Intrinsic Limits of Information Transmission in Biochemical Signaling Motifs

Supporting Information

Ryan Suderman[1] and Eric J. Deeds[1,3]

[1]Center for Computational Biology, The University of Kansas, 2030 Becker Dr., Lawrence, KS 66047
[3]Department of Molecular Biosciences, The University of Kansas, 1200 Sunnyside Ave., Lawrence, KS 66047

Email: Eric J. Deeds - deeds@ku.edu;

Contents

1 Estimating the channel capacity

In order to estimate the channel capacity from a set of dose-response data, we use the method described in ref. (1), the code for which is published online at https://github.com/ryants/EstCC. Here we provide an outline of our methodology, adapted from the calculations derived in ref. (2). Note that the units of information are bits, assuming a base-2 logarithm.

1.1 Binning

The first step is to take the dose-response data set and partition the dose (signal) and response (response) spaces into bins. This effectively results in a contingency table, showing frequencies of dose-response pairs, given some discretization of the data set. From this contingency table, we can perform a naive estimation of the mutual information:

$$I(S;R) = -\sum^{S} p(s) \log p(s) + \sum^{R} p(r) \sum^{S} p(s|r) \log p(s|r) \tag{1}$$

where we substitute the appropriate frequencies from the contingency table into the marginal and condtional probabilities seen in Eq 1.1.

The mutual information of a data set is dependent on the discretization, and so care must be taken when choosing the numbers of bins. To achieve a reasonable discretization, we perform the same mutual information calculation as above, but with shuffled pairings of the signal and response values. Randomization should theoretically result in 0 bits of information, and so we increase the number of bins in both dimensions until significant levels of information are observed in the randomized data set (an artifact of excessive binning). This provides a cutoff for defining this reasonable discretization of signal and response space.

1.2 Linear extrapolation

To remove bias that arises due to having a finite sample size, we use a bootstrapping procedure. After finding a reasonable set of bins for the data as described above, we calculate the naïve estimate of the mutual information for randomly sampled, smaller subsets of the data at varying percentages. We then model the mutual information as a linear function of inverse sample size, and extrapolate the value of the mutual information at infinite sample size.

1.3 Signal distribution trials

Since we are attempting to estimate the channel capacity (the supremum of the mutual information over all signal distributions)

$$C = \sup_{p_S(s)} I(S;R) \tag{2}$$

we perform a brute force optimization over a predefined set of signal distributions. We take the original data set and apply weights along the signal dimension to alter the frequencies in the contingency table, while maintaining a constant overall number of entries. Our estimated channel capacity, $\hat{C}$, is the greatest mutual information calculated over all attempted signal distributions. This procedure provides a reasonable lower-bound estimate of the channel capacity.

2 Framework

2.1 Model with low Hill coefficient

Shown in Figure S1 are the channel capacity as a function of the signal window (A) and the number of sampled signal values (B). These only differ from those in the main text in terms of the Hill coefficient used to generate the data. These plots have Hill coefficients of 6 as opposed to those in the main text that had Hill coefficients of 60. As can be clearly seen, there is no significant difference between the two data sets.

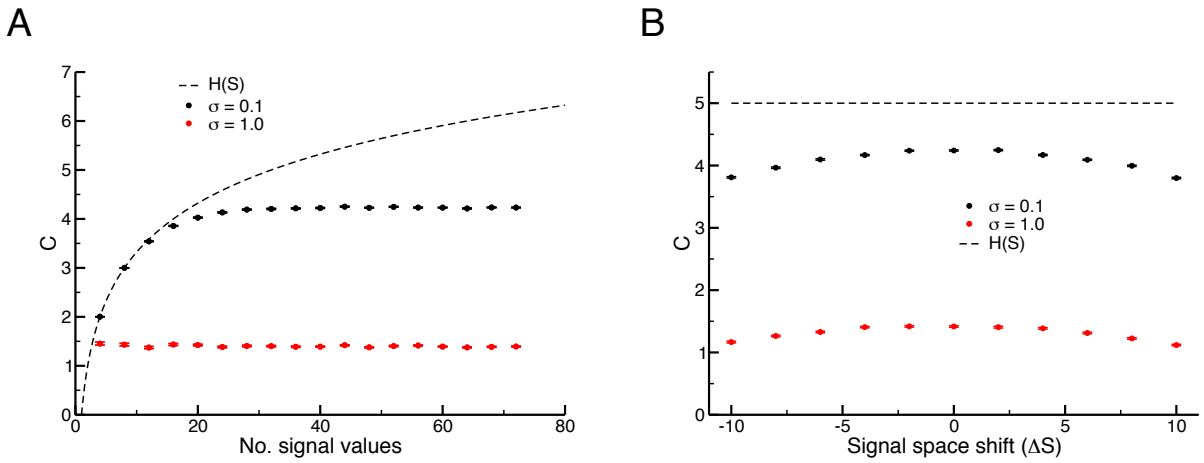

Figure S1: A & B. Similar to Figures 1C & D in the main text, but the data was generated with a Hill coefficient of 6 instead of 60.

