Biophysical Journal, Volume *120*

Supplemental Information

Trajectory-based energy landscapes of gene regulatory networks

Harish Venkatachalapathy, Samira M. Azarin, and Casim A. Sarkar

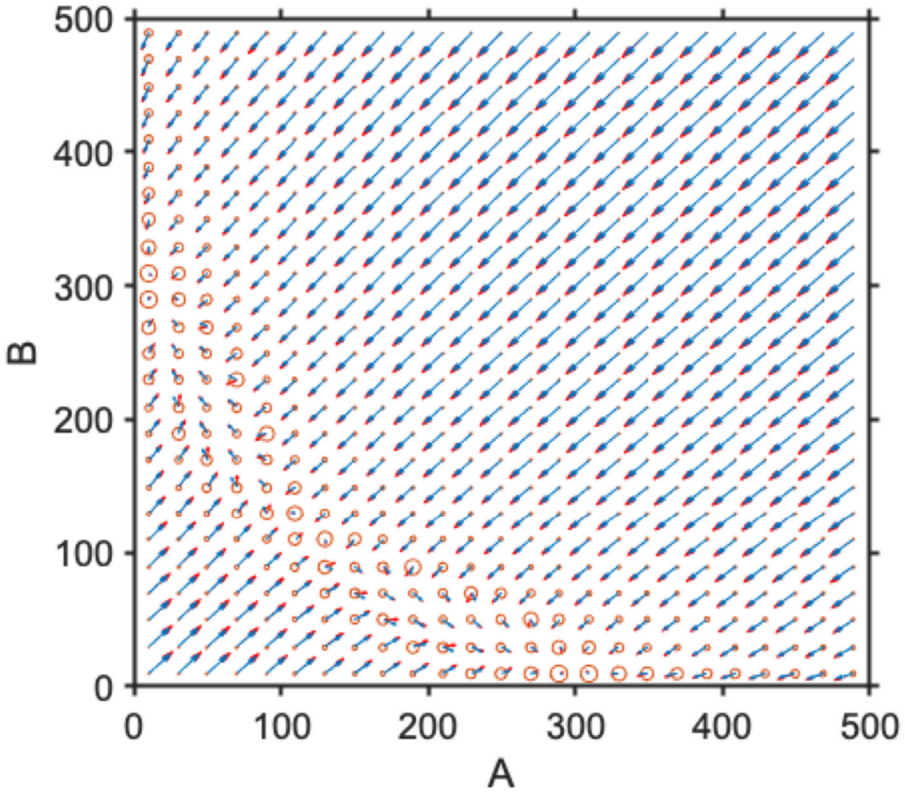

Fig. S1: Comparison between deterministic velocities (red) and trajectory-calculated velocities (blue).

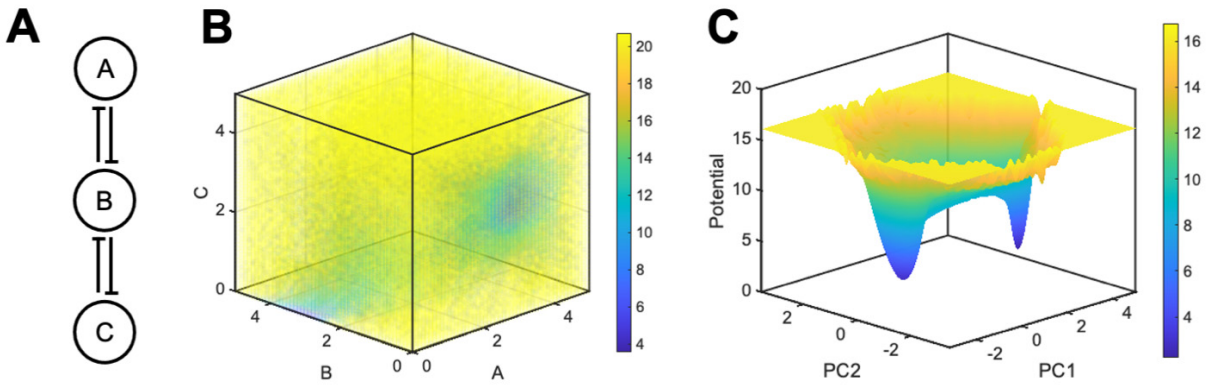

Fig. S2: Potential energy landscape representations of a three-species toggle switch. (A) Schematic of the three-species toggle switch. (B) A potential object for the three-species toggle switch, in which the color represents the potential energy at each coordinate. (C) Potential energy surface constructed using the first two principal components of the trajectories used to generate (B). The surface retains two steady states seen in (B) while retaining an intuitive 3D representation.

