nature portfolio

https://doi.org/10.1038/s41567-021-01380-3

Supplementary information

Escherichia coli chemotaxis is information limited

In the format provided by the authors and unedited

Supplementary Information for *E. coli* chemotaxis is information-limited

H.H. Mattingly[1,2,†], K. Kamino[1,2,†], B.B. Machta[3,4,*], T. Emonet[1,2,3,*]

[1] Department of Molecular, Cellular, and Developmental Biology, Yale University.

[2] Quantitative Biology Institute, Yale University.

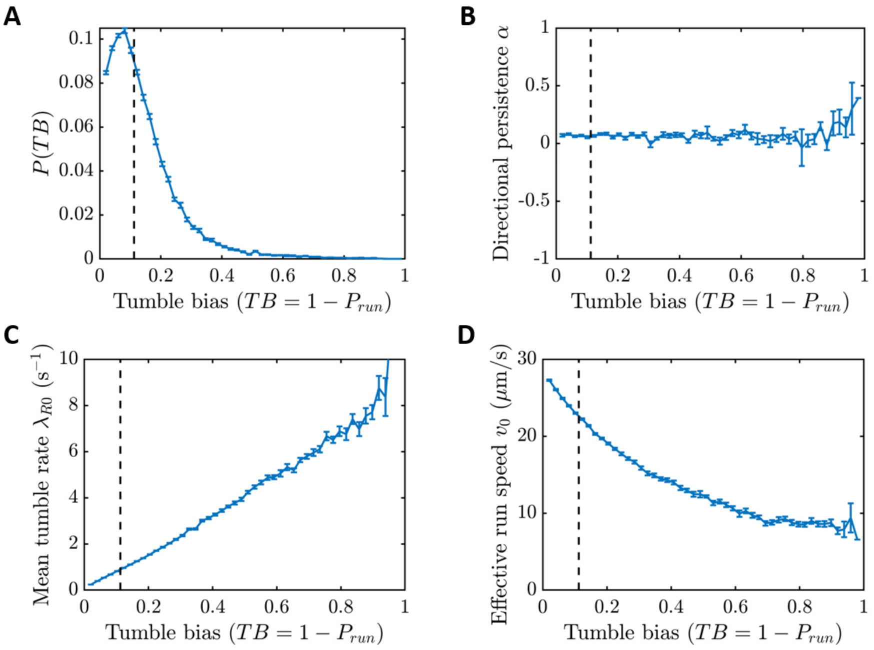
[3] Department of Physics, Yale University.

[4] Systems Biology Institute, West Campus, Yale University.

† These authors contributed equally.

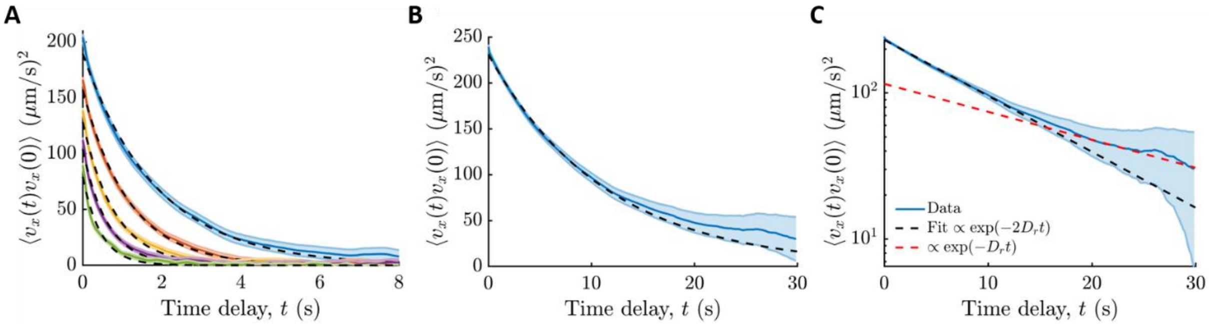
* Correspondence to: thierry.emonet@yale.edu and benjamin.machta@yale.edu

Table of Contents



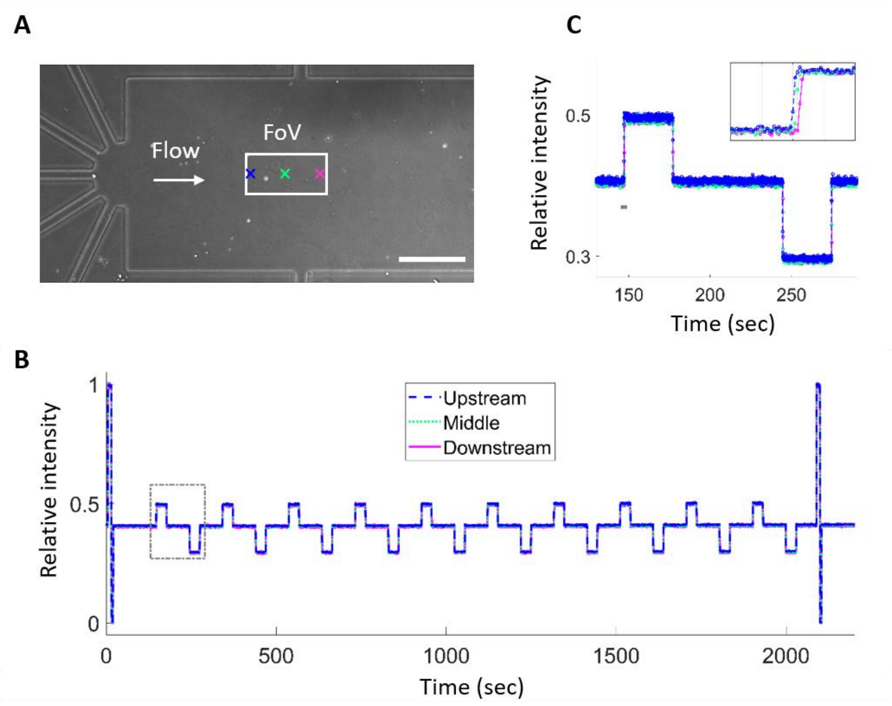
Supplementary Figure S1. Average swimming parameters in bins of tumble bias

A) Probability mass function of tumble bias $TB = 1 - P_{run}$ in populations of RP437 recorded with a 4X objective (Methods; bin size = 0.02). Each cell was weighted by its trajectory duration when generating this distribution. Error bars in each panel were determined by boot strapping (SI). In all panels, the vertical black dashed line indicates the population-median P_{run}. **B-D)** In bins of tumble bias: (B) Average directional persistence of tumbles, defined as $\langle \cos(\Delta\theta) \rangle$, where $\Delta\theta$ is the angle between the cell's heading vectors before and after runs; (C) Average tumble rate λ_{R0}; and (D) Effective run speed v_0. During transitions between run and tumble states, the swimming speed takes finite time to reach a steady value [1,2]. We lump these transitions into the run state, leading to effective run speeds that depend on how often the cell tumbles, i.e. P_{run}. Excluding these transitions would make cells appear to climb gradients artificially slowly compared to their run speeds. It would also lead to over-estimated information rates.



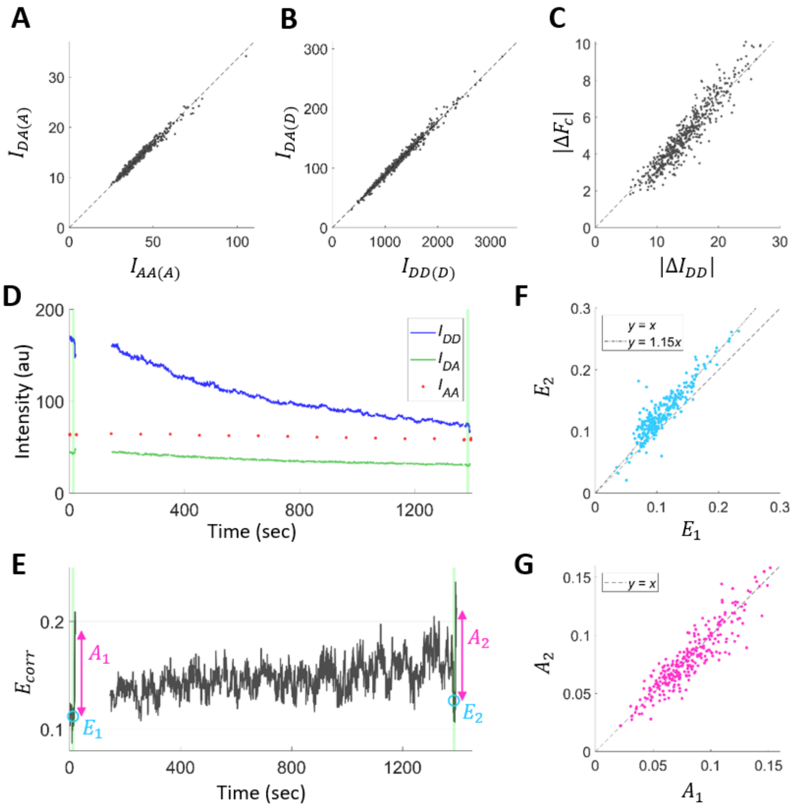
Supplementary Figure S2. Velocity autocorrelation functions for RP437 and VS100

A) Autocorrelations of the x-velocity of wild type (RP437) *E. coli* cells in the absence of a gradient, averaged over cells in bins of P_{run} (bin size 0.005). From blue to green, P_{run} values are: $0.93, 0.89, 0.84, 0.79, 0.74$. The orange line is the correlation function the median P_{run}. Curves in this panel are reproduced from Fig. 2C of the main text. The black dashed lines are the best fits of a decaying exponential to each correlation function (SI). Shading is ± one standard error, which was computed from the standard deviation of the samples of $v_x(0)v_x(t)$ at each time delay t, divided by $\sqrt{n_i(t)}$. $n_i(t) = n(t)\frac{\Delta t}{\Delta t+2/\lambda_{tot}}$ is the effective number of independent measurements at time delay t [3], where $n(t)$ is the total number of observations at delay t, Δt is the imaging interval (50 ms), and λ_{tot} is the best fit to the decay rate of each exponential. Since samples come from multiple cells, whose velocities are uncorrelated, this underestimates the actual n_i. **B)** Population-averaged x-velocity autocorrelation function of VS100 cells, which lack the *cheY* gene and therefore cannot tumble. Their correlation function is expected to decay exponentially with rate $\lambda_{tot} = 2\, D_r$, from which we inferred the rotational diffusion coefficient, D_r. Black dashed line is an exponential fit to the first 10 seconds of time delay. **C)** Same as (B) (velocity autocorrelation function for VS100 cells), but with the y-axis on log-scale. At long time delays, there is a bias for cells that remained visible, i.e. cells that had small vertical component to their velocity and therefore remained in the depth of field of our microscope objective. These cells by chance appear to be undergoing rotational diffusion in two dimensions instead of three; therefore, they lose direction at a rate D_r instead of $2\, D_r$. This transition did not affect our inference of D_r.