2.2 Varying the transition zone bounds

We performed a brief analysis of the simple model to characterize the effects of varying the response range of the transition zone that is shown in Figure S2. In the main text all transition zones are constructed using bounds of 10% and 90% maximal response (after subtracting basal response). We thus varied the percentage of response space removed from consideration to examine its effect on the transmission of information in this simple model. Though there is some variation, the trends are relatively flat and appear to depend somewhat on the variability in response. As a result, we will maintain our focus on the 10%-90% transition zone with the caveat

that modifying this parameter could introduce minor fluctuations in the resulting channel capacity values.

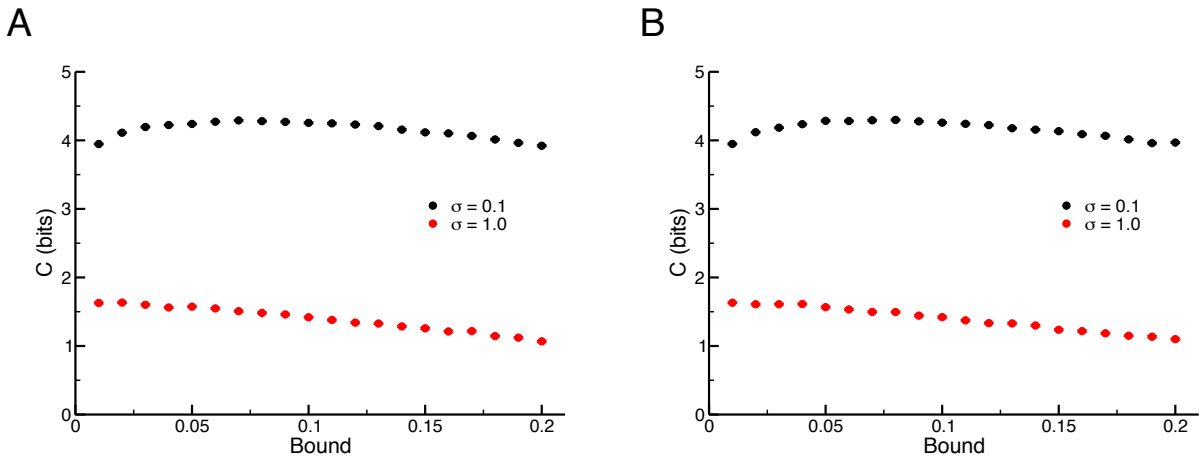

Figure S2: Varying the relative width of the transition zone shows minimal impact on the simple model regardless of utilizing high (A) or low (B) Hill coefficients in generating the data. The x-axis shows the fraction of response values that are removed from both high and low ends of response space (i.e.if the response interval is [0,10] and the bound is set to 0.1, then the transition zone interval is [1,9]). As can be seen, the choice of bound impacts the information estimation to some degree, but not sufficiently to warrant the computationally intensive search for the bounds optimizing each calculation. We use a bound of 0.1 or 10% throughout this work, and in this example the deviation from the optimal values is marginal at best.

2.3 Finding the transition zone empirically

For the majority of stochastically simulated models, we empirically determined the bounds of the transition zone by sampling a range of values that span the increasing regime of the dose-response curve and appear to approach the minimum and maximum responses. In the case of the kinase cascade models and the covalent modification cycles, the dose-response curves were sufficiently sigmoid in nature to allow a least squares regression fit to a Hill function. The resulting functional fit was inverted to find the signal values which bound the transition zone. Since the recruitment of Sos to the EGFR molecule is not so well represented by a simple sigmoid function, we took a different, but equally straightforward approach. We similarly characterized the majority of the dose-response curve in signal space and proceeded to use linear interpolation to estimate values corresponding to the signal bounds of the transition zone.

