The Influence of Decoys on the Noise and Dynamics of Gene Expression

Anat Burger*,[1] Aleksandra M. Walczak,[2] and Peter G. Wolynes[3]

[1] *Center for Theoretical Biological Physics, University of California San Diego, La Jolla, CA.*
[2] *CNRS and Laboratoire de Physique Théorique de l'École Normale Supérieure, Paris, France.*
[3] *Center for Theoretical Biological Physics, Rice University, Houston, TX.*

(Dated: October 8, 2012)

Many transcription factors bind to DNA with a remarkable lack of specificity, so that regulatory binding sites compete with an enormous number of non-regulatory 'decoy' sites. For an auto-regulated gene, we show decoy sites decrease noise in the number of unbound proteins to a Poisson limit that results from binding and unbinding. This noise buffering is optimized for a given protein concentration when decoys have a 1/2 probability of being occupied. Decoys linearly increase the time to approach steady state and exponentially increase the time to switch epigenetically between bistable states.

PACS numbers:

INTRODUCTION

A transcription factor must bind to a specific site in the genome to regulate the expression of a gene. This process does not occur in isolation. Instead, actual regulatory target sequences must be distinguished from an entire genome of alternative possible binding sites. In prokaryotes, the typical transcription factor binding motif is sufficiently specific that a regulatory target can be distinguished from decoys by its binding free energy alone as a roughly unique location in the genome [1]. Although eukaryotic genomes are much longer, the binding specificity of some eukaryotic transcription factor binding motifs can be so low that up to millions of consensus sequence binding sites can be expected by pure chance [2]. Recent experiments that measure *in vivo* binding occupancy for large numbers of transcription factors across various cell types and developmental contexts [3] make it now possible to investigate the nature of transcription factor-DNA binding.

Although usually only a subset of the predicted binding sites for a transcription factor are found to be occupied *in vivo* [4], some transcription factors have been found to bind to tens of thousands of sites, such as the muscle differentiation factor MyoD [5]. Certain developmental master regulators may have widespread regulatory binding because they encode positional information within an organism, and are thus implicated in the regulation of a majority of genes in order to modulate subtle differences between the cells of a particular tissue [6, 7]. Estimates for the mean fraction of occupied sites in regulatory regions that are functional (as determined by evolutionary conservation or gene expression assays) range between 10-40% [6, 8]. Alternative roles for transcription factors bound to DNA *in addition to* the canonical function of

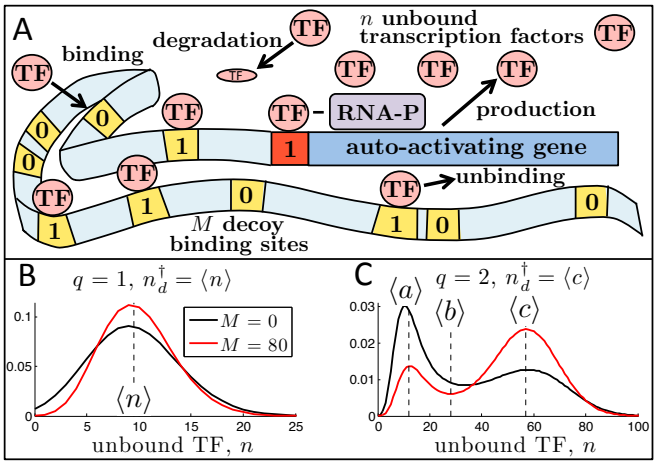

FIG. 1: **A.** Model of a generic auto-activating gene where transcription factors bind to a regulatory promoter site (*red*) as well as M identical, non-regulatory decoy binding sites (*yellow*) **B.** Since they protect bound proteins from degradation, decoy binding sites do not alter the steady state mean unbound copy number of a unimodal probability distribution, $\langle n \rangle$, yet they decrease the variance σ_n^2. **C.** Similarly, the deterministic fixed points of a bistable system, $\{\langle a \rangle, \langle b \rangle, \langle c \rangle\}$ do not change, but when decoys are added the relative stability of the expression states, LOW ($n < \langle b \rangle$) and HIGH ($n > \langle b \rangle$), is altered.

modification of transcriptional initiation [9] have been proposed, for example chromatin remodeling [6] or DNA repair [10]. There may be other advantages to widespread transcription factor binding, such as to specify regulatory regions [2] or to facilitate in the evolution of regulatory elements [11]. Decoy sites have also been identified in repetitive non-coding regions [12]. Mutations in these regions have been implicated in several diseases, suggesting that the non-regulatory binding of transcription factors to DNA could serve some currently unknown function, a question that is being explored in synthetically engineered systems [13].

Several studies have suggested that an additional con-

*Current affiliation: Department of Physics, FAS Center for Systems Biology, Harvard University, Cambridge, MA.

sequence of non-functional binding may be in maintaining a large abundance of transcription factors in a cell [14, 15] and buffering noise in gene expression [9, 16, 17]. In this paper we provide an analytical theory of how the noise characteristics and approach to steady state of the system are altered by non-regulatory binding (decoy sites) that confer stability to the system.

Previously we have shown [18] that when DNA-bound transcription factors are protected from degradation, which may be the case for several eukaryotic transcription factors including MyoD [19], the mean steady state concentration of *unbound* transcription factors, $\langle n \rangle$, does not change as decoys are added. Instead, the *total* number of transcription factors, N, adjusts to satisfy the binding to decoys and thus decoys do not change the deterministic behavior of the system. Here we exploit timescale separation to isolate and compare the properties of two contributions to the noise in transcription factor expression. The first is the intrinsic noise from production and degradation of transcription factors which is buffered by decoys. The second source of noise results from binding and unbinding of transcription factors to DNA, which becomes Poissonian for large numbers of decoy binding sites. We show that the optimum noise buffering decoys for a given concentration of transcription factors have a binding affinity that ensures they have an equal probability of being occupied and not occupied. Additionally, decoy binding affinity can alter the probability of occupancy of expression states in systems that exhibit multistability. For simplicity, we choose to study a ubiquitous network motif of an auto-regulated gene, but our results have a wide-ranging applicability to many biological systems.