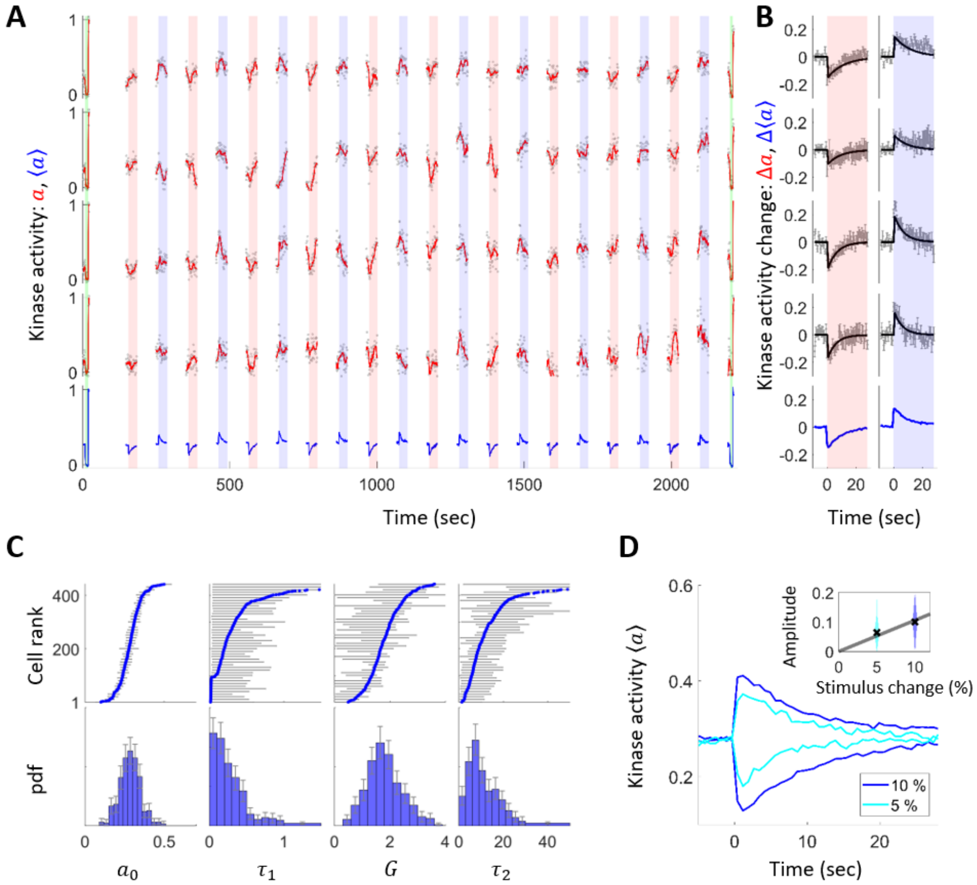
Supplementary Figure S3. Stimulus control by microfluidics

A) An image of the microfluidic device used in the FRET experiments (Kamino et al, *Sci. Adv.* 2020). The white scale bar at the bottom right is 200 µm. The direction of flow is shown by a white arrow and the position of a field of view (FoV) for FRET measurements is shown by a white rectangle. The three locations at which the temporal profiles of delivered chemical solutions were measured and plotted in panels (B) and (C) are indicated by the three colored crosses. Different solutions were delivered into the FoV from the five inlet channels shown on the left (Methods). **B)** The temporal profile of chemoattractant stimuli that the cells experience during the FRET experiment corresponding to main text Fig. 2D and Supplementary Fig. S5. Each solution delivered to the FoV through each inlet channel was visualized by using different concentrations of fluorescein, and the relative fluorescence intensities of the solutions were measured at three different positions in a 60X field of view (see panel (A)) and the measurement at each location is plotted in the color of its corresponding cross in panel (A). Five solutions were delivered to the imaging region by computer-controlled solenoid valves, which control the pressure applied to each channel (Methods). **C)** A magnified view of the gray box part of panel (B). The inset shows further magnified plots of the time window indicated by the horizontal gray bar to resolve the transient rise of fluorescence intensity. The X-grid in the inset is 1 s interval. Intensities were obtained at 75 ms intervals, indicated by the circles. The transient rise and delay between positions within the FoV are on the order of 0.1 s.



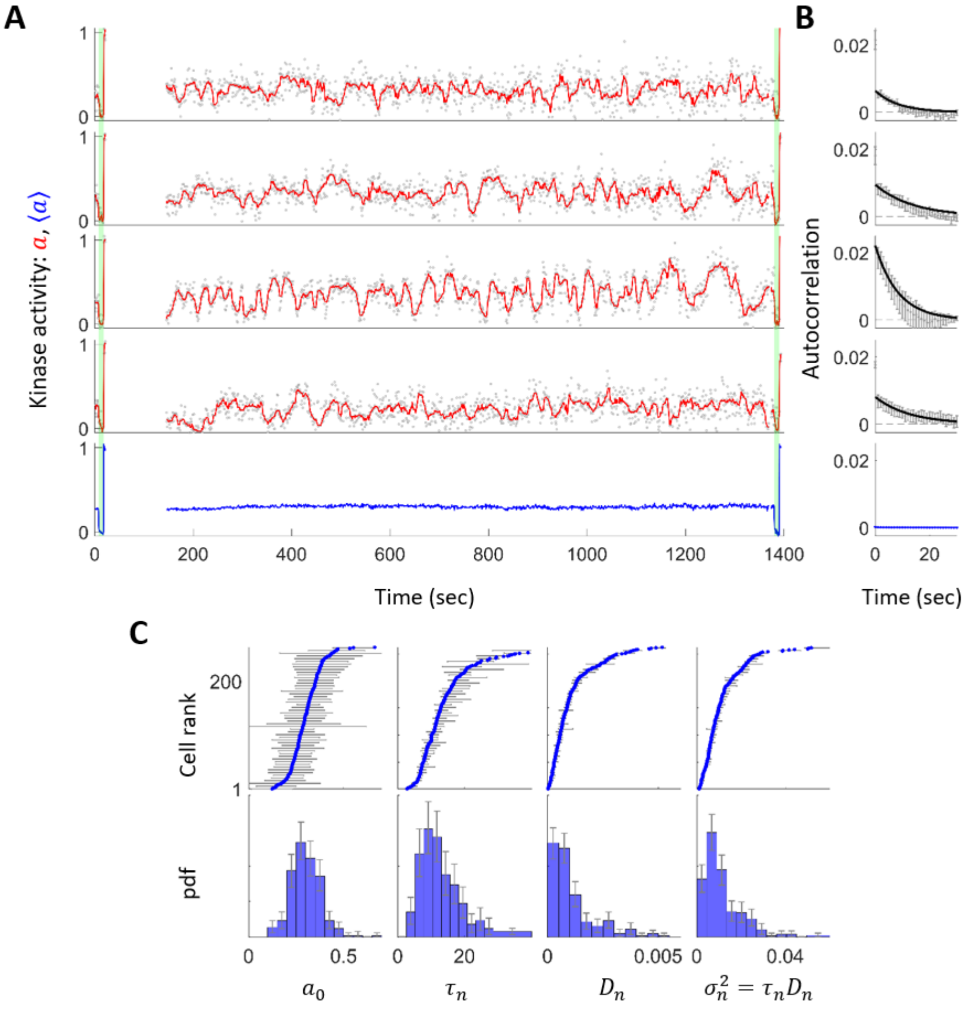
Supplementary Figure S4. FRET analysis

A) Single-cell fluorescence intensities from a strain that only expresses the acceptor (mRFP), obtained through the acceptor channel $I_{AA(A)}$ and the FRET channel $I_{DA(A)}$. The slope gives an estimate of the cross-excitation coefficient $a_E = 0.337$. **B)** Single-cell fluorescence intensities from a strain that only expresses the donor (mYFP), obtained through the donor channel $I_{DD(D)}$ and the FRET channel $I_{DA(D)}$. The slope gives the estimation of the bleedthrough coefficient $d_E = 0.089$. **C)** Absolute changes in the donor fluorescent signal $|\Delta I_{DD}|$ and the sensitized emission $|\Delta F_c|$ before and just after the removal of a saturating chemoattractant stimulus obtained from a FRET strain that expresses both the donor and acceptor. The slope gives the parameter $G_E = 0.35$. **D)** Representative time series of background-subtracted fluorescent signals $I_{DD}(t)$, $I_{DA}(t)$, and $I_{AA}(t)$ from a FRET experiment. Green shading at the start and end of the experiment indicate times when a saturating stimulus (1 mM MeAsp and 100 µM serine) was applied, which were followed by the removal of all attractants (i.e., [MeAsp] = [serine] = 0 M) for 5 seconds. After this, the rest of the experiment was performed in a background of 100 µM MeAsp. The fluorescence intensities decrease over time due to photobleaching. We show in the SI that the photobleaching and the finite precision of the parameter estimation (A-C) results in a bias in FRET estimation, but we also show that it can be corrected (E-G). **E)** The FRET index $E_{corr}(t)$ (SI) computed from the fluorescent signals $I_{DD}(t)$, $I_{DA}(t)$, and $I_{AA}(t)$ in panel (D). **F)** The median of $E_{corr}(t)$ during the saturating stimuli at the beginning (E_1) and at the end (E_2) of the measurement in panel (E). Consistent with the theoretical analysis (SI), the level of E_2 shows systematic deviation from the level of E_1 (i.e., from the line $y = x$) as a result of photobleaching of the fluorescent proteins. **G)** The change in E_{corr} induced by the removal of the saturating stimulus at the beginning (A_1) and at the end (A_2) of the measurement defined in panel (E). Consistent with the theoretical analysis (SI), the signal changes show undetectable bias after photobleaching, distributing around the line $y = x$. Thus, we can correct for the slowly-increasing bias in the absolute value of $E_{corr}(t)$ over time (E and F) by subtracting the trend, while also preserving information about signaling-induced changes in E_{corr} (SI).