3 Ligand-Transmembrane Receptor model

The motif underlying this model is a simple, reversible binary interaction between two proteins; while we refer to these two proteins as the Ligand and Transmembrane Receptor (hence *LT* model), this really could represent any binary molecular interaction. As mentioned in the main

text, the parameters we used for this model were not chosen to represent any specific case of such an interaction in biology, but to rather range across a series of reasonable values. A binary interaction model like this has two rate parameters (the association rate k_+ and the dissociation rate k_-. Since we are only considering steady-state responses here, the only parameter of importance is the dissociation constant, $K_D = k_-/k_+$. In order to maintain a constant relative level of saturation in the binding interaction as we increase the total number of T molecules in the system, we kept the ratio T_T/K_D constant and equal to 1 for all of the cases we considered here. To do that, we set the stochastic dissociation rate in our simulations as $k_- = 0.01$, and set the association rate for each simulation equal to the numerical value of T_T. For each value of T_T we considered, we used the approach described in the main text to define a set of input ligand numbers (i.e. the set of L_T values) to use for the simulations. For each T_T, L_T pair, we determined the length of time needed to reach steady-state in the simulaiton by visual inspection of the average number of LT complexes over time. After determining a time that was sufficient for the system to reach steady state, we ran a set of independent simulations, and used the final value of LT from each simulation to create the data set needed for the channel capacity calculation. The 'LT.ka' file provided as additional supplementary material contains a template model, written in version 3 of the Kappa rule-based modeling language, for this system. The simulations we performed were achieved by modifying the initial number of L and T agents, and the kinetic parameters of the rules, as described above.

In oder to add synthesis and degradation to the model, we introduced four new rules: two that introduce the L and T agents to the system (with rates Q_L and Q_T), and two that remove them from the system (with rates δ_L and δ_T. In this model, the total number of agents is controlled not by changing the initial conditions, but by changing the ratio Q/δ. As above, since we are only interested in steady-state behavior, it is only this ratio that matters for determining the behavior of the system. In this model, we kept the Q parameters constant (equal to 1) and varied the δ parameters to change protein numbers.

In order to determine the bounds of the transition zone, we used the following system of ordinary differential equations for the LT model with synthesis and degradation:

$$\begin{aligned}
\frac{\mathrm{d}L}{dt} &= k_-[C] - k_+[L][T] - \delta_L\left([L] + [C]\right) + Q \\
\frac{\mathrm{d}R}{dt} &= k_-[C] - k_+[L][T] - \delta_T\left([R] + [C]\right) + Q \\
\frac{\mathrm{d}C}{dt} &= -k_-[C] + k_+[L][T] - \left(\delta_L + \delta_T\right)[C]
\end{aligned}$$

where $[L]$,$[R]$, and $[C]$ are the Ligand, Transmembrane Receptor, and Complex concentrations, δ_M and Q are the degradation and synthesis rates of some molecule M, and k_- and k_+ are unbinding and binding rates. Note that total ligand and receptor concentrations are defined as $[L_T] = [L] + [C]$ and $[T_T] = [T] + [C]$ and that synthesis rates of L and T are equivalent, but not degradation rates. Furthermore, we parameterized our model such that $\dfrac{Q}{\delta_M} = M_T$. We can then solve for C as a function of L_T at equilibrium and invert the equation to determine the transition zone bounds, where $C_{\max} = 0.9 \cdot T_T$ and $C_{\min} = 0.1 \cdot T_T$ are the maximum and minimum

responses in the transition zone, respectively. The equation for C given L_T is quadratic:

$$0 = k_+[L_T][T_T] - (k_+[L_T] + k_+[T_T] + k_- + \delta_L + \delta_T)[C] + k_+[C]^2$$

and finding L_T is trivial, with knowledge of the parameter values.

We can see in Figure S3 how the number of sampled signal values in the transition zone alters the estimated information transmission in the LT model that includes synthesis and degradation of both ligand and receptor components.

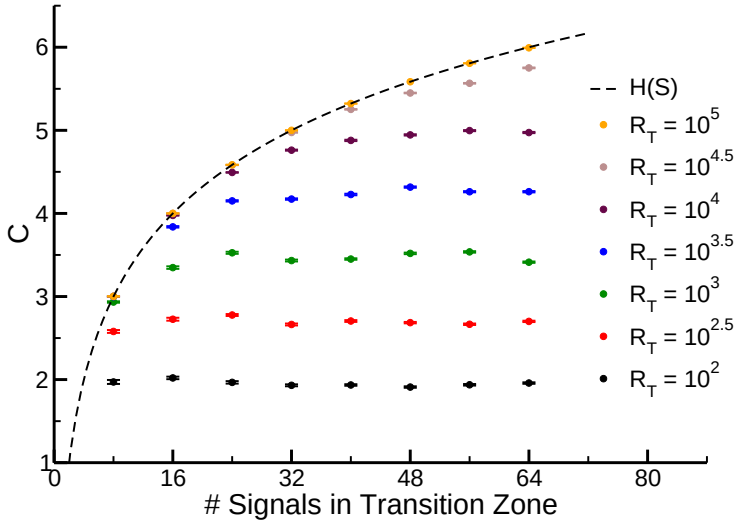

Figure S3: Similar to Figure 2C in the main text, but using the LT model that includes rules for synthesis and degradation.