THE MODEL

To elucidate the general effect of decoys on gene expression we model an auto-activated gene surrounded by a collection of M identical decoy binding sites that do not themselves directly regulate transcription but do protect bound proteins from degradation (Fig. 1). To describe this system we consider a master equation (Eqn. 1) for the time evolution of the joint probability distribution of the promoter occupancy, $i \in \{\text{unbound } (0), \text{bound } (1)\}$, the number of occupied decoys, m, and the number of unbound proteins, n:

$$\begin{aligned}
\partial_t p_{i,m,n} \;=\; & \Big[g_i p_{i,m,n-1} - g_i p_{i,m,n} \Big] \\
& + \Big[k(n+1) p_{i,m,n+1} - k n p_{i,m,n} \Big] \\
& + (-1)^{1-i} H_p(n+qi) p_{0,m,n+qi} \\
& + (-1)^{i} f_p p_{1,m,n-q(1-i)} \\
& + \Big[H_d(n+q) \Big(M - (m-1) \Big) p_{i,m-1,n+q} \\
& - H_d(n) \Big(M - m \Big) p_{i,m,n} \Big] \\
& + f_d \Big[\Big(m+1 \Big) p_{i,m+1,n-q} - m p_{i,m,n} \Big] . \quad (1)
\end{aligned}$$

The reactions represented in the master equation include protein production $n \xrightarrow{g_i} n+1$, degradation $n \xrightarrow{kn} n-1$, promoter binding, $(i,n) \xrightarrow{H_p(n)(1-i)} (i+1, n-q)$, promoter unbinding, $(i,n) \xrightarrow{f_p i} (i-1, n+q)$, decoy binding, $(m,n) \xrightarrow{H_d(n)(M-m)} (m+1, n-q)$, and decoy unbinding, $(m,n) \xrightarrow{f_d m} (m-1, n+q)$. The binding process encoded in the function H is described for $x \in \{p, d\}$ as $H_x(n) = h_x n$ for binding of monomers ($q = 1$) and $H_x(n) = \frac{1}{2} h_x n(n-1)$ for binding of dimers ($q = 2$). We define a site equilibrium constant $n_x^\dagger = f_x/h_x$ for $q = 1$ and $n_x^\dagger = \sqrt{2 f_x / h_x}$ for $q = 2$ that corresponds to a binding free energy E_x such that $n_x^\dagger = e^{\beta E_x}$, where $\beta = (k_B T)^{-1}$.

We solve this master equation numerically by matrix diagonalization to study properties of the steady state probability distribution over unbound copy numbers, $p_n = \sum_{i,m} p_{i,m,n} (t = \infty)$. To illustrate the invariant scalings it is convenient to introduce a factor S so that we write the production and promoter binding terms as $g_i = \widehat{g}_i S$ and $n_p^\dagger = \widehat{n}_p^\dagger S$. This results in $\langle n \rangle = \sum_n n p_n \approx \widehat{\langle n \rangle} S$. The equilibrium probability that a site is occupied is thus a Hill function,

$$\theta_x(\langle n \rangle) = \frac{\langle n \rangle^q}{(n_x^\dagger)^q + \langle n \rangle^q} \quad (2)$$

which can also be written in terms of energy, such that $\theta_x = 1/\left(1 + \exp\left[\beta q \Delta E\right]\right)$, where $\Delta E = E_x - \mu$ and $\mu = k_B T \ln \langle n \rangle$.

Dimensional reduction. We focus on the limiting case where binding and unbinding are both much faster than production and degradation; the case of so called "adiabatic" genes. We take advantage of this separation in timescales to treat separately the fast fluctuations in unbound copy number–due to binding and unbinding events–from the slow fluctuations in unbound copy number–due to production and degradation events. In this limit we are able to collapse the master equation (Eqn. 1) to a single dimension in terms of the slowly changing variable of the system, the *total* number of transcription factors, $N \equiv n + qi + qm$:

$$\partial_t p_N = \Big[G(N-1)p_{N-1} - G(N)p_N\Big] + \Big[K(N+1)p_{N+1} - K(N)p_N\Big], \quad (3)$$

which we write in terms of effective rates for the production, $G(N)$, and degradation, $K(N)$, of transcription factors. These rates are defined self-consistently as functions of the slowly varying component of the unbound transcription factors, $\bar{n}$, which depends on the total number of transcription factors, N:

$$N = \bar{n}(N) + qM\theta_d[\bar{n}(N)], \quad (4)$$

In Eq. 4 we neglect the term corresponding to binding of transcription factor proteins to the promoter, since we will be mainly interested in the limit of many decoy sites where the contribution of this term is small compared to the decoy binding term.