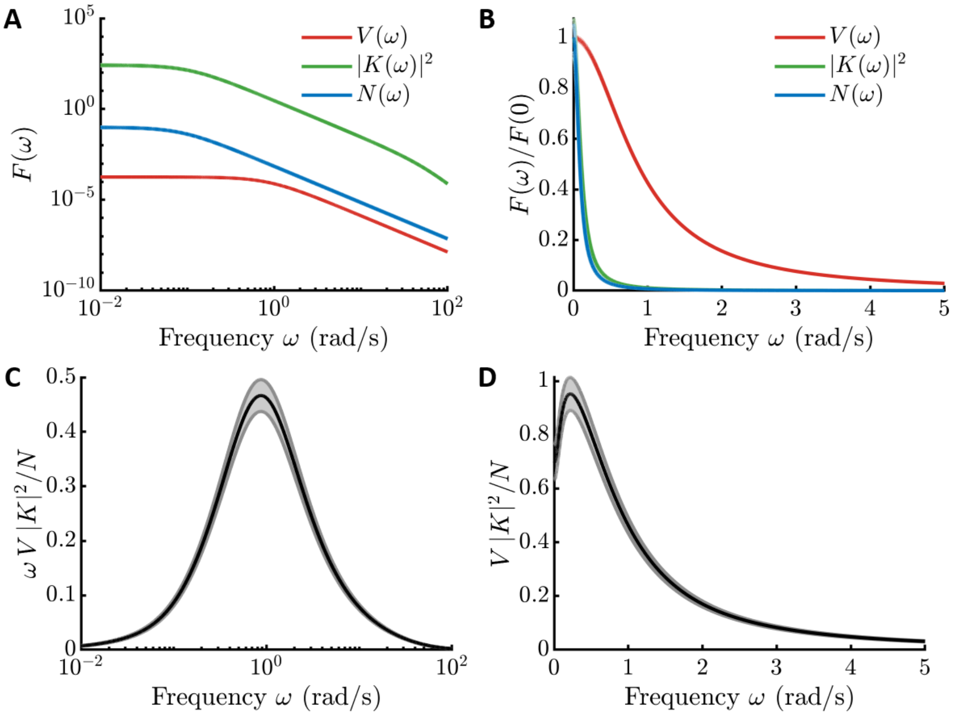
Supplementary Figure S5. Extracting linear response function

A) Representative time series of single-cell kinase activity a (SI for definition; red lines) and population average $\langle a \rangle$ (blue lines). Gray dots are raw data points and red lines are smoothed curves (10th order median filter; 7.5 s time window). Red and blue shading indicates times at which [MeAsp] is elevated by 10 μM and decreased by 10 μM from 100 μM, respectively. Green shading indicates when saturating stimuli (1 mM MeAsp and 100 μM serine) were applied to measure the minimum kinase activity, which were followed by the removal of the attractants (i.e., [MeAsp] = [serine] = 0 M) for 5 seconds to measure the maximum kinase activity. **B)** Averaged kinase responses. Gray lines with error bars are the within-cell average and standard error of the mean of the change in kinase activity Δa, defined as the change in a from its pre-stimulus value (left: step-up responses; right: step-down responses). Best fit model curves are shown in black (SI). Population-averaged kinase activity changes $\Delta\langle a \rangle$ are shown at the bottom in blue. **C)** (Top) Values of the extracted parameters sorted by the mean (for a_0) or maximum a posteriori probability (MAP) estimates (for τ_1, G, and τ_2). The errorbars are standard errors (for a_0) or 25 and 75 percentiles of the posterior distributions (for τ_1, G, and τ_2), and are shown for one in every ten cells. (Bottom) Marginal histograms of each parameter. 95% bootstrap confidence intervals are shown by the error bars. **D)** Time series of population-averaged kinase activity $\langle a \rangle$ upon 5% and 10% change in [MeAsp] from 100 μM background (changed at time zero; both step-up and step-down responses are shown). The inset shows the distribution of single-cell response amplitudes (defined as the average of $|\Delta a|$ over the first 3 s after the stimulus) upon step concentration changes. Both step-up and step-down responses are lumped into the same distribution. The gray line shows $y = a\,x$ fitted to the population average shown by black crosses. Consistent with previous work [4], the responses show quasi-linearity in this concentration range.

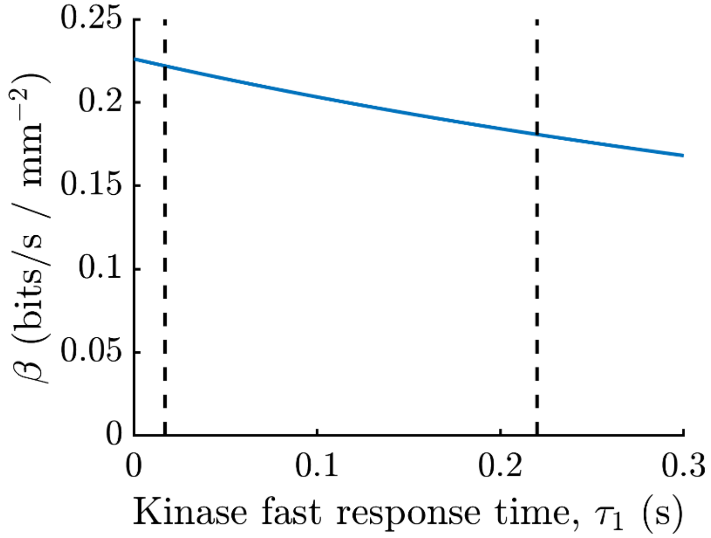


Supplementary Figure S6. Quantifying signaling noise

A) Representative time series of single-cell kinase activity a (red lines) and population average $\langle a \rangle$ (blue lines). Gray dots are raw data points and red lines are smoothed curves (10th order median filter; 10 sec time window). Green shading indicates when saturating stimuli (1 mM MeAsp and 100 μM serine) were applied, which were followed by the removal of the attractants (i.e., [MeAsp] = [serine] = 0 M) for 5 seconds before returning to the background of 100 μM MeAsp. **B)** Autocorrelation function of the kinase activity a of the cells shown in (A). Gray error bars are the average and standard error of the autocorrelation function computed from six segments of the time series with identical length. The black lines are the autocorrelation function of the Ornstein-Uhlenbeck (OU) process $C(t) = \tau_n D_n \exp(-t/\tau_n)$ with the extracted parameters using a Bayesian-filtering method (see SI). In blue at the bottom is the autocorrelation function of the population-averaged time series, with error bars. **C)** (Top) Values of extracted parameters sorted by the mean (for a_0) or MAP estimates (for τ_n, D_n, and $\sigma_n^2 = \tau_n D_n$). The error bars are the standard deviation (for a_0) or 25 and 75 percentiles of the posterior distributions (for τ_n, D_n, and $\sigma_n^2 = \tau_n D_n$), and are shown for one in every five cells for visualization purposes. (Bottom) Marginal histograms of each parameter. 95% bootstrap confidence intervals are shown by the error bars.