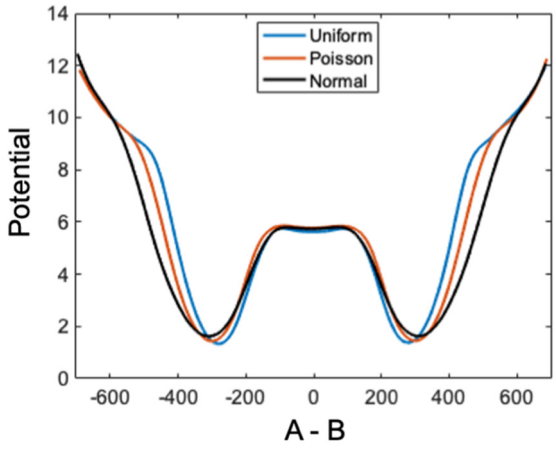

Fig. S3: Comparison between the potential energy surfaces generated using different mRNA synthesis distribution models. All distributions have been constrained to be positive and scaled to have the same mean and variance.

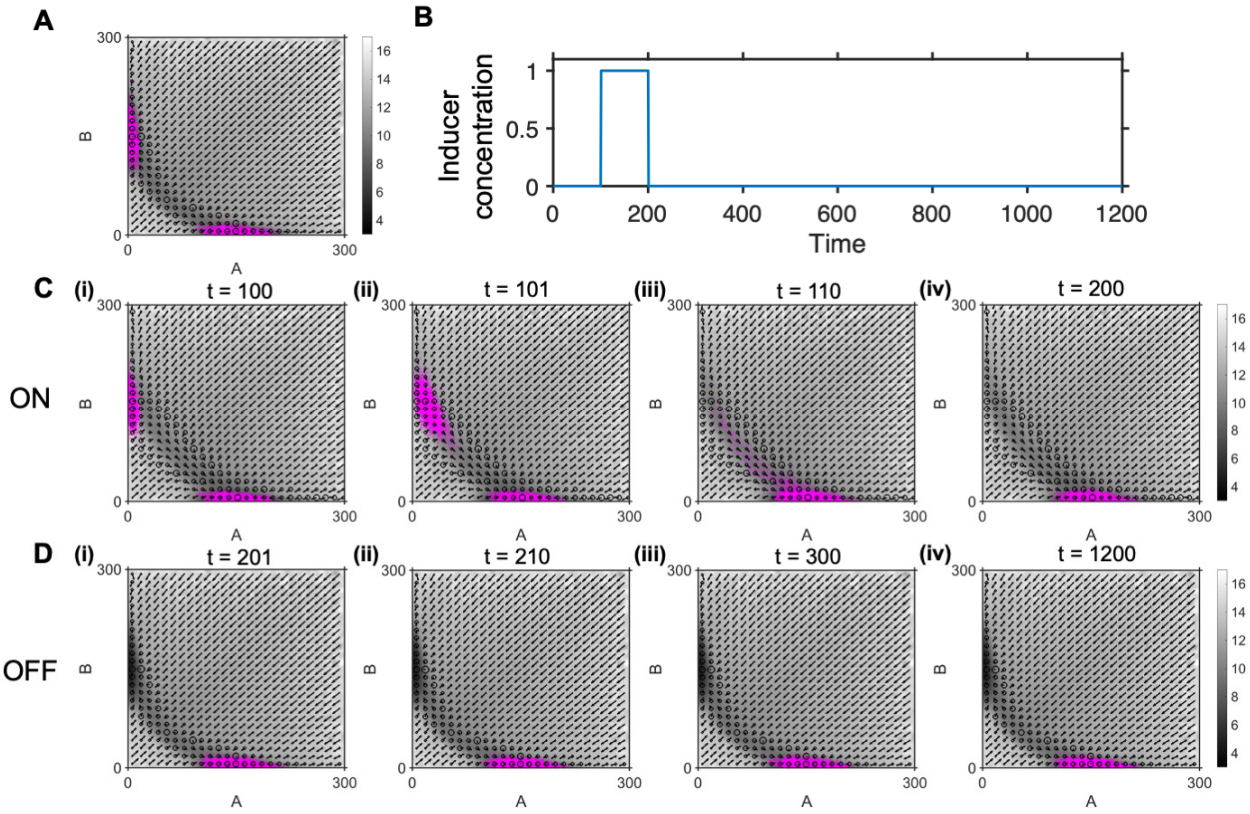

Fig. S4: Effect of inducer addition and removal on the toggle switch landscape and its effect on cells (magenta overlay) for burst size = 2. (A) Initial conditions for a toggle switch at t = 0 with cells in magenta. (B) Time course of inducer concentration applied to the system. (C) Effect of the change in landscape after inducer addition to the toggle switch system at different times: (i) t = 100, (ii) t = 101, (iii) t = 110, and (iv) t = 200. There is an instantaneous change in landscape followed by a transition to the induced valley. (D) Effect of the change in landscape after inducer removal to the toggle switch system at different times: (i) t = 201, (ii) t = 210, (iii) t = 300, and (iv) t = 1200. While the underlying landscape is reversible after inducer removal, cells do not transition to the other valley even after large amounts of time due to the lack of stochastic transitions.

