)$ is the probability of finding the j^{th} nucleotide at the i^{th} position in the ensemble of samples with the same target ΔG_{tot}. The nucleotide probabilities $p_i(n_j)$ are then visualized using sequence logos [17].

We then analyzed how the user-defined target translation rate affected the sequence entropy and the nucleotide conservation. At a low target translation initiation rate, there are a large number of RBS sequences (high entropy) that satisfy the functional constraints with little nucleotide conservation, indicating a highly degenerate sequence space (**Supplementary Figure S10A**). The maximum entropy of a sequence is $\log(4^{35}) = 48.52$. By comparing this maximum entropy to the entropy at the lowest target translation initiation rate, we observe that the constraints placed on synthetic RBS sequences eliminate about 60% of all possible sequences. These constraints include the absence of a large activation barrier in the unfolding of the 16S rRNA binding site and the absence of any long-range nucleotide base pairing; these constraints improve the thermodynamic model's accuracy. As the target translation initiation rate is increased, the optimization algorithm places an even tighter constraint on RBS function. As a result, the number of RBS sequences that satisfy a higher target translation initiation rate decreases (lower entropy) with greater nucleotide similarities between sequences. These nucleotide similarities appear at the locations that have the most effect on RBS function (**Supplementary Figure S10B**). Most notably, nucleotide conservation appears at the 16S rRNA binding site and at a short upstream region that forms non-occluding secondary structures. These non-occluding secondary structures prevent the formation of other secondary structures that compete with ribosome binding. Thus, without specifically requiring it, nucleotide conservation naturally emerges as a result of constraints on RBS function.

Next, we analyze how the strengths of the individual molecular interactions are affected by changing the target translation initiation rate. For each ribosome binding site sequence, we compare the values of their free energy terms with one another, taking two-dimensional snapshots of the five-dimensional free energy space. For each target ΔG_{tot}, four snapshots are shown: the ΔG_{mRNA} vs. $\Delta G_{mRNA:rRNA}$, the $\Delta G_{spacing}$ vs. $\Delta G_{mRNA:rRNA}$, the $\Delta G_{spacing}$ vs. $\Delta G_{standby}$, and the ΔG_{mRNA} vs. $\Delta G_{standby}$ (**Supplementary Figure S10C**). At the lowest target ΔG_{tot} (highest translation initiation rate), most ribosome binding site sequences occupy a small region of the free energy space, as they are constrained to minimize the magnitudes of the $\Delta G_{spacing}$ and $\Delta G_{standby}$ penalties,

prevent occluding secondary structures (maximize G_{mRNA}), and increase the 16S rRNA binding site affinity (minimize $G_{mRNA:rRNA}$). As the target G_{tot} increases (lower translation initiation rates), the generated ribosome binding site sequences are less constrained and will sample more of the available free energy space. At the highest target G_{tot} shown, the degeneracy in the strengths of the molecular interactions is quite large, indicating that the ribosome binding site sequences share little similarity in their molecular interactions. Thus, the design method allows us to connect the emergence of nucleotide conservation in RBS sequences to the functional constraints applied at the molecular level.