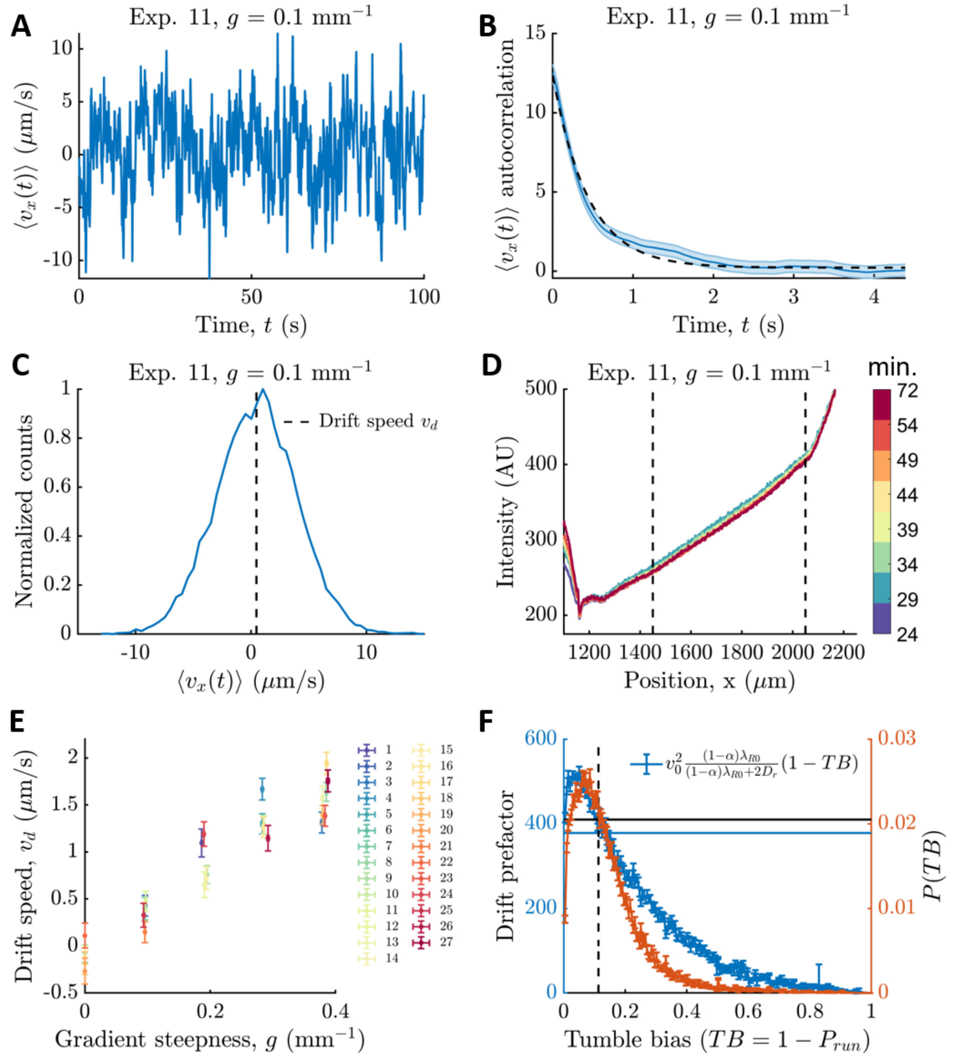


Supplementary Figure S7. Frequency domain representations of signal, response, and noise **A)** Inferred models for up-gradient velocity power spectrum $V(\omega)$ (red; units $(\mathrm{mm/s})^2$), frequency response squared $|K(\omega)|^2$ (green), and noise power spectrum $N(\omega)$ (blue) are shown (SI). Shading in all panels indicates ± one standard error (SI). **B)** Normalized frequency-space quantities plotted with linear axis scales. **C-D)** The integrand for computing β, the proportionality constant between squared gradient steepness g^2 and information rate $\dot{I}_{s\to a}$, shown in main text Fig. 2A (SI Eqn. (127)). The integrand is plotted on (C) linear-log scale $(\mathrm{bits/s})\ /\,(\mathrm{mm^{-2}})$ and (D) linear-linear scale (units $(\mathrm{bits/s})\ /\,(\mathrm{mm^{-2}\ rad/s})$). In (C), the integrand is multiplied by ω so that the area under the curve is equal to the integral.



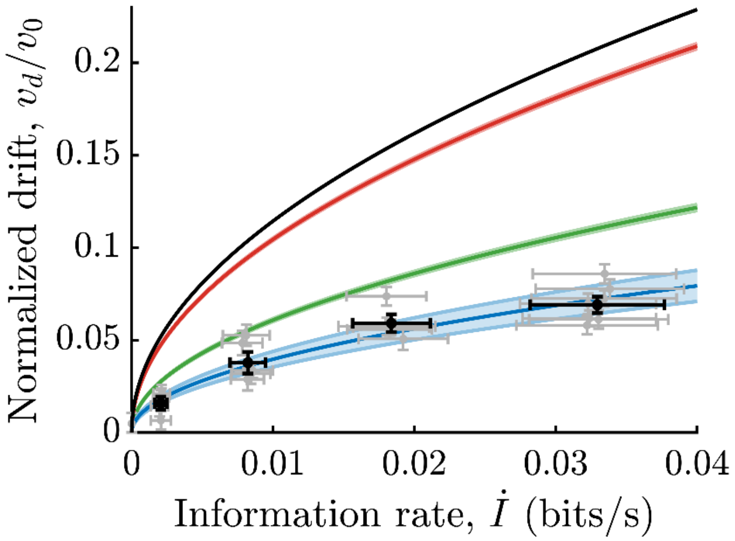
Supplementary Figure S8. Dependence of the information rate on τ_1

The information rate from signal to kinase activity $\dot{I}_{s\to a} = \beta\, g^2$ depends on the kinase fast response time τ_1 through β (see main text and SI). Plotted here is β for varying values of τ_1, showing weak, monotonically decreasing dependence. The dashed line at $\tau_1 = 0.22$ s is the kinase response time we measured, which contains the kinetics of CheY-CheZ binding and catalysis. The dashed line near $\tau_1 = 0.017 \sim 1/60$ s indicates the previously-measured kinase response time [5,6]. We used the latter value of τ_1 to calculate the information rate. This can only make cells less information efficient, achieving the same chemotactic performance with more information. Regardless, the effect on the information rate is small, and the effect on the efficiency is even smaller due to the square root dependence of the bound in Eqn. 1 the information rate.



Supplementary Figure S9. Computing average drift speeds

A) Up-gradient velocity $\langle v_x(t) \rangle$ averaged over cells in the field of view in each frame of a particular movie of cells climbing a chemical gradient. Only the first 100 s of the movie are shown. **B)** Autocorrelation function of $\langle v_x(t) \rangle$ from (A) is shown in blue (shading is standard error), and an exponential fit is shown in black. **C)** Histogram of all data points in (B) (bin size 0.5 μm/s). Drift speed v_d is the mean of the distribution. **D)** Example time course of fluorescein fluorescence intensity from a particular experiment. Color represents time since loading the cells into the device. In this experiment, a movie of swimming cells was recorded 54 minutes after loading them into the device, and a final fluorescein image was taken afterwards, 72 minutes after loading. Since this is raw fluorescence intensity, the profile deviates from a linear one outside of the region marked by black dashed lines because the depth of the device changes. Drift speed was computed from trajectories recorded inside the marked region. **E)** Drift speeds of individual, independent experiments, colored by experiment number. The same data is shown in gray in Fig. 3B of the main text. Uncertainties are standard errors. Colors in (D), (E) come from ref [7]. **F)** In theory, the population average drift speed could be different from the drift speed of the median phenotype (see SI). However, drift speed mostly depends on the combination of swimming parameters plotted here for different bins of tumble bias $TB = 1 - P_{run}$ (blue line; error bars from bootstrapping; see SI). Using this and the distribution of P_{run} (shown in orange; same as in Supplementary Fig. S1A), we compared the average value of this function to its value when evaluated at the median P_{run} (black dashed line) (see SI). We find that these two are similar: the population average gives $\left\langle v_0^2 \frac{(1-\alpha)\lambda_{R0}}{(1-\alpha)\lambda_{R0}+2\,D_r} P_{run} \right\rangle \sim 375 \pm 1 \left(\frac{\mu\mathrm{m}}{\mathrm{s}}\right)^2$ (solid blue horizontal line), whereas the value of the blue curve at the median bin P_{run} gives $v_0^2 \frac{(1-\alpha)\lambda_{R0}}{(1-\alpha)\lambda_{R0}+2\,D_r} P_{run} \sim 410 \pm 3 \left(\frac{\mu\mathrm{m}}{\mathrm{s}}\right)^2$ (solid black horizontal line). This justifies our comparison of population-average drift speeds to bounds quantified using a median cell's parameters.



Supplementary Figure S10. Information bounds on performance

The upper bound on chemotactic drift speed set by information acquisition depends on the cell's behavioral response, $\lambda_R(\{s\})$, but also on its behavioral parameters, $\boldsymbol{\theta} = \{\lambda_{R0}, \alpha, P_{run}\}$ (alternative parameterization from the main text). The green line and blue lines/data points are reproduced from main text Fig. 3C. The green line is the maximum drift speed possible for a cell that has the same behavioral parameters as those we measure in RP437 *E. coli*. This bound was found by optimizing over responses $\lambda_R(\{s\})$. In principle, a cell with different behavioral parameters has a different bound on its drift speed. Optimizing the mean tumble rate λ_{R0}, but keeping the remaining behavioral parameters held fixed at their measured values, gives the red curve. Finally, if the remaining parameters are optimized, this gives the black curve. No cell's drift speed can exceed this bound. See SI for derivations and expressions for each bound. Shading and error bars indicate ± one standard error.

Parameter	Value	Meaning	Source
P_{run}	0.89 ± 0.01	Median run bias	This study
λ_{R0}	$0.893 \pm 0.006\ \mathrm{s}^{-1}$	Baseline tumble rate	This study
α	0.06 ± 0.01	Directional persistence of tumbles	This study
v_0	$22.61 \pm 0.07\ \mu\mathrm{m/s}$	Speed during runs	This study
D_r	$0.0441 \pm 10^{-4}\ \mathrm{rad}^2/\mathrm{s}$	Rotational diffusion coefficient	This study
a_v	$157.1 \pm 0.5\ \left(\frac{\mu\mathrm{m}}{\mathrm{s}}\right)^2$	Fit prefactor of the x-velocity autocorrelation function $V(t)$	This study
λ_{tot}	$0.862 \pm 0.005\ \mathrm{s}^{-1}$	Fit decay rate of the x-velocity autocorrelation function $V(t)$	This study
G	1.73 ± 0.03	Gain of kinase output	This study
τ_2	$9.9 \pm 0.3\ \mathrm{s}$	Adaptation time	This study
τ_1	$0.22 \pm 0.01\ \mathrm{s}$	Time for kinase response and CheY/CheZ binding	This study
τ_1	$0.017 \pm 0.004\ \mathrm{s}$ $\sim 1/60\ s$	Kinase response time (used to compute the information rate $\dot{I}_{s\to a}$)	Refs [5,6]
D_n	$(7.2 \pm 0.3) \times 10^{-4}\ \mathrm{s}^{-1}$	Diffusivity of kinase activity fluctuations	This study
τ_n	$11.75 \pm 0.04\ \mathrm{s}$	Correlation time of kinase activity fluctuations	This study
σ_n	0.092 ± 0.002	Standard deviation of kinase activity fluctuations	This study
a_0	0.29 ± 0.07; 0.30 ± 0.08	Baseline kinase activity	This study
β	$0.225 \pm 0.032\ \mathrm{bits/s}\ /\ \mathrm{mm}^{-2}$	*E. coli*'s information rate from signal to kinase activity per squared gradient steepness, $\dot{I}_{s\to a} = \beta\, g^2$	This study
χ	$4300 \pm 150\ \mu\mathrm{m}^2/\mathrm{s}$	Drift speed per unit gradient steepness (chemotactic coefficient)	This study

Supplementary Table S1. Parameter values.