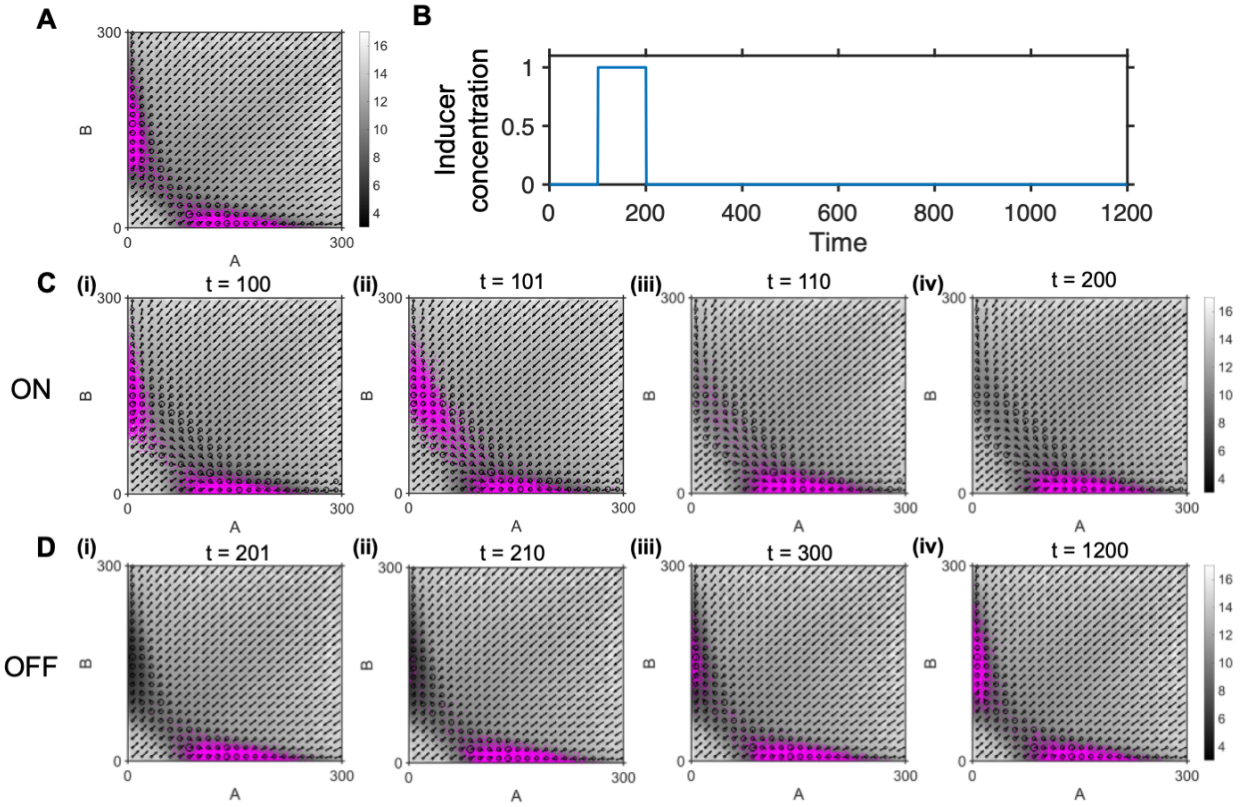

Fig. S5: Effect of inducer addition and removal on the toggle switch landscape and its effect on cells (magenta overlay) for burst size = 10. (A) Initial conditions for a toggle switch at t = 0 with cells in magenta. (B) Time course of inducer concentration on the system. (C) Effect of the change in landscape after inducer addition to the toggle switch system at different times: (i) t = 100, (ii) t = 101, (iii) t = 110, and (iv) t = 200. There is an instantaneous change in landscape followed by a transition to the induced valley. (D) Effect of the change in landscape after inducer removal to the toggle switch system at different times: (i) t = 201, (ii) t = 210, (iii) t = 300, and (iv) t = 1200. The underlying landscape is reversible after inducer removal and, further, cells transition back to the other valley given large amounts of time due to stochastic transitions. Based on the cell positions in D (ii-iv), this path is along the bridge between the two valleys.