The effective production rate is a function of the probability that the promoter is occupied:

$$G(N) = g_0\Big(1 - \theta_p[\bar{n}(N)]\Big) + g_1\theta_p[\bar{n}(N)] \quad (5)$$

The effective degradation rate is proportional to the net unbound copy number, which excludes the mean number of transcription factors bound to the promoter:

$$K(N) = k\Big(\bar{n}(N) - q\theta_p[\bar{n}(N)]\Big). \quad (6)$$

NUMERICAL RESULTS

To gain intuition we first numerically solve the master equations for two cases that are known to have qualitatively different dynamical and noise properties without decoys: monomer ($q = 1$) and dimer ($q = 2$) binding (see caption of Fig 2 for details). We compare the numerical solutions for the full and reduced model in Fig. 4. Dimer binding allows for bistability and switching between the two attractors, whereas in the adiabatic limit monomer binding yields a unimodal distribution easily characterized by simple measures such as the Fano factor for noise ($\sigma_n^2/\langle n\rangle$) and the mean relaxation time to steady state. In Fig 2 we see that adding decoys with a fixed binding energy (we use decoys that are half bound at steady state [29]) quantitatively affects the gene expression properties. However when the number of decoys is rescaled by the mean number of unbound proteins, the results for different choices of S collapse onto a common plot (see Fig 2 insets) indicating general principles that we explore below.

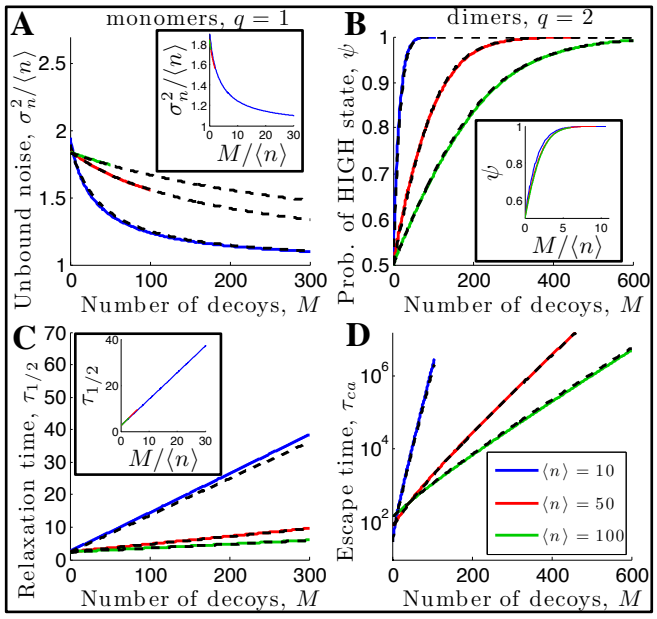

FIG. 2: Comparison of numerical (*solid curves*) and analytical (*dashed curves*) results for gene expression properties as decoys are added for systems with varying mean unbound numbers of protein copies, $\langle n\rangle$. **A.** The Fano factor; **B.** The probability for the bistable system to be in the HIGH protein expression state, ψ; **C.** time for the mean total copy number to reach half the steady state value; **D.** epigenetic escape time. Numerical results in **A** are calculated by projecting the solutions of the $3D$ master equation for $p_n = \sum_{i,m} p_{i,m,n}$, whereas the $1D$ master equation for p_N is accurate for the results plotted in **B, C, D** (see Fig. 4 for details). Analytical calculations follow from Eqns. 18, 21, 22, and 23 using numerical calculations for a gene without decoys. *Parameters:* $g_1 = 100S$, $g_0 = 8S$, $k = 1$, $n_P^\dagger = 53.2S$ for $q = 1$ which gives $\langle n\rangle = 50S$. For $q = 2$, $\psi_0 = 0.5$ is fixed such that $n_P^\dagger = 10.3$ for $S = .2$, $n_P^\dagger = 21.0$ for $S = 1$, and $n_P^\dagger = 106.8$ for $S = 2$.

We plot the dependence of the noise and dynamical properties of the system on the binding free energy of decoys E_d in Fig 3. In prokaryotic genomes, there is typically a free energy penalty of 1 to $2k_BT$ per mismatch with respect to the consensus binding motif. When there are 4 to 5 mismatches the binding becomes characteristic of background DNA [1]. Since most decoys will have a weaker binding affinity than the promoter, we concentrate on discussing the large M, large $n_d^\dagger$ limit [15].

Noise Buffering. The steady state unbound Fano factor, $\sigma_n^2/\langle n\rangle$, plotted in Fig. 2A approaches Poisson noise as decoys are added, such that $\sigma_n^2 \xrightarrow{M\to\infty} \langle n\rangle$. In the limit of large numbers of decoys the slow fluctuations in unbound copy number resulting from production and degradation events are dominated by an effective birth-death process in which a relatively small number of particles bind and unbind to a large reservoir of sites. We see that systems having smaller mean numbers of proteins approach the Poisson limit for smaller values of M (com-

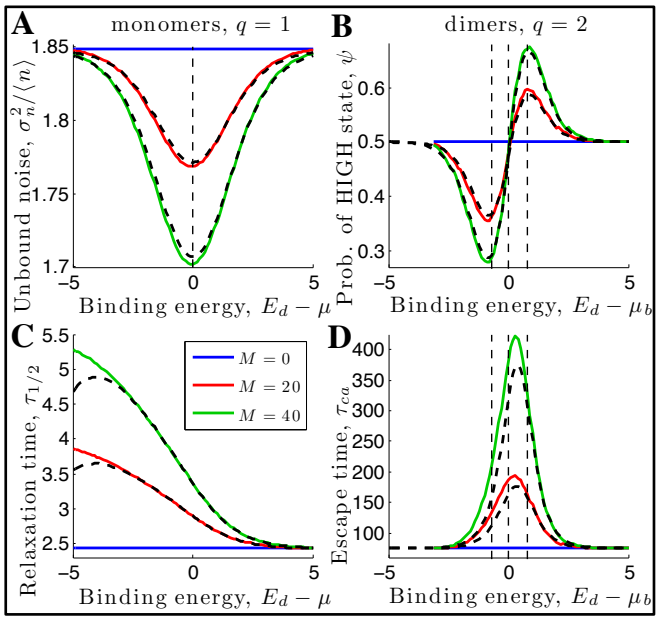

FIG. 3: Comparison of numerical (*solid curves*) and analytical (*dashed curves*) for the same properties as in Fig. 2 as a function of the decoy binding energy E_d, for fixed numbers decoys, M. The vertical dashed lines indicate the energies that correspond to the fixed points of the system. Parameters are the same as in Fig. 2 for $\langle n\rangle = 50$.

pare blue and green curves in Fig. 2A) than those with larger mean protein numbers. Figure 3A shows that noise buffering is optimized for a particular value of the decoy binding energy, $E_d^* = \mu$. This corresponds to the case where decoys are half bound at steady state ($n_d^{\dagger *} = \langle n\rangle$). Intuitively, the potential to buffer noise is maximized at $E_d^* = \mu$ since binding and unbinding events are most probable when sites are on average half-occupied.