Implicit in the units of parameters G and σ_n are the dimensionless units of kinase activity. Uncertainties for a_0 are its standard deviation over an isogenic population. The first value comes from the linear response experiments, and the second one from the noise measurement experiments.

Supplementary Information Text

Section 1: Relationship between mutual information and transfer entropies

Consider two time-dependent random variables $X(t)$ and $Y(t)$, and let time be discrete with $X(k) = X(t_k) = X(k\,\Delta t)$. The set of values of X from time $t = 0$ to time $t = t_{k-1}$ (inclusive) will be denoted X_0^{k-1}. Then, the mutual information between $X(t)$ and $Y(t)$ from $t = 0$ to $t = t_N$ is:

$$MI(X_0^N; Y_0^N) = \int dX_0^N\, dY_0^N\, P(X_0^N, Y_0^N) \log\left(\frac{P(X_0^N, Y_0^N)}{P(X_0^N)\, P(Y_0^N)}\right) \tag{1}$$

where the probability distributions and integral are over trajectories. Each of the probability distributions can be decomposed into a product of conditional distributions, chosen such that the probability of each X and Y is conditioned only on those at earlier times:

$$P(X_0^N, Y_0^N) = \prod_{k=0}^{N} P\big(X(k), Y(k)\big|X_0^{k-1}, Y_0^{k-1}\big) \tag{2}$$

where the $k = 0$ term is just $P\big(X(0), Y(0)\big)$. Doing this for each of the distributions in Eqn. (1) gives:

$$MI(X_0^N; Y_0^N) = \int dX_0^N\, dY_0^N\, P(X_0^N, Y_0^N) \log\left(\frac{\prod_{k=0}^{N} P(X(k), Y(k)\big|X_0^{k-1}, Y_0^{k-1})}{\prod_{k'=0}^{N} P\big(X(k')\big|X_0^{k'-1}\big)\ \prod_{k''=0}^{N} P\big(Y(k'')\big|Y_0^{k''-1}\big)}\right). \tag{3}$$

Writing the log of the products as a sum of the logs and grouping probabilities of X and Y at the same time step:

$$MI(X_0^N; Y_0^N) = \int dX_0^N\, dY_0^N\, P(X_0^N, Y_0^N) \sum_{k=0}^{N} \log\left(\frac{P\big(X(k), Y(k)|X_0^{k-1}, Y_0^{k-1}\big)}{P\big(X(k)|X_0^{k-1}\big)\, P\big(Y(k)|Y_0^{k-1}\big)}\right) \tag{4}$$

Moving the sum outside of the integral, the k'th term in the sum only depends on X and Y at earlier times; therefore $P(X_0^N, Y_0^N)$ can be marginalized for X and Y at all times larger than t_k:

$$MI(X_0^N; Y_0^N) = \sum_{k=0}^{N} \int dX_0^k\, dY_0^k\, P\big(X_0^k, Y_0^k\big) \log\left(\frac{P\big(X(k), Y(k)|X_0^{k-1}, Y_0^{k-1}\big)}{P\big(X(k)|X_0^{k-1}\big)\, P\big(Y(k)|Y_0^{k-1}\big)}\right) \tag{5}$$

Multiplying the numerator and denominator inside the log by $P\big(X(k)\big|X_0^{k-1}, Y_0^{k-1}\big)$ and $P\big(Y(k)\big|X_0^{k-1}, Y_0^{k-1}\big)$ and then separating the log of the product into a sum of logs gives:

$$
\begin{aligned}
&MI(X_0^N;Y_0^N) \\
&= \sum_{k=0}^{N} \int dX_0^k\, dY_0^k\, P\big(X_0^k, Y_0^k\big) \log\left(\frac{P\big(X(k)\big|X_0^{k-1}, Y_0^{k-1}\big)}{P\big(X(k)|X_0^{k-1}\big)}\right) \\
&+ \sum_{k=0}^{N} \int dX_0^k\, dY_0^k\, P\big(X_0^k, Y_0^k\big) \log\left(\frac{P\big(Y(k)\big|X_0^{k-1}, Y_0^{k-1}\big)}{P\big(Y(k)|Y_0^{k-1}\big)}\right) \\
&+ \sum_{k=0}^{N} \int dX_0^k\, dY_0^k\, P\big(X_0^k, Y_0^k\big) \log\left(\frac{P\big(X(k), Y(k)|X_0^{k-1}, Y_0^{k-1}\big)}{P\big(X(k)|X_0^{k-1}, Y_0^{k-1}\big)\, P\big(Y(k)|X_0^{k-1}Y_0^{k-1}\big)}\right)
\end{aligned} \tag{6}
$$

$$
= I_{Y\to X} + I_{X\to Y} + \sum_{k=0}^{k=N} I\big(X(k); Y(k)|X_0^{k-1}, Y_0^{k-1}\big) \tag{7}
$$

$$
= I_{Y\to X} + I_{X\to Y} \tag{8}
$$

The first term is the transfer entropy from Y to X, and the second is the transfer entropy from X to Y. The last term measures the conditional dependence of one variable at the current time on another at the current time, given their histories, and is zero for a stationary, causal, finite-order Markov system.

We use “transfer entropy” to refer to the cumulative amount of statistical influence of one variable onto another over some period of time. This is in contrast to Schreiber [8], who used transfer entropy to refer to a rate—Schreiber’s “transfer entropy” is our “transfer entropy rate”.

The steady state mutual information rate and transfer entropy rates are obtained by dividing by the total time $t_N = N\,\Delta t$ and taking N to infinity.

$$
\dot{I}_{X\to Y} = \lim_{N\to\infty} \frac{1}{N\,\Delta t} \sum_{k=0}^{N} \int dX_0^k\, dY_0^k\, P\big(X_0^k, Y_0^k\big) \log\left(\frac{P\big(Y(k)\big|X_0^{k-1}, Y_0^{k-1}\big)}{P\big(Y(k)|Y_0^{k-1}\big)}\right) \tag{9}
$$

For a stationary process, after sufficient time has passed (sufficiently large N), each term in the sum is equivalent, and we can reassign the indices so that $k = 0$ is the current time, $k = 1$ is the next future time step, and negative indices indicate time in the past:

$$
\dot{I}_{X\to Y} = \lim_{N\to\infty} \frac{1}{\Delta t} \int dX_{-N}^1\, dY_{-N}^1\, P(X_{-N}^1, Y_{-N}^1) \log\left(\frac{P(Y(k=1)|X_{-N}^0, Y_{-N}^0)}{P(Y(k=1)|Y_{-N}^0)}\right) \tag{10}
$$

Taking $\Delta t \to dt$, the continuous time limit, $k = 0$ is replaced with the current time t, $k = 1$ becomes time $t + dt$, and we get

$$
\dot{I}_{X\to Y} = \frac{1}{dt} \int d\{X(t+dt)\}\, d\{Y(t+dt)\}\, P(\{X(t+dt)\}, \{Y(t+dt)\}) \log\left(\frac{P(Y(t+dt)|\{X(t)\}, \{Y(t)\})}{P(Y(t+dt)|\{Y(t)\})}\right) \tag{11}
$$

$$\dot{I}_{X\to Y} = \frac{1}{dt}\left\langle D_{KL}\big(P(Y(t+dt)|\{X(t)\},\{Y(t)\})||P(Y(t+dt)|\{Y(t)\})\big)\right\rangle_{\{X(t)\},\ \{Y(t)\}} \tag{12}$$

where $\{X(t)\}$ is the full history of $X(t)$ up to time t (inclusive), and D_{KL} is the Kullback-Leibler divergence.

Section 2: Drift speed in the regime of small chemotactic bias

The first step in connecting information transfer to chemotactic performance is to construct a model of chemotaxis and derive expressions for the two quantities. Here we draw on past work [9,10] to derive the drift speed of a chemotactic *E. coli* cell in shallow, static gradients. In this section, rather than re-derive that result, we will cast it in a form that will be convenient later on, when we derive the upper bound that information transfer places on the maximum drift speed a cell can achieve.

We model chemotaxis, like others have done before [9–13], as follows. During runs, the cell swims with speed v_0 and is subject to rotational diffusion with coefficient D_r. Many experiments to date [14–18] are consistent with a Monod-Wyman-Changeux (MWC) model [19] of cells' transmembrane receptor activity. In this description, cells sense concentration through its effects on the free energy difference between the inactive and active states of the receptors, $f_c = \log\left(\frac{1+c(t)/K_i}{1+c(t)/K_a}\right)$. Here, K_i and K_a are the receptor dissociation constants for the ligand when the receptors are in the inactive or active state [20,21], and cells are log-sensing over a wide range of concentrations $K_i \ll c(t) \ll K_a$[16,22]. In a static, shallow, exponential gradient, we have $c(x) = c_0 e^{g\,x}$, where c_0 is the background concentration and g is the gradient steepness. Cells respond to time changes of concentration through time changes of receptor state $\frac{d}{dt}f_c$, which in the log-sensing regime is $\sim \frac{d}{dt}\log c(t) = v_x(t)\frac{d}{dx}\log\big(c(x(t))\big) = g\,v_0\cos(\theta(t))$. Here, $\theta(t)$ is the angle between the cell's heading at time t and the direction of the concentration gradient (the x-axis). Therefore, we define the signal to be $s(t) = \frac{d}{dt}\log c(t)$. In shallow gradients (small g), we could just as well define the signal as $s(t) = \frac{dc}{dt}$, lumping c_0 into the cell's response to the signal. Since signal and response always come as a pair, this distinction does not affect our results.