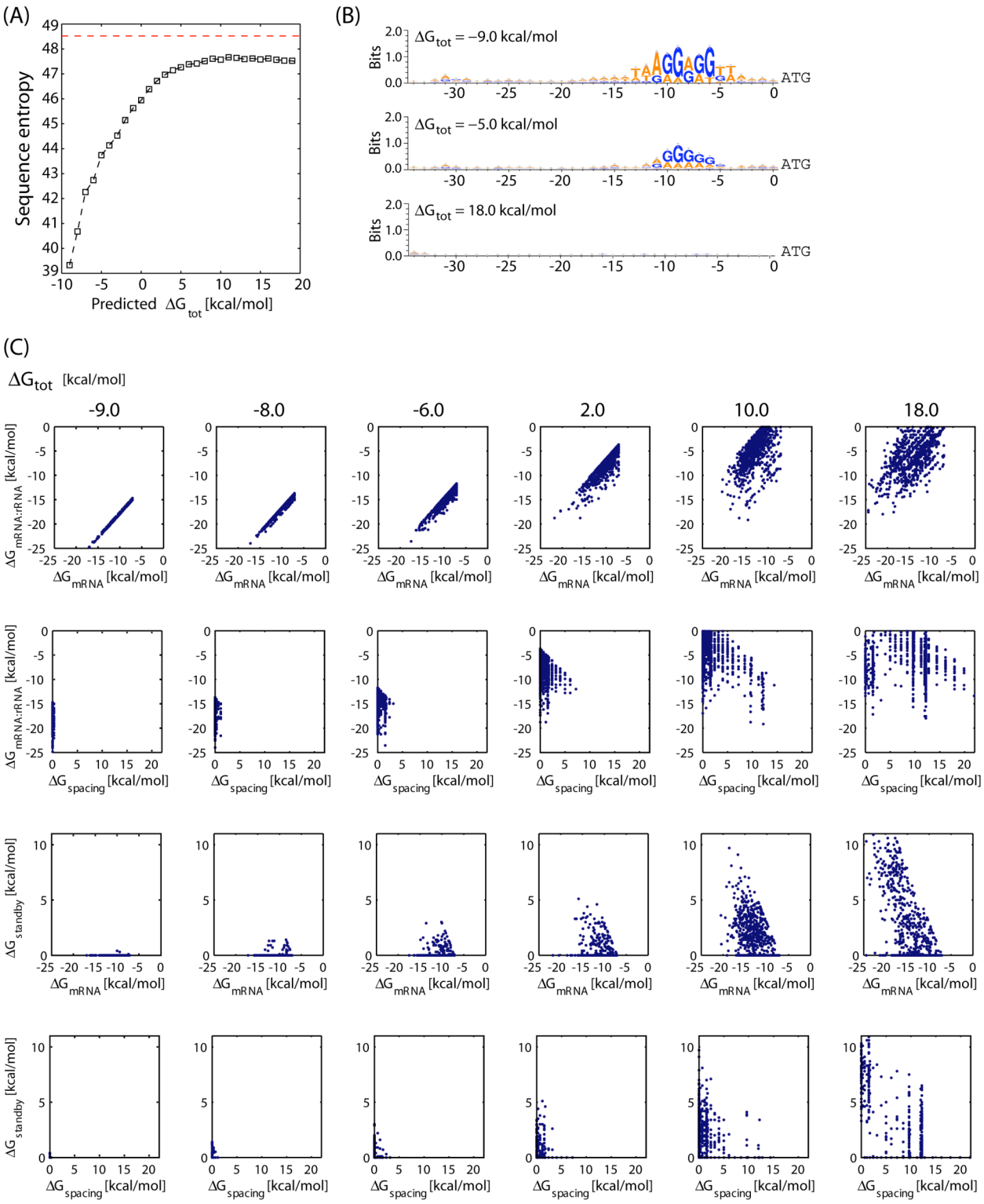

Figure S10: Ensemble analysis of synthetic ribosome binding sites. (A) The sequence entropy of an ensemble of synthetic RBS sequences decreases with higher translation initiation rates (lower target G_{tot}) (squares, dashed line). The maximum entropy for a 35 nucleotide RBS is $\log(4^{35}) = 48.52$ (red dashed line). (B) Nucleotide conservation emerges as the target translation rate increases, most notably at the 16S rRNA binding site and a short upstream region that forms

non-occluding secondary structures. (C) Each box compares the values of two different free energy terms of each sequence in an ensemble of synthetic ribosome binding sites, designed with the shown target G_{tot}. At high target translation rates, the constraints cause the optimization algorithm to generate sequences with free energies that remain clustered in a small portion of the available free energy space. As the target G_{tot} increases, the design method's optimization algorithm is less constrained and finds sequence solutions that sample more of the available space.

10. A Table of All Ribosome Binding Site Sequences

The ribosome binding site sequences in this study are located in an Excel worksheet (**Supplementary Table I**). The DNA (RNA) sequence begins at the transcriptional start site (+1) and continues into the middle of the protein coding sequence. All thymine nucleotides (T) are converted to uracil (U) when inputted into the thermodynamic model for the conversion from DNA to RNA. The predicted G_{tot} [kcal/mol], average protein fluorescence [au], and the standard deviation of fluorescence [au] are shown for each sequence.