Approach to Steady State. Although the mean steady state *unbound* protein copy number, $\langle n\rangle$, remains constant, adding decoys increases the mean steady state *total* protein number, $\langle N\rangle = \langle n\rangle + M\theta_d(\langle n\rangle)$. The relaxation time, $\tau_{1/2}$, (the time to reach $\langle N(\tau_{1/2})\rangle = \langle N\rangle/2$, from an initial condition of $\langle N(0)\rangle = 0$) increases linearly as decoys are added (Fig 2C) due to the time required to produce the proteins needed to satisfy binding equilibrium. Strongly binding decoys ($E_d << \mu$) increase $\tau_{1/2}$ the most because more proteins must be created (Fig 3C).

Epigenetic Escape. In a bistable system where proteins bind as dimers, the addition of decoys does not alter the three deterministic fixed points corresponding to the stable low expression, unstable intermediate expression, and stable high expression levels, $n = \{\langle a\rangle, \langle b\rangle, \langle c\rangle\}$. However, decoys are able to influence the ability of the system to stochastically transition between the stable global phenotypic states which we call the LOW and HIGH expression states (See Fig 1C). The binding affinity of the decoys determines the change in the likelihood of observing the different expression states. In Fig. 2B we see that decoys with a binding energy $E_d = \mu_c \equiv k_BT\ln\langle c\rangle$ increase the probability to be in the HIGH protein copy expression state by preferentially decreasing fluctuations in the protein buffer in the vicinity of $n = \langle c\rangle$, such that $\psi \xrightarrow{M\to\infty} 1$ where $\psi = \sum_{n>\langle b\rangle} p_n$. On the other hand, decoys with a binding energy $E_d = \mu_a \equiv k_BT\ln\langle a\rangle$ will act to stabilize the LOW protein copy number expression state, such that $\psi \to 0$. We see that the epigenetic escape times, defined as the mean first passage times between the two steady states, $\tau_{ac} : n = \langle a\rangle \to n = \langle c\rangle$ and $\tau_{ca} : n = \langle c\rangle \to n = \langle a\rangle$, increase exponentially as decoys are added (Fig. 2D). The variation of ψ with decoy binding energy (Fig 3B) shows that decoys with binding energy $E_d = \mu_b$ stabilize neither state, however, they significantly increase the epigenetic escape rate by effectively stabilizing the transition state (Fig 3D).

ANALYTICAL RESULTS

Noise Buffering. To understand the numerical observations in Figs. 2 and 3 we note the variance in the unbound protein concentration depends on both fast and slow fluctuations through the law of total variance, $\sigma_n^2 = \sigma_{n,slow}^2 + \sigma_{n,fast}^2$, where the slow fluctuations are due to production and degradation events and the fast fluctuations are due to binding and unbinding events.

The slow contribution to the variance can be obtained by approximating the master equation for p_N (Eqn 3) by a Fokker-Planck equation:

$$\frac{\partial}{\partial t}p_N = -\frac{\partial}{\partial N}\left[v(N) - \frac{1}{2}\frac{\partial}{\partial N}D(N)\right]p_N, \qquad (7)$$

with the drift, $v(N) \equiv G[\bar{n}(N)] - K[\bar{n}(N)]$, and diffusion, $D(N) \equiv G[\bar{n}(N)] + K[\bar{n}(N)]$. The steady state probability distribution of Eq. 7 is given by:

$$p(N) = \frac{\mathcal{N}}{D(N)}\exp\left[\int_0^N dN' \frac{2v(N)}{D(N)}\right]. \qquad (8)$$

Within a Gaussian approximation around $N = \langle N\rangle$, Eqn. 8 yields the variance in the total protein copy number:

$$\sigma_N^2 = \left|\frac{D(N)}{-2\partial_N[v(N)]}\right|_{N=\langle N\rangle}. \qquad (9)$$

One can obtain the variance in the slowly varying component of the unbound protein copies, $\bar{n}$, by performing a change of variables on Equation 9 from N to $\bar{n}$. The drift

and diffusion functions evaluated for $\bar{n}$ are equivalent to that of a gene without decoys ($v_0(\bar{n})$ and $D_0(\bar{n})$). The derivative, $\mathcal{J}(\bar{n}) \equiv \partial N/\partial \bar{n}$, is calculated from Eqn. 4 [30]:

$$\mathcal{J}(\bar{n}) = \begin{cases} 1 + M\frac{n_d^\dagger}{(n_d^\dagger+\bar{n})^2}, & \text{for } q=1 \\ 1 + M\frac{4(n_d^\dagger)^2\bar{n}}{((n_d^\dagger)^2+\bar{n}^2)^2}, & \text{for } q=2 \end{cases} \quad (10)$$

After the change of variables,

$$\sigma^2_{n,slow} = \left|\frac{D_0(\bar{n})/\mathcal{J}(\bar{n})}{-2\partial_{\bar{n}}[v_0(\bar{n})]}\right|_{\bar{n}=\langle n\rangle} = \frac{\sigma_0^2}{\mathcal{J}(\langle n\rangle)}, \quad (11)$$

where σ_0^2 is the variance of the gene without decoys [31].