The cell's swimming state $m(t)$ can be either run (R) or tumble (T). We assume that the cell has zero speed during tumbles [23]. Previous studies have shown that slow fluctuations in signaling activity drive long tails in the counterclockwise rotation durations of single motors [24–27]. Here and in our experiments, we only consider shallow gradients, where these slow fluctuations are predicted to increase drift speed by at most 10% [28,29]. For simplicity, here we model run-tumble transitions as Poisson processes with rates that do not fluctuate in the absence of signal. As a result, the tumble rate should be understood as an effective tumble rate that lumps together the effects of ligand arrival noise, noisy internal state, and multiple motors. This model fits well our measurements of cells' velocity autocorrelation functions, which on the time scale we are able to measure exhibits an exponential decay (Supplementary Fig. S2A; see also refs [1,23]). The run-to-tumble transition occurs with rate

$$\lambda_R(t) = \lambda_{R0}\left(1 - \epsilon(\{s(t)\})\right), \tag{13}$$

depending on the history of signal seen, $\{s(t)\}$. Curly brackets { }, such as $\{s(t)\}$, indicate a trajectory of a quantity up to and including the indicated time (in this case, t). We define λ_{R0} as the average tumble rate in the absence of a gradient. Tumble to run transitions occur with constant rate λ_T, which is taken to be independent of $s(t)$, valid in shallow gradients [9,23]. As a result of the Poisson assumption, run and tumble transitions have no memory of how long the cell has been in its current state.

Tumbles can partially reorient the cell [23,30], quantified by $\alpha = \langle\cos(\gamma)\rangle$, where γ is the angle between the cell's swimming direction before and after the tumble. When $\alpha = 0$, tumbles fully reorient the cell. We also assume that the cell is navigating a shallow gradient, therefore the tumble rate modulation is small, $\epsilon(\{s(t)\}) \ll 1$.

In general, $\epsilon(\{s(t)\})$ might be a complicated function of the past signals, but in shallow gradients, only a linear response approximation is needed to calculate the cell's drift speed. Therefore, we consider responses of the form

$$\epsilon(\{s(t)\}) = \int_{-\infty}^{t} K_b(t - t')\, s(t')\, dt', \tag{14}$$

with kernel $K_b(T)$ mapping signal to behavior.

By integrating by parts, one can show that responses to $s(t)$ with kernel $K_b(T)$ are equivalent responses to (log) concentration with a different kernel:

$$\epsilon(\{c(t)\}) = \int_{-\infty}^{t} R_b(t - t')\left(\log\left(c(t')\right) - \log(c_0)\right) dt', \tag{15}$$

where c_0 is the background concentration. These are equivalent for $s(t) = \frac{d}{dt}\log\left(c(t)\right)$, $R_b(T) = K_b(0)\,\delta(T) + \frac{dK_b(T)}{dT}$, and $K_b(T) = \int_0^T R_b(T')\,dT'$. In these formulations, perfect adaptation [31,32] can be encoded by $\int_0^\infty R_b(T)\,dT = 0$ or $K_b(\infty) = 0$. See **Section 7: Imperfect Adaptation** for derivations.

Locsei [9] and Celani and Vergassola [10] showed that the drift speed is:

$$v_d = \frac{v_0}{3}\frac{(1-\alpha)\lambda_{R0}}{(1-\alpha)\lambda_{R0} + 2\,D_r} P_{run}\, \frac{v_0\, g}{(1-\alpha)\lambda_{R0} + 2\,D_r} \int_0^\infty R_b(T)\, e^{-((1-\alpha)\lambda_{R0} + 2\,D_r)T}\, dT. \tag{16}$$

$P_{run} = P(m(t) = R)$ is the fraction of time the cell spends in the run state, or $P_{run} = 1 - TB$, with tumble bias $TB = \frac{\lambda_{R0}}{\lambda_{R0} + \lambda_T}$. Plugging in the relationship between $R_b(T)$ and $K_b(T)$ above, and integrating by parts, one finds:

$$v_d = \frac{v_0}{3} \frac{(1-\alpha)\lambda_{R0}}{(1-\alpha)\lambda_{R0} + 2\, D_r} P_{run}\, v_0\, g \int_0^\infty K_b(T)\, e^{-\left((1-\alpha)\lambda_{R0} + 2\, D_r\right)T}\, dT. \tag{17}$$

This expression applies whether or not the response kernel adapts perfectly [10,12] (see **Section 7: Imperfect Adaptation**).

Below, we will show that in the drift speed expression:

$$\left\langle \epsilon(\{s(t)\}) \cos\big(\theta(t)\big) \right\rangle = \frac{1}{3} v_0\, g \int_0^\infty K_b(T)\, e^{-\left((1-\alpha)\lambda_{R0} + 2\, D_r\right)T}\, dT, \tag{18}$$

where the angled brackets are an average over trajectories, conditioned on the cell being in the run state at the current time t, $m(t) = R$. Writing the drift speed in this form,

$$v_d = v_0 \frac{(1-\alpha)\lambda_{R0}}{(1-\alpha)\lambda_{R0} + 2\, D_r} P_{run} \left\langle \epsilon(\{s(t)\}) \cos\big(\theta(t)\big) \right\rangle, \tag{19}$$

will simplify the derivations later when we upper bound the cell's drift speed at fixed information rate from signal to behavior.

To show this, we start from the definition of $\epsilon(\{s(t)\})$ in Eqn. (14) and split the average over trajectories into two expectations. The first is an average over past trajectories given that they end with signal $s(t)$, and the second is an average over current signals $s(t)$. The first expectation is:

$$\begin{aligned} &\left\langle \epsilon(\{s(t)\}) \cos\big(\theta(t)\big) \,\middle|\, s(t), m(t) = R \right\rangle \\ &\qquad = \left\langle \cos\big(\theta(t)\big) \int_{-\infty}^{t} K_b(t-t')\, s(t')\, dt' \,\middle|\, s(t), m(t) = R \right\rangle \end{aligned} \tag{20}$$

Given $s(t) = v_0\, g \cos\big(\theta(t)\big)$, only $\theta(t')$ is a random variable in this expression. Similarly, conditioning on $s(t)$ is the same as conditioning on $\cos\big(\theta(t)\big)$, so we can simplify to:

$$= g\, v_0 \cos\big(\theta(t)\big) \int_{-\infty}^{t} K_b(t-t') \left\langle \cos\big(\theta(t')\big) \,\middle|\, \cos\big(\theta(t)\big), m(t) = R \right\rangle dt'. \tag{21}$$

The heading at time t' in the past is $\cos\big(\theta(t')\big) = \cos\big(\theta(t') - \theta(t) + \theta(t)\big) = \cos\big(\theta(t') - \theta(t)\big) \cos\big(\theta(t)\big) - \sin\big(\theta(t') - \theta(t)\big) \sin\big(\theta(t)\big)$, using the angle sum formula. The product of sine terms will equal zero in the end, so we drop them here. Plugging this in:

$$\begin{aligned} &\left\langle \epsilon(\{s(t)\}) \cos\big(\theta(t)\big) \,\middle|\, s(t), m(t) = R \right\rangle \\ &\qquad = g\, v_0 \cos^2\big(\theta(t)\big) \int_{-\infty}^{t} K_b\big(t - t^{'}\big) \left\langle \cos\left(\theta\big(t^{'}\big) - \theta(t)\right) \,\middle|\, \cos\big(\theta(t)\big), m(t) = R \right\rangle\, dt^{'}. \end{aligned} \tag{22}$$

Since we assume the gradient is shallow, we will only keep terms up to first order in g. Therefore, $\langle\cos\big(\theta(t') - \theta(t)\big)\rangle$ is an average over swimming trajectories that experience rotational diffusion and tumbles but that don't respond to the gradient. Rotational diffusion is symmetric around the initial heading, which is why we excluded the term $\langle\sin\big(\theta(t') - \theta(t)\big)\rangle = 0$ above. Since $\langle\cos\big(\theta(t') - \theta(t)\big)\rangle = e^{-\left((1-\alpha)\lambda_{R0}+2\,D_r\right)(t-t')}$ for $t' < t$[30], we now have:

$$= g\, v_0\, \cos^2\big(\theta(t)\big)\int_{-\infty}^{t} K_b(t-t')\, e^{-\left((1-\alpha)\lambda_{R0}+2\,D_r\right)(t-t')}\, dt'. \tag{23}$$

Since the system is stationary and we conditioned on the state at time t, we can change variables to integrate over time into the past T, back from the current time t: $T = t - t'$ and $dT = -dt'$. This gives:

$$\begin{aligned}&\langle\epsilon(\{s(t)\})\cos\big(\theta(t)\big)\,\big|s(t), m(t) = R\rangle \\ &\qquad = g\, v_0\, \cos^2\big(\theta(t)\big)\int_{0}^{\infty} K_b(T)\, e^{-\left((1-\alpha)\lambda_{R0}+2\,D_r\right)T}\, dT.\end{aligned} \tag{24}$$

Then, taking the second expectation, over current signals $s(t)$, or current headings $\cos\big(\theta(t)\big)$ (with the conditioning on $m(t) = R$ implied):

$$\langle\epsilon(\{s(t)\})\cos\big(\theta(t)\big)\rangle = g\, v_0\, \langle\cos^2\big(\theta(t)\big)\rangle\int_{0}^{\infty} K_b(T)\, e^{-\left((1-\alpha)\lambda_{R0}+2\,D_r\right)T}\, dT. \tag{25}$$

Again, since we assume the gradient is shallow and only keep terms to first order in g, the expectation $\langle\cos^2\big(\theta(t)\big)\rangle$ is with respect to a uniform distribution of headings in three-dimensional space, giving:

$$\begin{aligned}&\langle\epsilon(\{s(t)\})\cos\big(\theta(t)\big)\rangle \\ &= g\, v_0\, \frac{1}{4\pi}\int_{0}^{2\pi}\int_{0}^{\pi}\cos^2(\theta)\sin(\theta)\, d\theta\;\, d\phi\int_{0}^{\infty} K_b(T)\, e^{-\left((1-\alpha)\lambda_{R0}+2\,D_r\right)T}\, dT\end{aligned} \tag{26}$$

or:

$$\langle\epsilon(\{s(t)\})\cos\big(\theta(t)\big)\rangle = \frac{1}{3} g\, v_0\int_{0}^{\infty} K_b(T)\, e^{-\left((1-\alpha)\lambda_{R0}+2\,D_r\right)T}\, dT. \tag{27}$$

Section 3: Information rate from signal to behavior

Although the information available to the cell to navigate is the transfer entropy from signal to kinase activity (discussed later), chemotaxis performance depends on the properties of the motor responses. Therefore, to connect information acquisition at the kinases to performance, we will need an expression for the rate of information transfer to behavior. The information acquired by the kinases upper bounds the information communicated to behavior, which in turn upper bounds chemotactic performance. Thus, information acquired by the kinases upper bounds performance.