11. A Summary of Predictions for Common Ribosome Binding Sites

RBS BioBrick[a] Name	RBS Sequence	Protein	First 35 nt of protein coding sequence	Predicted TIR [au]
BBa_J61101	TTCTAGAGAAAGACAGGACCCACTAGT	mRFP1	ATGGCGAGCTCTGAAGACGTTATCAAAGAGTTCAT	356
BBa_J61104	TTCTAGAGAAAGAAGGGACAGACTAGT	mRFP1	ATGGCGAGCTCTGAAGACGTTATCAAAGAGTTCAT	703
BBa_J61107	TTCTAGAGAAAGAAGAGACTCACTAGT	mRFP1	ATGGCGAGCTCTGAAGACGTTATCAAAGAGTTCAT	86
BBa_J61115	TTCTAGAGAAAGAAGGGATACACTAGT	mRFP1	ATGGCGAGCTCTGAAGACGTTATCAAAGAGTTCAT	735
BBa_J61120	TTCTAGAGAAAGACGCGAGAAACTAGT	mRFP1	ATGGCGAGCTCTGAAGACGTTATCAAAGAGTTCAT	228
BBa_J61130	TTCTAGAGAAAGAAACGACATACTAGT	mRFP1	ATGGCGAGCTCTGAAGACGTTATCAAAGAGTTCAT	62
BBa_B0030	TTCTAGAattaaagaggagaaattaagc	mRFP1	ATGGCGAGCTCTGAAGACGTTATCAAAGAGTTCAT	4100
BBa_B0031	TTCTAGAtcacacaggaaaccggttcg	mRFP1	ATGGCGAGCTCTGAAGACGTTATCAAAGAGTTCAT	495
BBa_B0032	TTCTAGAtcacacaggaaaggcctcg	mRFP1	ATGGCGAGCTCTGAAGACGTTATCAAAGAGTTCAT	146
BBa_B0033	TTCTAGAtcacacaggacggccgg	mRFP1	ATGGCGAGCTCTGAAGACGTTATCAAAGAGTTCAT	584

[a] the catalog of biobricks can be found at http://partsregistry.org/Main_Page

12. Supplementary References

1. Znosko, B.M., Silvestri, S.B., Volkman, H., Boswell, B. & Serra, M.J. Thermodynamic parameters for an expanded nearest-neighbor model for the formation of RNA duplexes with single nucleotide bulges. *Biochemistry* **41**, 10406-10417 (2002).
2. Xia, T. et al. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **37**, 14719-14735 (1998).
3. Mathews, D.H., Sabina, J., Zuker, M. & Turner, D.H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288**, 911-940 (1999).
4. Dirks, R.M., Bois, J.S., Schaeffer, J.M., Winfree, E. & Pierce, N.A. Thermodynamic Analysis of Interacting Nucleic Acid Strands. *SIAM Review* **49**, 65-88 (2007).

5. Kierzek, R., Burkard, M.E. & Turner, D.H. Thermodynamics of single mismatches in RNA duplexes. *Biochemistry* **38**, 14214-14223 (1999).
6. Miller, S., Jones, L.E., Giovannitti, K., Piper, D. & Serra, M.J. Thermodynamic analysis of 5' and 3' single- and 3' double-nucleotide overhangs neighboring wobble terminal base pairs. *Nucleic Acids Res* **36**, 5652-5659 (2008).
7. Christiansen, M.E. & Znosko, B.M. Thermodynamic characterization of the complete set of sequence symmetric tandem mismatches in RNA and an improved model for predicting the free energy contribution of sequence asymmetric tandem mismatches. *Biochemistry* **47**, 4329-4336 (2008).
8. Blose, J.M. et al. Non-nearest-neighbor dependence of the stability for RNA bulge loops based on the complete set of group I single-nucleotide bulge loops. *Biochemistry* **46**, 15123-15135 (2007).
9. Badhwar, J., Karri, S., Cass, C.K., Wunderlich, E.L. & Znosko, B.M. Thermodynamic characterization of RNA duplexes containing naturally occurring 1 x 2 nucleotide internal loops. *Biochemistry* **46**, 14715-14724 (2007).
10. O'Toole, A.S., Miller, S., Haines, N., Zink, M.C. & Serra, M.J. Comprehensive thermodynamic analysis of 3' double-nucleotide overhangs neighboring Watson-Crick terminal base pairs. *Nucleic Acids Res* **34**, 3338-3344 (2006).
11. Vecenie, C.J., Morrow, C.V., Zyra, A. & Serra, M.J. Sequence dependence of the stability of RNA hairpin molecules with six nucleotide loops. *Biochemistry* **45**, 1400-1407 (2006).
12. Andronescu, M., Condon, A., Hoos, H.H., Mathews, D.H. & Murphy, K.P. Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics* **23**, i19-28 (2007).
13. Bevilacqua, P.C. & Blose, J.M. Structures, kinetics, thermodynamics, and biological functions of RNA hairpins. *Annu Rev Phys Chem* **59**, 79-103 (2008).
14. Chen, H., Bjerknes, M., Kumar, R. & Jay, E. Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of Escherichia coli mRNAs. *Nucleic Acids Res* **22**, 4953-4957 (1994).
15. Anderson, J.C., Voigt, C.A. & Arkin, A.P. Environmental signal integration by a modular AND gate. *Mol Syst Biol* **3**, 133 (2007).
16. Shultzaberger, R.K., Bucheimer, R.E., Rudd, K.E. & Schneider, T.D. Anatomy of Escherichia coli ribosome binding sites. *J Mol Biol* **313**, 215-228 (2001).
17. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. WebLogo: a sequence logo generator. *Genome Res* **14**, 1188-1190 (2004).