To calculate the fast contribution to the variance in the number of unbound protein copies due to binding and unbinding of monomers, $\sigma^2_{n,fast}$, we consider a master equation indexed over the number of unbound transcription factors, n, given a *constant* total number of transcription factors, N:

$$\begin{aligned} \frac{dp_{n|N}}{dt} &= f_d\Big[(N-n+1)p_{n-1|N} \\ &\quad -(N-n)p_{n|N}\Big] \\ &+ h_d\Big[(n+1)(M-N+n+1)p_{n+1|N} \\ &\quad -n(M-N+n)p_{n|N}\Big] \end{aligned} \quad (12)$$

We neglect binding and unbinding to the promoter because we are interested in the limit of large numbers of decoy sites, $M \to \infty$. The steady state probability distribution is found by recursion:

$$\begin{aligned} p_{n|N} &= p_{0|N}\prod_{\ell=0}^{n-1}\frac{f(N-\ell)}{h(\ell+1)(M-N+\ell+1)} \\ &= p_{0|N}(n^\dagger)^n\frac{N!}{n!(N-n)!}\frac{(M-N)!}{(M-N+n)!} \\ &\equiv \exp\left[\mathcal{F}(n)\right] \quad (13) \\ &\approx \exp\left(\mathcal{F}(\bar{n})\right)\exp\left[\frac{1}{2}(n-\bar{n})^2\frac{\partial^2\mathcal{F}}{\partial n^2}\Big|_{n=\bar{n}}\right]. \quad (14) \end{aligned}$$

In the last step we Gaussian expand $\mathcal{F}$ for large M, N, and n within a Stirling expansion. Setting $\partial/\partial n\left[\mathcal{F}(\bar{n})\right] = 0$ recovers the deterministic result for the mean number of unbound protein copy numbers, $\bar{n} \approx \sum_n np_{n|N}$ for $\bar{n} >> 0$, given in Eqn. 4. The variance in the number of unbound protein copy numbers is:

$$\begin{aligned} \sigma^2_{n|N} &= \left(\frac{\partial^2\mathcal{F}}{\partial n^2}\Big|_{n=\bar{n}}\right)^{-1} \\ &= \bar{n}\left[\frac{Mn_d^\dagger}{\left(n_d^\dagger+\bar{n}\right)^2+Mn_d^\dagger}\right] \\ &= \bar{n}\left[1-\frac{1}{\mathcal{J}(\bar{n})}\right]. \quad (15) \end{aligned}$$

The fast contribution to the unbound fluctuations is found by averaging over the probability distributions of the total copy number, p_N, which is the steady state solution of Eqn. 3 :

$$\begin{aligned} \sigma^2_{n,fast} &= \sum_N \bar{n}\left[\frac{Mn_d^\dagger}{\left(n_d^\dagger+\bar{n}\right)^2+Mn_d^\dagger}\right]p_N \\ &\approx \langle n\rangle\left[\frac{Mn_d^\dagger}{\left(n_d^\dagger+\langle n\rangle\right)^2+Mn_d^\dagger}\right] \quad (16) \\ &= \langle n\rangle\left[1-\frac{1}{\mathcal{J}(\langle n\rangle)}\right]. \quad (17) \end{aligned}$$

where we have approximated the average of the function by the function of the average, which is valid for $\left[\left(n_d^\dagger+\langle n\rangle\right)^2+Mn_d^\dagger\right] >> 1$.

Combining the slow (Eqn. 11) and fast (Eqn. 17) contributions to the variance yields

$$\sigma_n^2 \approx \left(\sigma_0^2-\langle n\rangle\right)\left[\frac{\left(n_d^\dagger+\langle n\rangle\right)^2}{\left(n_d^\dagger+\langle n\rangle\right)^2+Mn_d^\dagger}\right]+\langle n\rangle \quad (18)$$

This formula agrees well in the appropriate limits with numerical solutions of the full master equation, as shown in Figs. 2A and 3A, and also holds for a model that includes translational bursting (see Appendix B). From Eq. 18 in the large M limit, we obtain the observed Poisson noise, $\sigma_n^2 \to \langle n\rangle$. Noise reduction is proportional to the deviation from Poisson noise in a system without decoys. Decoys will decrease noise for $\sigma_0^2 > \langle n\rangle$ [32]. Eq. 18 is minimized for $n_d^{\dagger *} = \langle n\rangle$. Eq. 18 can be written as a function of $M/\langle n\rangle$ and ΔE which results in the data collapse shown in the inset of Fig. 2A.

To describe the noise buffering efficacy we quantify the number of decoys needed to reduce the super-Poissonian noise by a half, $M_{1/2}$. We find $M_{1/2} = 2\langle n\rangle(1+\cosh\Delta E)$ is independent of σ_0^2. For decoys with optimum buffering capacities ($\Delta E^* = 0$), $M_{1/2} = 4\langle n\rangle$ and $M_{1/2}$ asymptotically doubles for every binding energy increase of $k_BT\ln 2$ (or doubling of $n_d^\dagger$).