To derive an expression for the information communicated to behavior, or the information rate for short in this section, we will use the differential expression for the steady state transfer entropy rate:

$$\dot{I}_{s\to m} = \frac{1}{dt}\left\langle D_{KL}\big(P(m(t+dt)|\{m(t)\},\{s(t)\})||P(m(t+dt)|\{m(t)\})\big)\right\rangle_{\{m(t)\},\{s(t)\}} \tag{28}$$

$$= \frac{1}{dt}\left\langle \log\left(\frac{P(m(t+dt)|\{m(t)\},\{s(t)\})}{P(m(t+dt)|\{m(t)\})}\right)\right\rangle_{\{m(t)\},\{s(t)\},m(t+dt)}. \tag{29}$$

Again, the cell's swimming state $m(t)$ can be either run (R) or tumble (T), $s(t)$ is the signal the cell perceives, and curly brackets $\{\ \}$, such as $\{s(t)\}$ indicate a trajectory of a quantity up to and including time t.

Deriving the information rate for a chemotactic *E. coli* requires deriving the transition probabilities above. Each of the transition probabilities can be written with the trajectories $\{m(t)\}$ and $\{s(t)\}$ separated by the current time and the past: $\{m(t)\} = (m(t), \{m(t-dt)\})$ and $\{s(t)\} = (s(t), \{s(t-dt)\})$. Furthermore, since we assumed in the previous section that transitions between runs and tumble are inhomogeneous Poisson processes given the trajectory $\{s(t)\}$, the transition probabilities are independent of how long the cell has been in its current state. Together, these give: $P(m(t+dt)|\{m(t)\},\{s(t)\}) = P(m(t+dt)|m(t), s(t), \{s(t-dt)\})$. In particular, the probability that a running cell transitions to tumbling in the next dt is $P(m(t+dt) = T|m(t) = R, \{s(t)\}) = \lambda_R(\{s(t)\})\, dt$ and the probability that a tumbling cell transitions to running in dt is $P(m(t+dt) = R|m(t) = T, \{s(t)\}) = \lambda_T\, dt$.

We also need the conditional probability $P(m(t+dt)|\{m(t)\})$. When the cell is in the tumble state at time t, $P(m(t+dt) = R|m(t) = T, \{m(t-dt)\})$, reduces to $P(m(t+dt) = R|m(t) = T) = \lambda_T\, dt$, since tumble-to-run transitions are independent of the past and of the signal. When the cell is the run state, the transition probability $P(m(t+dt) = T|m(t) = R, \{m(t-dt)\})$ is more complicated. This quantity measures the inferred probability that the cell will tumble in dt, given the past of motor states $\{m(t)\}$. The past motor states are informative of whether the cell will tumble only because they allow inference of the past signals $\{s(t)\}$. Writing this out mathematically:

$$P(m(t+dt) = T|m(t) = R, \{m(t-dt)\}) \tag{30}$$

$$= \int P(m(t+dt) = T|\{s(t)\}, m(t) = R, \{m(t-dt)\})\, P(\{s(t)\}|m(t) = R, \{m(t-dt)\})\, d\{s(t)\} \tag{31}$$

$$= \int P(m(t+dt) = T|\{s(t)\}, m(t) = R)\, P(\{s(t)\}|m(t) = R, \{m(t-dt)\})\, d\{s(t)\} \tag{32}$$

$$= \int \lambda_R(\{s(t)\})\, dt\, P(\{s(t)\}|m(t) = R, \{m(t-dt)\})\, d\{s(t)\}. \tag{33}$$

$P(\{s(t)\}|m(t) = R, \{m(t-dt)\})$ is the distribution of past signal trajectories that can be inferred from knowledge that the cell is currently running and from its past behavior. Since the behavioral response $\epsilon(\{s(t)\})$ is small and behavior only depends on the history of signal through it, $P(\{s(t)\}|m(t) = R, \{m(t-dt)\})$ can be written as an asymptotic series solution in $\epsilon(\{s(t)\})$:

$$P(\{s(t)\}|m(t) = R, \{m(t-dt)\}) \sim P_0(\{s(t)\}|m(t) \\ = R)\big(1 + O\big(\epsilon(\{s(t)\})\big) + O\big(\epsilon^2(\{s(t)\})\big) + \cdots\big), \tag{34}$$

where $P_0(\{s(t)\}|m(t) = R)$ is the distribution of signal trajectories when the cell does not respond to the signal. This is a probability distribution, so it must integrate to one. Since the zeroth order term integrates to one, all higher order terms must either integrate to zero or equal zero. Since terms with $\epsilon(\{s(t)\})$ raised to an even power are nonnegative, their integrals can only equal zero if they are zero. Therefore, we have:

$$P(\{s(t)\}|m(t) = R, \{m(t-dt)\}) \sim P_0(\{s(t)\}|m(t) \\ = R)\big(1 + O\big(\epsilon(\{s(t)\})\big) + \cdots\big). \tag{35}$$

We assumed that λ_{R0} is the average tumble rate in the absence of a gradient, or:

$$\lambda_{R0} = \int P_0(\{s(t)\}|m(t) = R)\ \lambda_R(\{s(t)\})\ d\{s(t)\} \tag{36}$$

$$= \lambda_{R0} \int P_0(\{s(t)\}|m(t) = R)\ \big(1 - \epsilon(\{s(t)\})\big)\ d\{s(t)\} \tag{37}$$

$$= \lambda_{R0}\ (1 - \langle\epsilon(\{s(t)\})\rangle) \tag{38}$$

$$\rightarrow \quad \langle\epsilon(\{s(t)\})\rangle = 0 \tag{39}$$

Therefore, $\big\langle O\big(\epsilon(\{s(t)\})\big)\big\rangle = 0$. Below, angled brackets without a subscript will denote an average with respect to $P_0(\{s(t)\}|m(t) = R)$, which is the same usage as in the derivation of the drift speed in the section above.

Using the expression for $P(\{s(t)\}|m(t) = R, \{m(t-dt)\})$ above, we get:

$$P(m(t+dt) = T|m(t) = R, \{m(t-dt)\}) = \\ dt \int \lambda_{R0}\big(1 - \epsilon(\{s(t)\})\big)\ P_0(\{s(t)\}|m(t) = R)\big(1 + O\big(\epsilon(\{s(t)\})\big) + \cdots\big)\ d\{s(t)\}. \tag{40}$$

Keeping terms up to order $\epsilon^2(\{s(t)\})$ (and dropping (t) for clarity):

$$= \lambda_{R0}\, dt \int P_0(\{s(t)\}|m(t) \\ = R)\left(1 - \epsilon(\{s\}) + O\big(\epsilon(\{s\})\big) - O\big(\epsilon(\{s\})\big)^2\right) d\{s(t)\}. \tag{41}$$

Using $\langle \epsilon(\{s(t)\})\rangle = 0$, this is:

$$= \lambda_{R0}\, dt\, \left(1 - \left\langle O\big(\epsilon(\{s\})\big)^2 \right\rangle\right) = \lambda_{R0}\, dt\, (1 - A), \tag{42}$$

where we define $A = \left\langle O\big(\epsilon(\{s\})\big)^2 \right\rangle > 0$ as the expectation of the first-order terms squared that come from $P(\{s(t)\}|m(t) = R, \{m(t-dt)\})$. This definition will be useful because A will ultimately drop out in the final expression for $\dot{I}_{s\to m}$. To understand the expression for $P(\{s(t)\}|m(t) = R, \{m(t-dt)\})$ above, observing that the cell has been running for any finite amount of time increases the likelihood that it has been running up the gradient. The amount by which that likelihood increases is order $\epsilon(\{s(t)\})$. Then, since it is more likely going up-gradient than down-gradient, the likelihood that it will tumble in dt should be strictly lower than average. Predicting whether the cell will tumble in dt produces another factor of $\epsilon(\{s(t)\})$ in A from the definition of $\lambda_R(t)$.