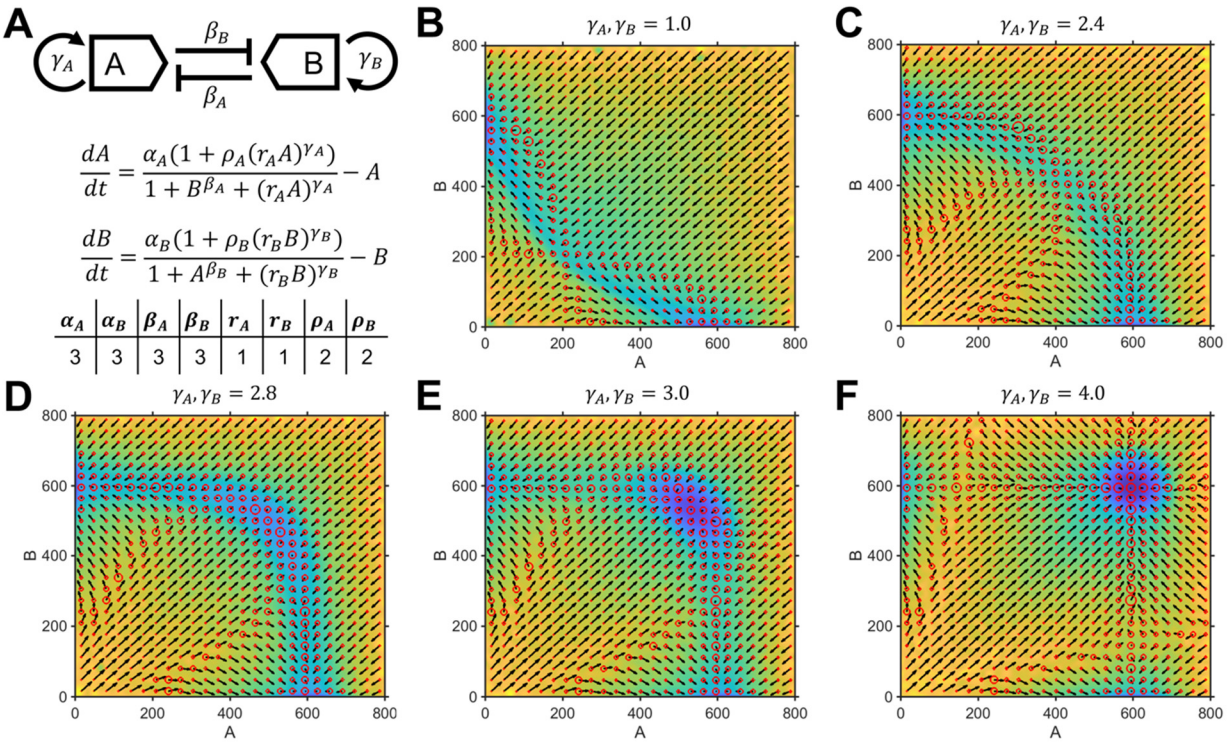

Fig. S6: Effect of cooperativity on a self-activating toggle switch (A) Schematic of a self-activating toggle switch with the corresponding equations and parameter values. r represents the relative ratio of activating binding and inhibitory binding; ρ is the relative increase in production rate upon self-activation. (B – F) Landscapes for different values of self-activation binding cooperativity. Increase in cooperativity initially results in increase of A and B at the saddle point, eventually resulting in a third stable steady state at high A and high B. Further increases in cooperativity lead to stronger demarcations between the stable steady states leading to higher potential barriers.