4 Covalent modification cycle model (the GK Loop)

In their seminal work on the covalent modification cycle, Goldbeter & Koshland mathematically characterized the steady-state level of phosphorylation of a substrate protein W given two enzymes, K and P which catalyze the addition and removal of the modification, respectively, and the enzymatic parameters of those enzymes (their catalytic rates, the k_cat's, and their Michaelis constants, the K_M's) (3). Their analysis identified the primary variable that controls the steady-state response of the substrate, which they termed "r":

$$r \equiv \frac{k_{cat,K} \cdot K_T}{k_{cat,P} \cdot P_T}$$

where K_T and P_T are the total numbers of kinase and phosphatase molecules, respectively. The r parameter is traditionally considered to be the input signal to a covalent modification cycle (3). Traditionally, r is varied by modifying the copy number/concentration of the kinase, but this approach would introduce three problems in our calculations. For one, varying the number of K molecules in the simulation would simultaneously change both the input signal value *and* the statistical properties of the system that might arise from low copy numbers. Secondly, a stochastic simulation like the ones we performed can only have an integer number of each type of species, so, at the low end of the signal spectrum, we could only ever realize signal values corresponding to 0 or 1 K molecule. In performing our simulations, there were many cases in which we needed to sample signal values that would correspond to numbers of kinases between 0 and 1, which is simply impossible in this kind of simulation framework. Finally, the fact that we are limited to integer values of K_T restricts to total number of signal values we could have in the transition zone, limiting the maximum value of $H(S)$ and thus the calculated channel capacity as described in the main text.

Since we are only interested in steady-state responses in this case, only the parameter r, the K_M's, and the total amount of K, P and W matter for determining the response of the system (3). We thus chose to vary r by varying the kinetic rate of the phosphatase, $k_{cat,P}$. Of course, varying this parameter effects not only r, but also the K_M of the Phosphatas-substrate reaction:

$$K_{M,P} = \frac{k_{cat,P} + k_{-,P}}{k_{+,P}}$$

where the kinetic constants represent the standard steps of a traditional Michaelis-Menten enzyme reaction (3). To keep the level of saturation of the phosphatase constant while varying r, we used the following procedure. First, we set the catalytic rate of the kinase, $k_{cat,S}$ to be 1 s^{-1}, and we set both of the off rates to be $k_{-,K} = k_{-,p} = 0.01\ s^{-1}$. Our simulations considered a variety of total substrate numbers (W_T values), and, for each W_T, we also considered a set of saturation levels for the enzymes (the K_M/W_T values). For any given K_M/W_T ratio that we wanted to consider, we first defined the requisite K_M for the kinase by setting $k_{+,K}$ to whatever value was necessary to give us the K_M that we wanted. We then varied r for that system by changing $k_{cat,P}$ as described above. To ensure that $K_{M,K} = K_{M,P}$, we varied the association rate of the phosphatase:

$$k_{+,P} = \frac{k_{-,P} + k_{cat,P}}{K_{M,K}}.$$

which allows us to maintain a desired level of saturation for both enzymes regardless of the r value we choose.

As described in the main text, we used a set of preliminary simulations at each value of W_T and the K_M's to determine the boundaries of the transition zone, and then generated a set of signal values within that transition zone to use for the channel capacity calculation. We ran simulations at each value of signal as described for the LT model above. A template Kappa file for this model is included as supplementary material, and one can obtain any simulation that we ran with the appropriate modification of the kinetic rates for the rules.