Approach to Steady State. The time to reach half of the mean steady state expression, $\tau_{1/2}$, starting from a mean of zero protein copies is found from the deterministic equation for the mean total copy number, $d_t\langle N(t)\rangle = v(N) = v_0\left[\bar{n}(N)\right]$, to be:

$$\tau_{1/2} = \int_0^{\langle N\rangle/2} dN \frac{1}{v_0\left[\bar{n}(N)\right]}. \tag{19}$$

Performing a change of variables from N to $\bar{n}$ yields:

$$\tau_{1/2} = \int_0^{\bar{n}(\langle N\rangle/2)} d\bar{n} \frac{\mathcal{J}(\bar{n})}{v_0(\bar{n})} \tag{20}$$

where the upper boundary is the mean unbound copy number $\bar{n}$ such that Eqn. 4 is evaluated for $N = \langle N\rangle/2$ for binding of monomers. In the limit of weak decoys, $E_d > \mu$, we find

$$\tau_{1/2} = \tau_{0,1/2} + M\Delta\tau_{1/2} \tag{21}$$

where $\Delta\tau_{1/2}$ is a correction due to adding the decoys, recovering the linear increase of $\tau_{1/2}$ with decoys seen in Fig. 2C. For very weak decoys, $E_d >> \mu$, (or $n_d^\dagger >> \langle n\rangle$), $\mathcal{J}(\bar{n}) \approx 1 + M/n_d^\dagger = \text{const}$. Hence $\Delta\tau_{1/2} \approx \tau_{1/2,0}/n_d^\dagger$ (see Appendix C for details).

Epigenetic Escape. Within the Fokker-Planck approximation the epigenetic escape time can be found by expanding the effective potential about the fixed points to second order. In the limit that the barrier height is sufficiently large one finds:

$$\tau_{ca} = \tau_{ca,0}\sqrt{\mathcal{J}_2(\langle c\rangle)\mathcal{J}_2(\langle b\rangle)}e^{M\zeta_{bc}}, \tag{22}$$

where $\tau_{ca,0}$ is the escape time without decoys and ζ_{bc} is a correction to the escape path action due to a single decoy (see Appendix D for details). An analogous expression holds for escape from $\langle a\rangle$ to $\langle c\rangle$. The escape times increase exponentially for large M as decoys are added.

Since the moel has been reduced to one dimension, the bimodal system obeys an effective detailed balance such that $\psi\tau_{ac} = (1-\psi)\tau_{ca}$, where ψ is the probability to be in the HIGH protein copy number expression state. Using the previous results for the escape times,

$$\psi = \frac{\psi_0\sqrt{\mathcal{J}(\langle a\rangle)/\mathcal{J}(\langle c\rangle)}e^{M\zeta_{ac}}}{1 + \psi_0\left(\sqrt{\mathcal{J}(\langle a\rangle)/\mathcal{J}(\langle c\rangle)}e^{M\zeta_{ac}} - 1\right)}. \tag{23}$$

When $n_d^\dagger \lessgtr \langle b\rangle$, $\zeta_{ac} \lessgtr 0$ such that $\psi \xrightarrow{M\to\infty} {}^{0}_{1}$. Thus the binding affinity of the decoys can determine which expression state is favored.

DISCUSSION

In summary, when there is a sufficient separation of timescales between slow protein production-degradation and fast binding-unbinding to the DNA, we have shown that decoys buffer gene expression noise. The fluctuations in binding and unbinding act as an effective birth-death process that imposes the Poisson limit on noise reduction. Noise buffering is optimized for decoys that are half-occupied at the appropriate protein concentration.

Not all gene regulatory systems function in the fully adiabatic limit explored here [21–23]. If binding and unbinding to decoys is much slower than the fluctuations in total copy number, decoys are unable to influence the steady state unbound protein expression. If binding and unbinding to the promoter become much slower than the fluctuations in total copy number, there are effectively two gene states with constant production rates. In this case the decoys have no impact on the steady state unbound protein expression.

ACKNOWLEDGMENTS

We thank Thierry Mora, Vincent Hakim, and Marc Santolini for helpful discussions, and support from the Center for Theoretical Biological Physics sponsored by the NSF (PHY-0822283), the D.R. Bullard-Welch Chair at Rice University, and the ICAM Junior Travel Grant.

APPENDIX A: VALIDATION OF DIMENSIONAL REDUCTION

In the limit that binding and unbinding are much faster than production and degradation we can reduce the master equation to a one dimensional master equation indexed over the total number of transcription factors. We verify that the validity of the dimensional reduction in Fig. 4 by comparing solutions of the reduced master equation for the probability distribution of the total copy number N (Eqn. 3, black dashed curves) to solutions of the three dimensional master equation indexed over the total copy number, N (Eqn. 1 colored curves).

APPENDIX B: TRANSLATIONAL BURST NOISE

Another source of noise in gene expression comes from multiple translation events of a single mRNA copy, so that proteins are effectively produced in bursts rather than one at a time [20]. Although our model does not include mRNA, we mimic the effects of bursting by specifying that each production event results in an instantaneous burst of B transcription factors with a reduced production rate of transcription factors, $g \to g/B$, such