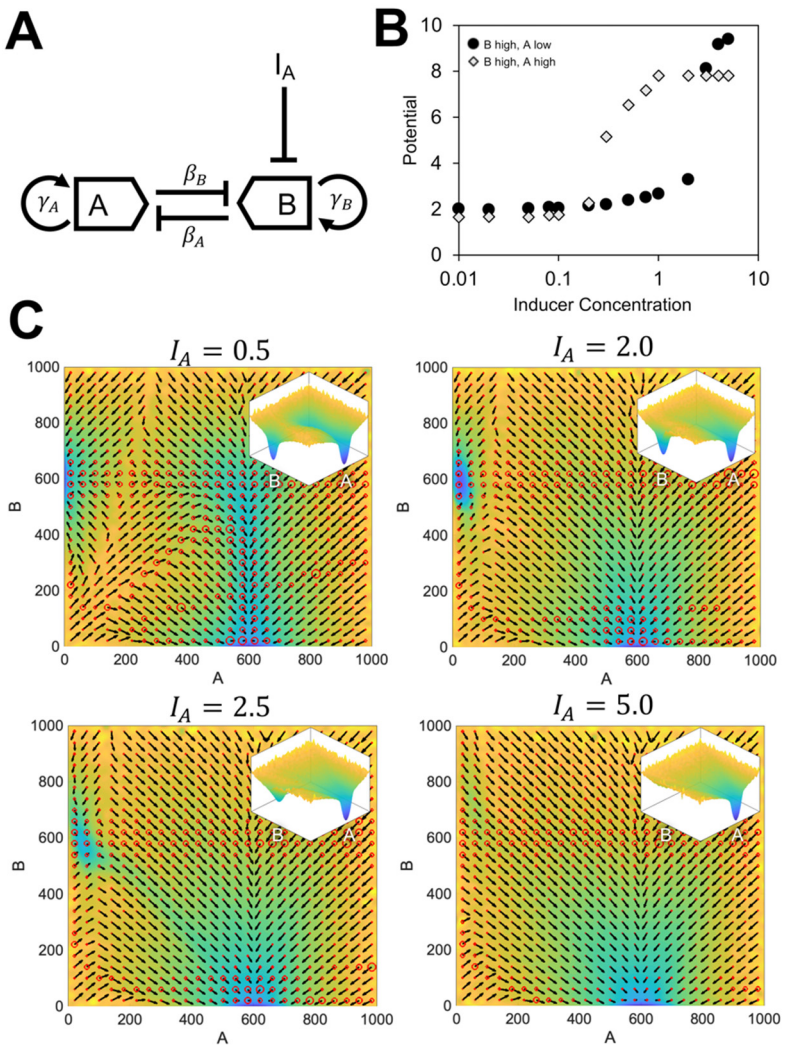

Fig. S7: Effect of inducer on a tristable self-activating toggle switch (A) Schematic of a self-activating toggle switch with inducer (B)Effect of inducer on the potential of the two non-induced states showing preferential destabilization of the A high, B high state followed by the A low, B high state. (C) Landscapes for different inducer concentration showcasing sequential destabilization and reduction in potential barriers with increasing inducer.

Supporting Text

Methodology for landscape generation

Overview

Multiple simulations of the model are run using kinetic Monte Carlo with input conditions that are sampled from the phase space of interest. These individual simulations are tracked through time and used to generate a probability distribution of the existence of a cell at a given point in the phase space. This probability (P) is then converted to a potential (U) described by $U = -\ln(P)$.

Input

Input conditions for the simulations are sampled randomly from a uniform distribution across the phase space of interest. A uniform distribution ensures that all points in the phase space are covered and there is minimal bias in generating the landscape.

Simulation methodology

For the simulations, we have chosen a kinetic Monte Carlo method with no approximations (Gillespie algorithm). This methodology captures the features required for analysis, including bistability and biological noise. However, other simulation methods can also be used for landscape generation if the resulting trajectories sufficiently capture the characteristic features needed to understand and analyze the system. In this paper, to convert the deterministic equations based on concentration into individual propensity functions based on number of molecules, a system Avogadro number (N_A) was employed. A smaller N_A leads to higher stochasticity while a larger N_A approaches a deterministic regime (**Fig. S8**). This parameter should be chosen to reflect the average number of molecules within the system and/or to capture the associated stochasticity and variance. The mean number of molecules in the system is proportional to N_A while the coefficient

of variation scales inversely with the square root of N_A. This follows from the fact that the underlying probability distribution in the Gillespie algorithm follows a Poisson distribution for which the coefficient of variation is $\lambda^{-\frac{1}{2}}$, where λ is the mean.

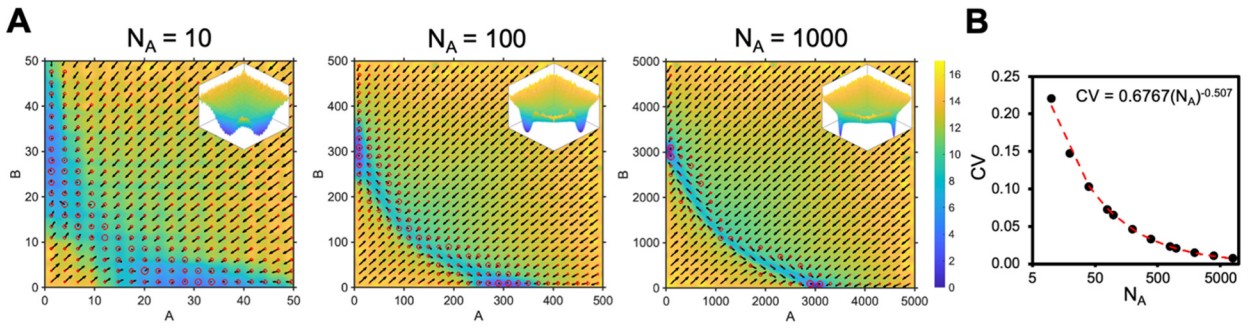

Fig. S8: (A) Landscapes generated for different system N_A where the x-axis is number of molecules of A and y-axis is number of molecules of B. (B) Plot of coefficient of variation of the number of molecules at steady state with the fitted power law curve.

Number of trajectories

The number of trajectories should be chosen so that it offers good coverage of the entire phase space. Practically, this represents the limit beyond which additional trajectories do not qualitatively change the landscape significantly. This limit in our system is reached with 10,000 cells. Beyond this, additional cell trajectories do not change the landscapes significantly (**Fig. S9**). This is also confirmed by a quantitative convergence analysis (1) that tracks the standard deviation of the percentage of cells in each steady state (**Fig. S10**) and shows that 10,000 is the least number of cells with a low standard deviation in steady-state occupancy frequency while having a mean steady state occupancy frequency consistent with a higher number of cells.