Addition of synthesis and degradation in this model was done in a manner similar to that for the LT case. We introduced new rules for the synthesis of the basic species (K, P and W) and

Parameter	Value
Association	$10^{-7}\,(\text{molec}\cdot\text{sec})^{-1}$
Dissociation	$0.1\ \text{sec}^{-1}$
Catalysis	$1\ \text{sec}^{-1}$
Kinase ($1 \leq i < d$)	10^4 copies
Kinase (d)	10^5 copies
Scaffold	10^4 copies
Phosphatase (1)	10^5 copies
Phosphatase ($1 < i \leq d$)	10^3 copies

Table S1: Table of parameter values for the scaffold and solution kinase cascade models, where d denotes the depth of the cascade. Note that the association rate here is in units of "per molecule per second" because the simulations are stochastic. If we posit a particular volume in which the reactions are occurring (say, the volume of a yeast cell), it is straightforward to convert these units to the typical $\text{M}^{-1}\text{s}^{-1}$ used for deterministic simulations.

degradation rules for all of the relevant species and complexes. As in the LT case, we set the Q parameters uniformly to 1, and varied the total number of each relevant molecule type by varying its corresponding degradation rate δ. Our approach to sampling signal space by varying $k_{cat,P}$ while keeping the K_M's constant was the same as for the model without synthesis and degradation as described above. A template Kappa file for this model is also included as additional supplementary material.

5 Kinase cascade models

The two scaffold models we considered are also meant to be schematic representations of the types of signaling networks found in biological systems, but their complexity prevented us from exploring the influence of parameters like copy number and saturation on information transmission as completely as we did in the LT and GK Loop models. As discussed in the main text, we parameterized these models using parameters broadly inspired by copy numbers and catalytic rates observed in the yeast MAP Kinase signaling cascade. The parameters themselves may be found in Table S1. As with all of the models discussed here, template rule files are available as additional supplementary material.

In this model, we used the copy number of an explicit "activator" agent as the input signal. While this more closely mimics the realistic biological situation observed in, e.g. MAPK cascades, the issues with this approach that we discussed in the section on GK loops discussed above still apply. In particular, the lowest possible non-zero input signal value in this case is $S = 1$. As the depth of the solution cascade increases, the response becomes extremely sensitive to input signals; in other words, it takes very little input signal to fully activate K_F in larger solution cascade models (Figure 4C in the main text). We used two parameters in the model to counteract this problem. For one, the association rate between the kinases is relatively slow (Table S1), though still well within the range of biologically feasible values. Perhaps more importantly, however, we introduced a large number of phosphatases for the first kinase in the cascade (K_1, see Table **??**). This

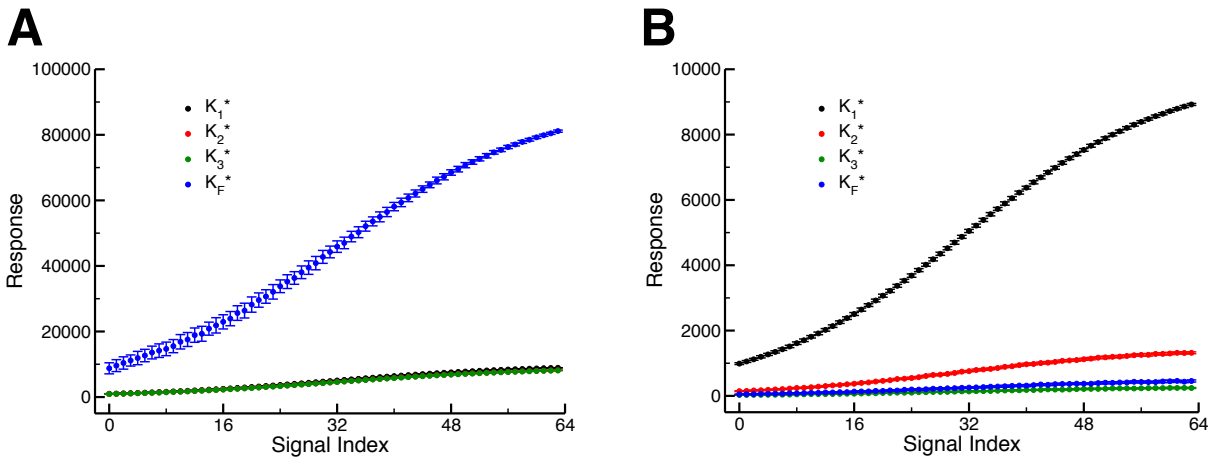

Figure S4: (A) Dose-response data for the solution model using the VTZ approach. Though difficult to see, all observables approach 90% activation. (B) As (A) but for the scaffold model. The trends here are distinct from (A), where only K_1 approaches 90% activation. The other observables, including K_F, fall well short of maximal activation.

allowed us to capture all the transitions in activity in all the intermediates in the solution model (Figure 4C in the main text) using an explicit activating molecule for the cascade.