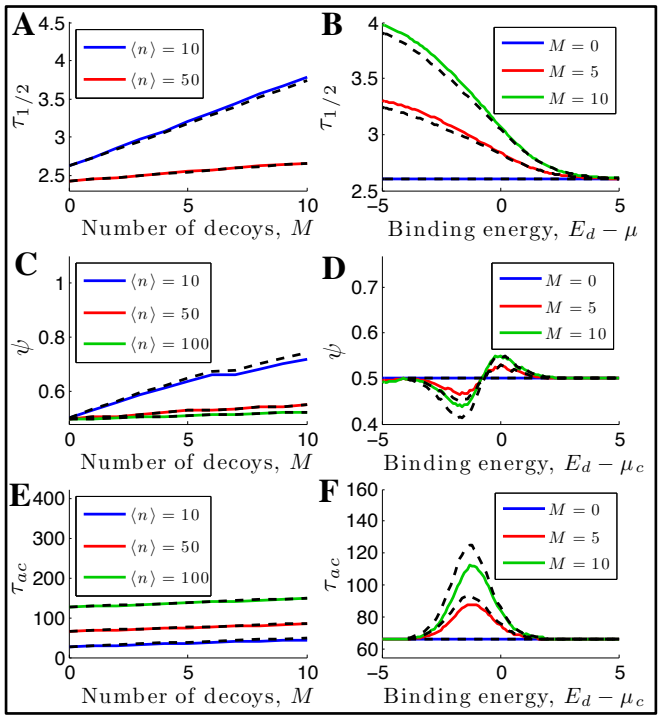

FIG. 4: **Validation of Dimensional Reduction.** Here we compare calculations from the full master equation indexed over total copy number, Eqn. 1, (*colored curves*) with calculations from the one dimensional master equation, Eqn. 3, (*black dashed curves*). We see that the dimension reduction breaks down for small system size, $\langle n \rangle$, or strong decoys, $E_d << \mu$. *Parameters:* $g_1 = 100S, g_0 = 8S =$ and in panels **A** and **B**, $S = 1$, $n_p^\dagger = 53.2$ $n_d^\dagger = 10$, in panels **C**, **D**, **E**, and **F** $n_p^\dagger = 10.3$ for $S = .2, n_p^\dagger = 21.0$ for $S = 1$, and $n_p^\dagger$ =106.8 for S =2, with $n_d^\dagger = \mu_c$.

that the average unbound number of transcription factors $\langle n \rangle$ does not change even though the variance *increases*. For a constitutively produced gene (where $g_0 = g_1$) the variance without decoys becomes $\sigma_0^2/\langle n \rangle = (B+1)/2$ [24]. Decoy binding sites that protect transcription factors from degradation have the opposite effect on the variance to bursts – they *decrease* the variance without changing the mean expression $\langle n \rangle$. The noise buffering formula derived above for $\sigma_n^2 = \sigma_{n,slow}^2 + \sigma_{n,fast}^2$ can be applied to a constitutively produced bursty gene as follows:

$$\begin{aligned} \sigma_n^2 &= \left(\sigma_0^2 - \langle n \rangle\right) \mathcal{J}^{-1}(\langle n \rangle) + \langle n \rangle \\ &= \langle n \rangle \left[\left(\frac{B-1}{2} \right) \mathcal{J}^{-1}(\langle n \rangle) + 1 \right]. \end{aligned} \quad \text{(B1)}$$

There are similar opposing effects between decoys and bursts when one considers the bimodal probability distribution. Large bursts can eliminate bimodality by decreasing the typical number of production events needed reach the transition state from a fixed point [25], such that the probability of the HIGH state decreases. Adding decoys that stabilize the HIGH state ($n_d^\dagger > \langle b \rangle$) can restore bimodality in a bursty bimodal system. Similarly, bursts exponentially decrease the time to escape between states [26], whereas decoys exponentially increase the time to escape between states.

APPENDIX C: APPROACH TO STEADY STATE

In this appendix we further discuss the limiting behaviour of the approach to steady state for the cases of weak and strong decoys.

Limit of weak decoys. For weak decoys ($n_d^\dagger >> \langle n \rangle$), approximating $\theta_d(\bar{n}) \approx \bar{n}/n_d^\dagger$ in Eqn 4 results in $N \approx \bar{n}(1 + M/n_d^\dagger)$ and $\mathcal{J}(\bar{n}) = \partial N/\partial \bar{n} = 1 + M/n_d^\dagger = const.$ In this limit the upper boundary of the integral becomes $\bar{n}(\langle N \rangle/2) \approx \langle n \rangle/2$ and $\tau_{1/2} = \tau_{1/2,0} + M\Delta\tau_{1/2}$ where $\Delta\tau_{1/2} \approx \tau_{1/2,0}/n^\dagger$.

Limit of strong decoys. For strong decoys ($n_d^\dagger << \langle n \rangle$), Eqn. 4 becomes $N \approx \bar{n} + M(1 - n^\dagger/\bar{n})$, and $\mathcal{J}(\bar{n}) \approx 1 + M n_d^\dagger/\bar{n}^2$. Therefore, unlike weak decoys that influence $\tau_{1/2}$ independently of $\bar{n}$, strong decoys have the most significant effect of increasing the time to reach the steady state (compared to the gene with no decoys) when $\bar{n}$ is small.

In the limit of extremely strong decoys, each transcription factor that is produced binds to a decoy site and remains bound. As a result, until all decoys are saturated, the unbound copy number will be zero. There will be no transcription factors available to bind to the promoter and the production will be fixed at the basal production level, g_0. After saturation, however, strong decoys no longer influence the dynamics of the system. Therefore the time to approach steady state can be broken up into a basal production stage and an isolated gene stage.

For $M > \langle n \rangle$, the time to reach half of the steady state number of proteins happens before the decoys are saturated - in the regime when transcription factors are produced with a rate g_0 per unit time,

$$\tau_{1/2} = \frac{\langle N \rangle}{2g_0} = \frac{M + \langle n \rangle}{2g_0}, \text{ for } M > \langle n \rangle >> n_d^\dagger. \quad \text{(C1)}$$