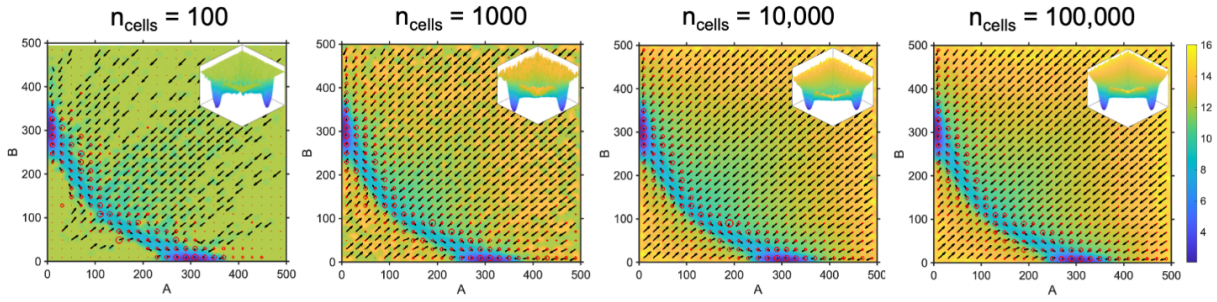

Fig. S9: Landscapes generated for different numbers of cells (or alternatively, different numbers of initial conditions).

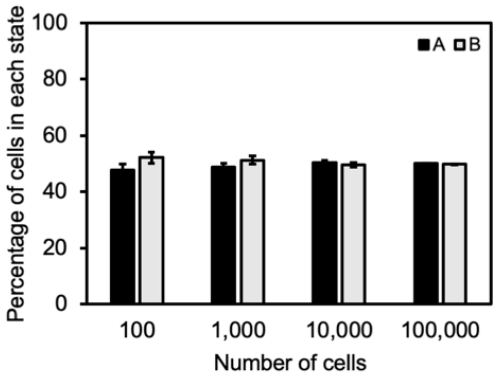

Fig. S10: Quantitative convergence analysis showing the percentage of cells in each steady state with error bars showing the standard deviation across three independent replicates.

Simulation length

Simulation time must be chosen such that both steady state and intermediate details are captured in the landscape. This is the minimum time beyond which there is negligible change in steady state potentials. Choosing longer simulation times would result in underweighting the approach to steady state within the probability distribution used to construct the potential energy function, which would obscure intermediate dynamics on the potential surface. The simulation length should be maintained across multiple parameter changes to provide a commensurate comparison between landscapes. In our simulations, this time was determined to be 100 (dimensionless units) (**Fig. S11**).

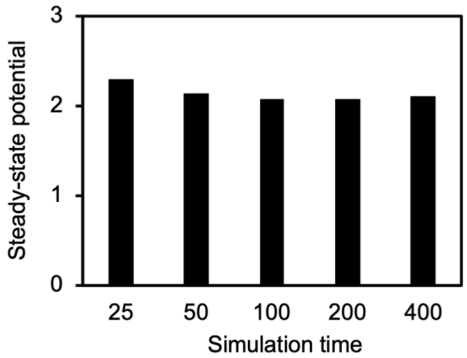

Fig. S11: Mean steady-state potential on the landscape for different simulation lengths.

Trajectory sampling

To visualize the landscape, we generate a probability distribution based on the probability of a cell existing with a particular phase space specification at any point during the simulation. While every point in the simulation trajectories could be used for generating the probability distribution, this results in exceedingly large file sizes and proves difficult for analysis. To overcome this, we sample each trajectory at specific intervals of time. This sampling can change between systems and can be determined by identifying the minimum time interval below which there are no qualitative changes in the landscape. In our system, this time interval is 0.1. Finer sampling, as in the 0.01 case, does not notably change the landscapes (**Fig. S12**).

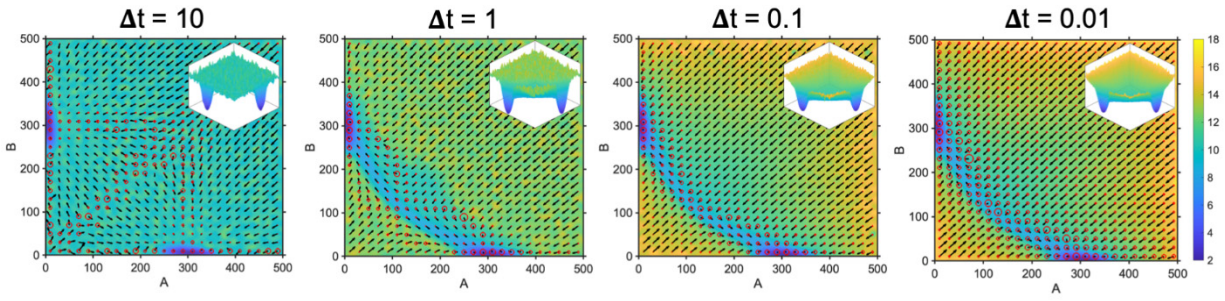

Fig. S12: Landscapes generated for different sampling intervals.