To simplify the comparison between the solution and scaffold models, we used the same parameter sets for both. A result of this, however, is the fact that, while the kinases in the solution cascade reach full activation, those in the scaffold model do not S4. This lower level of overall response may be one of the reasons that the scaffold model consistently encodes lower levels of information than the solution model (Figure 5 in the main text). As mentioned above, these models are large and have many parameters, so we have left a systematic exploration of the sensitivity of information transfer to parameter values in this model to future work. Such an exploration will be necessary in order to definitively understand the relative information transmission potential of scaffold- vs. solution-based kinase cascades. Regardless, it is clear from our analysis that both solution and scaffold-based kinase cascades can encode much more information than has ever been observed experimentally for cell signaling at the molecular level.

The results presented in the main text focus on a single type of cascade with four kinases (i.e. $F = 4$). We also considered other cascade depths (namely $F = 2$ and $F = 3$). As we can see from Figre S4, the general trend here is similar to that seen for a single cascade: as the number of interactions in the cascade increases, overall information transfer capacity decreases.

5.1 Two component signaling

Bacterial two component signaling motifs are similar in certain aspects to the covalent modification cycle described above. A key distinction, however, is the fact that the histidine kinase, when inactive, operates as the phosphatase in the cycle. In contrast to the models discussed above, this model is meant to represent a specific biological example in cell signaling,

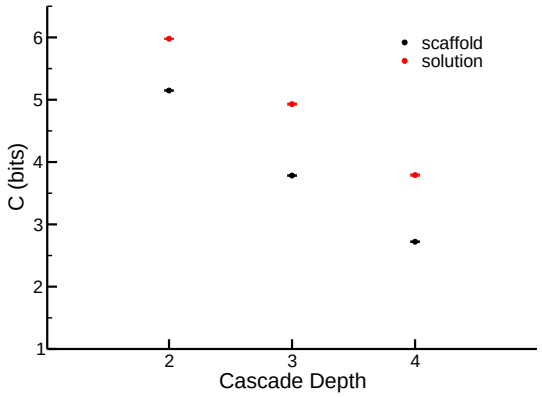

Figure S5: Information transmission to final kinase ($C(S; K_F)$) for various cascade depths

and so our parameters were taken from experimental data on TCS signaling in bacterial cells. The specific example in this case was the EnvZ/OmpR system in *E. coli*, but those parameters are applicable to many other TCS systems (4; 5; 6). The specific values of all of the parameters (off rates, on rates, etc.) may be found in the Kappa rule file provided as additional supplementary material.

Although this is a specific biological example, we did perform a local exploration of how changes in enzyme saturation and total substrate copy number would influence information transmission. As described in the main text, the autophosphorylation rate of the HK is the biologically realistic input parameter for this model (4; 5), so less care was needed here when changing saturation than was required in the GK Loop model. Changes in saturation were achieved by varying the association rate between the HK and the RR substrate (keeping the ratio between the K_M's for the active vs. inactive HK constant). The most likely parameter regime in terms of K_M's and substrate copy numbers is highlighted with the red box in Figure 6C in the main text.

5.2 EGFR model

We obtained this model from ref. (7) and translated it into Kappa without any modification of the rules or parameters. The Kappa file we used as the basis for these simulations is provided as additional supplementary material, and all the parameters we used may be found there. A full discussion of how this model was developed and parameterized, as well as an analysis of a number of its dynamical properties, may be found in ref. (7).

References

1. Suderman, R, Bachman, J. A, Smith, A, Sorger, P. K, & Deeds, E. J. (2017) *Proceedings of the National Academy of Sciences* **114**, 5755–5760.

2. Cheong, R, Rhee, a, Wang, C. J, Nemenman, I, & Levchenko, a. (2011) *Science* **334**, 354–358.

3. Goldbeter, A & Koshland, D. E. (1981) *Proceedings of the National Academy of Sciences of the United States of America* **78**, 6840–6844.

4. Rowland, M. A & Deeds, E. J. (2014) *PNAS* **111**, 5550–5555.

5. Batchelor, E & Goulian, M. (2003) *Proceedings of the National Academy of Sciences of the United States of America* **100**, 691–696.

6. Skerker, J. M, Prasol, M. S, Perchuk, B. S, Biondi, E. G, & Laub, M. T. (2005) *PLoS Biology* **3**, e334.

7. Blinov, M. L, Faeder, J. R, Goldstein, B, & Hlavacek, W. S. (2006) *Bio Systems* **83**, 136–151.