APPENDIX D: EPIGENETIC ESCAPE

To calculate the epigenetic escape times in Eq. 22, we define fixed points in total copy number, $N = \{\langle A \rangle, \langle B \rangle, \langle C \rangle\}$, that correspond to the fixed points in unbound copy number, $n = \{\langle a \rangle, \langle b \rangle, \langle c \rangle\}$. The mean escape time from $N = \langle A \rangle$ to $N = \langle C \rangle$ is [27]:

$$\tau_{AC} = 2\int_{\langle A\rangle}^{\langle C\rangle} dY \exp\big[W(Y)\big]\int_0^Y \frac{dZ}{D(Z)}\exp\big[-W(Z)\big], \tag{D1}$$

where

$$W(N) = -\int_0^N dN' \frac{2v(N')}{D(N')}. \tag{D2}$$

Within a Gaussian approximation about $N = \langle A\rangle$ and $N = \langle B\rangle$, Eqn. D1 becomes

$$\tau_{AC} = \frac{2\pi}{D(\langle A\rangle)}\sqrt{\frac{D(\langle A\rangle)D(\langle B\rangle)}{|\partial_N v(\langle A\rangle)|\,|\partial_N v(\langle B\rangle)|}}e^{-\int_{\langle A\rangle}^{\langle B\rangle} dN\frac{2v(N)}{D(N)}},$$

Performing a change of variables from N to $\bar{n}$, the escape time becomes

$$\tau_{ac} = \tau_{ac,0}\sqrt{\mathcal{J}(\langle a\rangle)\mathcal{J}(\langle b\rangle)}e^{-M\zeta_{ab}}, \tag{D3}$$

where $\tau_{ac,0}$ is the mean escape time without decoys and ζ_{ab} is the decoy perturbation to the action over the interval $[\langle a\rangle, \langle b\rangle]$:

$$\zeta_{ab} = \int_{\langle a\rangle}^{\langle b\rangle} d\bar{n}'\frac{2v_0(\bar{n}')}{D_0(\bar{n}')}\left[\frac{4(n_d^\dagger)^2\bar{n}'}{((n_d^\dagger)^2+\bar{n}'^2)^2}\right]. \tag{D4}$$

Likewise we find $\tau_{ca} = \tau_{ca,0}\sqrt{\mathcal{J}_2(\langle c\rangle)\mathcal{J}_2(\langle b\rangle)}e^{M\zeta_{bc}}$.

[1] U. Gerland, J.D. Moroz, T. Hwa, PNAS, **99**, 19 (2002).
[2] Z. Wunderlich and L.A. Mirny, Trends Genet., **25**, 10 (2009).
[3] J. R. Ecker *et al.*, Nature, **489**, (2012).
[4] T. Kaplan, *et al.*, PLoS Genetics, **7**, (2011).
[5] Y. Cao, *et al.*, Dev. Cell, **18**, (2010).
[6] P. J. Farnham, Nat. Rev. Gen. **10**, (2009).
[7] M. D. Biggin, Dev. Cell, **18**, (2010).
[8] L. D. Ward and H. J. Bussemaker, Bioinformatics, **24**, (2008).
[9] K. L. MacQuarrie, A. P. Fong, R. H. Morse, and S. J. Tapscott, Trends Genet. **27**, 4 (2011).
[10] S. Sengupta and C. C. Harris, Nat. Rev. Mol. Cell Biol.. **6**, (2005).
[11] S. B. Carroll, PLoS Biology, **3**, 7, (2005).
[12] J.T. Horng *et al.* IEEE T. Inf. Technol. B **7**, 2 (2003).
[13] T.H. Lee and N. Maheshri, Mol. Sys. Biol., **8**, 576 (2012).
[14] S. Lin and A. D. Riggs, Cell, **4**, (1975).
[15] P.H. von Hippel and O.G. Berg, PNAS **83** (1986).
[16] X.Y. Li *et al.* PLoS Biol. **6**, e27 (2008)
[17] A. Tanay, Genome Res., **16**, (2006).
[18] A. Burger, A.M. Walczak, P.G. Wolynes, PNAS, **107**, 9 (2010).
[19] O. Abu Hatoum *et al.*, Mol. Cell Biol., **18**, 10 (1998) .
[20] M. Kaern, *et al.*, Nat. Rev. Genet. **6**, (2005).
[21] A. M. Walczak, J.N. Onuchic, P.G. Wolynes, PNAS, **103**:52 (2005).
[22] J.E.M. Hornos *et al.*, Phys. Rev. E, **72** (2005).
[23] J.Elf, G.W. Li, X.S. Xie, Science, **316** (2007).
[24] M. Thattai and A. van Oudenaarden, PNAS **98**, 15 (2001).
[25] M. Scott, T. Hwa, and B. Ingalls, PNAS **104**, 18 (2007).
[26] P. Mehta, R. Mukhopadhyay, and N.S. Wingreen, Phys. Biol. **5**, (2008).
[27] C. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences* (Springer; 3rd edition, 2004).
[28] E. Aurell *et al.* Phys. Biol. **4**, 134 (2007).
[29] We set $n_d^\dagger = \langle n\rangle$ for $q = 1$ and $n_d^\dagger = \langle c\rangle$ for $q = 2$.
[30] Technically we could include an extra term here corresponding to the promoter occupancy, however we neglect this because we are interested in the limit of including large numbers of decoys.
[31] Equivalently, the same formula for $\sigma^2_{n,slow}$ can be obtained by first performing the change of variables on Eqn. 7, obtaining the effective drift $\widetilde{v}(\bar{n}) = v_0(\bar{n})\mathcal{J}^{-1}(\bar{n}) + 1/2D_0(\bar{n})\mathcal{J}^{-1}(\bar{n})\partial_{\bar{n}}\mathcal{J}^{-1}(\bar{n})$ and effective diffusion $\widetilde{D}(\bar{n}) = D_0(\bar{n})\mathcal{J}^{-2}(\bar{n})$, followed by the small noise approximation.
[32] For a burst size of one, decoys will *increase* noise for an auto-repressing gene (where $\sigma_0^2 < \langle n\rangle$) and have no effect on a constitutively produced gene (where $\sigma_0^2 = \langle n\rangle$).