Potential energy generation

Using the common statistical mechanics conceptualization, we convert the probability function to a pseudo-potential energy using $U = -\ln(P)$. This energy surface is then "smoothed" using spline interpolation and visualized in MATLAB using the *surf* function.

Velocity generation

To generate the velocities, we use a finite-difference formula on the trajectories. For a cell whose trajectories are represented in terms of two variables, A and B, both of which are functions of time, the velocity $\vec{v}$ at a time t_i is given by

$$\vec{v} = \overrightarrow{v_A} + \overrightarrow{v_B} \tag{1}$$

$$\overrightarrow{v_A}(t_i) = \left.\frac{dA}{dt}\right|_{t_i} \hat{a} = \frac{A(t_{i+1}) - A(t_i)}{t_{i+1} - t_i}\hat{a} \tag{2}$$

$$\overrightarrow{v_B}(t_i) = \left.\frac{dB}{dt}\right|_{t_i} \hat{b} = \frac{B(t_{i+1}) - B(t_i)}{t_{i+1} - t_i}\hat{b} \tag{3}$$

where $\hat{a}$ and $\hat{b}$ are unit vectors in the directions of increasing A and B, respectively.

The phase space is then discretized into unit blocks, and the average and coefficient of variation of all the velocity vectors originating within each block is calculated. These velocity averages are represented by the relative lengths of each arrow shown in the landscape (as visualized by the *quiver* function in MATLAB). The coefficient of variation is then represented by the relative size of the circle around the base of each arrow (visualized using *scatter* in MATLAB). The discretization of the phase space depends on the level of detail required from the landscape. Larger unit blocks are better used for landscapes where less variability is expected while smaller unit blocks are better for visualizing regions where the velocity vectors have large variances (e.g., at the valley edges) (**Fig. S13**).

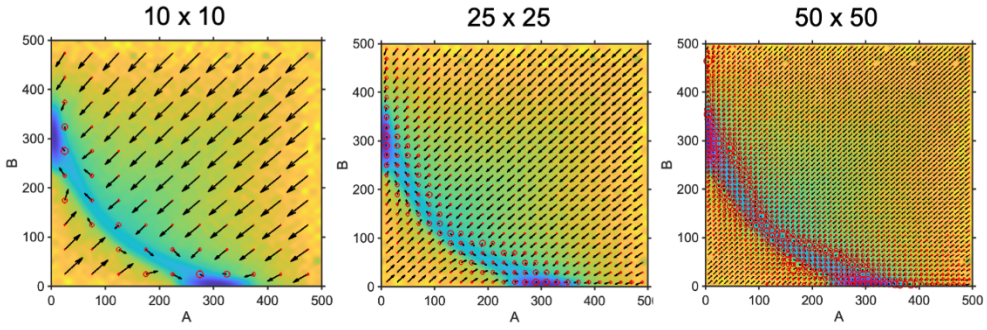

Fig. S13: Landscapes generated for different grid sizes.

Models used for simulation

Genetic toggle switch by Gardner et al. (2)

For two proteins, A and B, that cross-antagonize by binding to the opposite DNA and preventing subsequent protein production, the rates of change are given by

$$\frac{dA}{dt} = \frac{\alpha_A}{1 + B^{\beta_A}} - A \tag{4}$$

$$\frac{dB}{dt} = \frac{\alpha_B}{1 + A^{\beta_B}} - B \tag{5}$$

where α_i and β_i are the basal rates of production and cooperativity of cross-antagonistic binding, respectively.

DNA copy number

The production term in the original model is valid for regimes in which there are a large number of DNA molecules so that the rate can be determined as a weighted average of the production rates during the unbound (active) and bound (inactive) states. However, in cases with a single DNA copy number, the DNA is either bound or unbound and thus the production rate cannot be considered an average of the two states. We explicitly model this behavior by setting the production term for molecule i (being repressed by molecule j) as either 0 or α_i with the following probabilities based on the cross-antagonistic binding propensities

$$P(production\ propensity = \alpha_i) = \frac{1}{1 + j^{\beta}}\ (DNA_i\ is\ unbound) \tag{6}$$

$$P(production\ propensity = 0) = \frac{j^{\beta}}{1 + j^{\beta}}\ (DNA_i\ is\ bound) \tag{7}$$

For two copies of DNA, as in the S/G2/M phases of the cell cycle, we simulate them by independently considering two molecules of DNA that are either bound or unbound based on the probabilities above. These two DNA molecules are then each active for half as much time as when a single molecule is within the system, and this propensity is given by $\alpha_i/2$.

Transcriptional bursting

Under bursty transcription, multiple protein molecules are formed under a single burst of

production. Instead of explicitly modeling two-stage transcription and translation, we posit that every time the protein production term is chosen by the Gillespie algorithm, there are multiple protein molecules produced. This number is sampled from a Poisson distribution with different means depending on the "burst size" of the protein. This assumption is valid under cases where the mRNA degradation rates are much greater than the protein degradation rates (see derivation below).

The two-stage model proposed by Strasser et al. (3) is

$$\frac{d(d_i)}{dt} = \tau_j^-(1 - d_i) - \tau_j^+ d_i n_j \tag{8}$$

$$\frac{d(m_i)}{dt} = \alpha_i d_i - \gamma_i m_i \tag{9}$$

$$\frac{d(n_i)}{dt} = \beta_i m_i - \delta_i n_i + \tau_i^-\left(1 - d_j\right) - \tau_i^+ d_j n_i \tag{10}$$

where d_i, m_i, and n_i are the DNA, mRNA, and protein concentrations respectively. τ_j^- and τ_j^+ are the unbinding and binding rate constants of protein n_j and DNA d_i. α_i and γ_i represent the production and degradation rate constants of mRNA. Similarly, β_i and δ_i represent the production and degradation rate constants of protein. We consider both $(i = A,\ j = B)$ and $(i = B, j = A)$ pairings.

For this model to reduce to the form of the Gardner model, we apply the quasi-steady-state assumption (QSSA) to the mRNA and DNA rate equations. This reduces the protein rate equations to

$$\frac{d(n_i)}{dt} = \frac{{\alpha_i \beta_j}/{\gamma_i}}{1 + \frac{\tau_j^+}{\tau_j^-} n_j} - \delta_i n_i \tag{11}$$

For further congruence through scaling, we let $n_i' = n_i\left(\frac{\tau_i^+}{\tau_i^-}\right)$, $n_j' = n_j\left(\frac{\tau_j^+}{\tau_j^-}\right)$ and $t' = t\delta_i$. Then,

$$\frac{d(n'_i)}{dt'} = \frac{\alpha_i \beta_i \tau_i^+ / \gamma_i \delta_i \tau_i^-}{1 + n'_j} - n'_i \tag{12}$$

To determine the validity of the QSSA, we apply a scaling analysis to the DNA and mRNA equations. Let $d'_i = d_i$ and $m'_i = \frac{\gamma_i}{\delta_i} m_i$. Then,

$$\left(\frac{\delta_i}{\tau_j^-}\right) \frac{d(d'_i)}{dt'} = 1 - d'_i - d'_i n'_j \tag{13}$$

$$\left(\frac{\delta_i}{\gamma_i}\right) \frac{d(m'_i)}{dt'} = d'_i - m'_i \tag{14}$$

Thus, the QSSA is valid when $\delta_i \ll \tau_j^-$ and $\delta_i \ll \gamma_i$. In other words, the DNA and mRNA kinetics have to be much faster than the protein kinetics.

Toggle switch with inducer

In this version of the toggle switch, we explicitly incorporate inducer molecules that increase synthesis of one protein by binding to the other protein and abrogating its ability to bind its target DNA to repress gene expression. The following equations represent this augmented toggle switch with inducers I_A and I_B of proteins A and B, respectively. In deriving these equations, we have assumed that the rates of inducer binding and unbinding to their respective proteins are much faster than other processes in the system.

$$\frac{dA}{dt} = \frac{\alpha_A}{1 + \left(\frac{B}{(1 + (z_B I_A)^{a_B})}\right)^{\beta_A}} - A \tag{15}$$

$$\frac{dB}{dt} = \frac{\alpha_B}{1 + \left(\frac{A}{(1 + (z_A I_B)^{a_A})}\right)^{\beta_B}} - B \tag{16}$$

where I_i represents the concentration of inducer of i, z_i represents the selectivity of j binding to

the inducer of i compared to cross-antagonistic binding to the DNA of i, and a_i represents the cooperativity of inducer binding. All conditions within the paper involving an inducer have been simulated with $I_B = 0$, $z_B = 1$, and $a_B = 1$.

Supporting References

1. Hari, K., B. Sabuwala, B.V. Subramani, C.A.M. La Porta, S. Zapperi, F. Font-Clos, and M.K. Jolly. 2020. Identifying inhibitors of epithelial–mesenchymal plasticity using a network topology-based approach. *npj Syst. Biol. Appl.* 6:1–12.
2. Gardner, T.S., C.R. Cantor, and J.J. Collins. 2000. Construction of a genetic toggle switch in *Escherichia coli*. *Nature*. 403:339–342.
3. Strasser, M., F.J. Theis, and C. Marr. 2012. Stability and multiattractor dynamics of a toggle switch based on a two-stage model of stochastic gene expression. *Biophys. J.* 102:19–29.

Additional Supporting Files

Video S1. Dynamically changing landscape under sinusoidal input

Video S2. Dynamically changing landscape with cells overlaid at frequency = 0.05

Video S3. Dynamically changing landscape with cells overlaid at frequency = 0.5

Video S4. Dynamically changing landscape with cells overlaid at frequency = 0.1

Video S5. Dynamically changing landscape with cells overlaid at frequency = 0.2

Data S1 (TrajectoryBasedLandscapes.zip). Compressed folder with MATLAB code for visualization and C++ code for toggle switch trajectory generation