With these transition probabilities in hand, we now derive the transfer entropy rate. As a reminder, the information rate is:

$$\dot{I}_{s\to m} = \frac{1}{dt}\left\langle \log\left(\frac{P(m(t+dt)|\{m(t)\},\{s(t)\})}{P(m(t+dt)|\{m(t)\})}\right)\right\rangle_{\{m(t)\},\{s(t)\},m(t+dt)}. \tag{43}$$

Writing out the expectation with respect to $m(t)$:

$$= \frac{1}{dt}\left(P_{run}\left\langle \log\left(\frac{P(m(t+dt)|m(t)=R,\{m(t-dt)\},\{s(t)\})}{P(m(t+dt)|m(t)=R,\{m(t-dt)\})}\right)\right\rangle_{\{s(t)\},m(t+dt),\{m(t-dt)\}|m(t)=R} \\ + P_{tumble}\left\langle \log\left(\frac{P(m(t+dt)|m(t)=T,\{m(t-dt)\},\{s(t)\})}{P(m(t+dt)|m(t)=T,\{m(t-dt)\})}\right)\right\rangle_{\{s(t)\},m(t+dt),\{m(t-dt)\}|m(t)=T}\right), \tag{44}$$

where $P_{tumble} = 1 - P_{run}$. Since tumble durations are assumed to not depend on past behaviors or on the signal, the second term in the parentheses is zero. We also simplify the numerator in the remaining logarithm:

$$\dot{I}_{s\to m} \\ = \frac{1}{dt}P_{run}\left\langle \log\left(\frac{P(m(t+dt)|m(t)=R,\{s(t)\})}{P(m(t+dt)|m(t)=R,\{m(t-dt)\})}\right)\right\rangle_{\{s(t)\},m(t+dt),\{m(t-dt)\}|m(t)=R}. \tag{45}$$

Writing out the expectation with respect to $m(t+dt)$,

$$
\begin{aligned}
&= \frac{1}{\mathrm{dt}} P_{run} \Bigg\langle P(m(t+dt)=T|m(t)=R,\{s(t)\}) \log\left(\frac{P(m(t+dt)=T|m(t)=R,\{s(t)\})}{P(m(t+dt)=T|m(t)=R,\{m(t-dt)\})}\right) \\
&+ P(m(t+dt)=R|m(t)=R,\{s(t)\}) \log\left(\frac{P(m(t+dt)=R|m(t)=R,\{s(t)\})}{P(m(t+dt)=R|m(t)=R,\{m(t-dt)\})}\right)\Bigg\rangle_{\{s(t)\},\{m(t-dt)\}|m(t)=R} .
\end{aligned}
\tag{46}
$$

Plugging in the transition probabilities derived above:

$$
\begin{aligned}
&= \frac{1}{dt} P_{run} \Bigg\langle \lambda_{R0}\big(1-\epsilon(\{s(t)\})\big)dt \, \log\left(\frac{\lambda_{R0}\big(1-\epsilon(\{s(t)\})\big)\,dt}{\lambda_{R0}\,dt\,(1-A)}\right) \\
&\qquad + \big(1-\lambda_{R0}\big(1-\epsilon(\{s(t)\})\big)dt\big)\log\left(\frac{\big(1-\lambda_{R0}\big(1-\epsilon(\{s(t)\})\big)\,dt\big)}{\big(1-\lambda_{R0}\,dt\,(1-A)\big)}\right)\Bigg\rangle_{\{s(t)\},\{m(t-dt)\}|m(t)=R} .
\end{aligned}
\tag{47}
$$

Next, expanding the second logarithm to first order in dt:

$$
\begin{aligned}
&= \frac{1}{dt} P_{run} \Bigg\langle \lambda_{R0}\,dt\,\big(1-\epsilon(\{s(t)\})\big)\,\log\left(\frac{1-\epsilon(\{s(t)\})}{1-A}\right) \\
&\qquad + \Big(1-\lambda_{R0}\,dt\,\big(1-\epsilon(\{s(t)\})\big)\Big)\Big(-\lambda_{R0}\,dt\,\big(1-\epsilon(\{s(t)\})\big) \\
&\qquad + \lambda_{R0}\,dt\,(1-A)\Big)\Bigg\rangle_{\{s(t)\},\{m(t-dt)\}|m(t)=R}
\end{aligned}
\tag{48}
$$

$$
\begin{aligned}
&= \frac{1}{dt} P_{run} \Bigg\langle \lambda_{R0}\,dt\,\big(1-\epsilon(\{s(t)\})\big)\,\log\left(\frac{1-\epsilon(\{s(t)\})}{1-A}\right) \\
&\qquad + \Big(-\lambda_{R0}\,dt\,\big(1-\epsilon(\{s(t)\})\big)+\lambda_{R0}\,dt\,(1-A)\Big)\Bigg\rangle_{\{s(t)\},\{m(t-dt)\}|m(t)=R}
\end{aligned}
\tag{49}
$$

$$
= \lambda_{R0}\,P_{run}\left\langle \big(1-\epsilon(\{s(t)\})\big)\,\log\left(\frac{1-\epsilon(\{s(t)\})}{1-A}\right)+\epsilon(\{s(t)\})-A\right\rangle_{\{s(t)\},\{m(t-dt)\}|m(t)=R}
\tag{50}
$$

For shallow gradients and small $\epsilon(\{s(t)\})$, we can expand the remaining logarithm to second order in $\epsilon(\{s(t)\})$. Recalling that A is already second order in $\epsilon(\{s(t)\})$, we get:

$$
\begin{aligned}
&= \lambda_{R0}\,P_{run}\,\Big\langle\big(1-\epsilon(\{s(t)\})\big)\Big(-\epsilon(\{s(t)\})-\frac{1}{2}\epsilon^2(\{s(t)\})+A\Big)+\epsilon(\{s(t)\}) \\
&\qquad - A\Big\rangle_{\{s(t)\},\{m(t-dt)\}|m(t)=R} .
\end{aligned}
\tag{51}
$$

Keeping to second order in $\epsilon(\{s(t)\})$:

$$
= \lambda_{R0}\,P_{run}\,\Big\langle -\epsilon(\{s(t)\})-\frac{1}{2}\epsilon^2(\{s(t)\})+A+\epsilon^2(\{s(t)\})+\epsilon(\{s(t)\})-A\Big\rangle_{\{s(t)\},\{m(t-dt)\}|m(t)=R}
\tag{52}
$$

$$
= \frac{1}{2}\lambda_{R0}\,P_{run}\,\langle\epsilon^2(\{s(t)\})\rangle_{\{s(t)\},\{m(t-dt)\}|m(t)=R} .
\tag{53}
$$

Next we separate the remaining expectation into two parts:

$$= \frac{1}{2} \lambda_{R0}\, P_{run} \left\langle \langle \epsilon^2(\{s(t)\}) \rangle_{\{s(t)\}|\{m(t-dt)\},m(t)=R} \right\rangle_{\{m(t-dt)\}|m(t)=R}. \tag{54}$$

Since the term inside the angled brackets is already order $\epsilon^2(\{s(t)\})$, only the zeroth order term of $P(\{s(t)\}|\{m(t-dt)\}, m(t) = R)$ contributes to the inner expectation:

$$= \frac{1}{2} \lambda_{R0}\, P_{run} \left\langle \langle \epsilon^2(\{s(t)\}) \rangle \right\rangle_{\{m(t-dt)\}|m(t)=R}. \tag{55}$$

The inner expectation is taken with respect to $P_0(\{s(t)\}|m(t) = R)$. What remains there does not depend on $\{m(t - dt)\}$, so we can integrate out $\{m(t - dt)\}$, leaving:

$$\begin{aligned} \dot{I}_{s\to m} &= \frac{1}{2} \lambda_{R0}\, P_{run} \langle \epsilon^2(\{s(t)\}) \rangle \\ &= \frac{1}{2} \lambda_{R0}\, P_{run} \left\langle \left(\frac{\lambda_R(\{s(t)\}) - \lambda_{R0}}{\lambda_{R0}} \right)^2 \right\rangle. \end{aligned} \tag{56}$$

This equation indicates that in the linear regime, the information rate measures the signal-induced variation in the tumble rate $\lambda_R(\{s\})$. The factor of $\lambda_{R0}\, P_{run}$ indicates that $\dot{I}_{s\to m}$ measures the long-time, signal-induced variation in $\lambda_R(\{s\})$ times the frequency of tumble events. Larger signal-induced variations in the tumble rate indicates larger information transfer from signal to behavior. When tumbles occur more frequently on average (larger λ_{R0}), those variations in tumble rate cause more apparent changes in the run-tumble statistics, increasing the transfer entropy rate.

Section 4: Information-performance bound with fixed behavioral parameters

Here we derive the bound in Eqn. 1 of the main text and in Figs. 1 and 3. As we've shown above, the drift speed v_d and the information rate $\dot{I}_{s\to m}$ depend on the behavioral response $\lambda_R(\{s\})$, or $\epsilon(\{s\})$. However, while some responses correspond to high information rates, they don't necessarily generate high drift speed. We reiterate that we are looking for the behavioral response that achieves the highest drift speed with a given information rate—this is not necessarily the response that produces the highest information rate or the highest drift speed.

To find the behavioral response $\epsilon(\{s(t)\})$ that maximizes the drift speed at fixed information rate, we solve:

$$\max_{\epsilon(\{s(t)\})} \mathcal{L}(\epsilon(\{s(t)\}), \boldsymbol{\theta}) \tag{57}$$

where $\boldsymbol{\theta} = \{\alpha, \lambda_{R0}, P_{run}\}$ are the behavioral parameters characterizing the cell's swimming in absence of a gradient (this is an alternative parameterization from the one in the main text). The Lagrangian is:

$$\mathcal{L} = v_d - \beta_1\, \dot{I}_{s\to m} - \beta_2 \langle \epsilon(\{s(t)\}) \rangle. \tag{58}$$

β_1 and β_2 are Lagrange multipliers (and since $\dot{I}_{s\to m} \geq 0$, we must have that $\beta_1 \geq 0$). The second term constrains the information rate, while the third term enforces the constraint that $\langle \epsilon(\{s(t)\}) \rangle = 0$ (see previous section). The full expressions for the quantities above are:

$$v_d = v_0 \frac{(1-\alpha)\,\lambda_{R0}}{(1-\alpha)\,\lambda_{R0} + 2\,D_r} P_{run} \left\langle \epsilon(\{s(t)\}) \cos\big(\theta(t)\big) \right\rangle, \tag{59}$$

$$\dot{I}_{s\to m} = \frac{1}{2}\,\lambda_{R0}\,P_{run}\,\langle \epsilon^2(\{s(t)\}) \rangle. \tag{60}$$

Our efforts in previous sections to write v_d and $\dot{I}_{s\to m}$ in terms of $\epsilon(\{s(t)\})$ and to write expectations with respect to the same distributions make taking the functional derivative with respect to $\epsilon(\{s(t)\})$ easier here. Setting it equal to zero gives:

$$\frac{\delta \mathcal{L}}{\delta \epsilon(\{s(t)\})} = 0$$
$$\to v_0 \frac{(1-\alpha)\,\lambda_{R0}}{(1-\alpha)\,\lambda_{R0} + 2\,D_r}\, P_{run} \cos\big(\theta(t)\big) - \beta_1\,\lambda_{R0}\,P_{run}\,\epsilon^*(\{s(t)\}) + \beta_2 = 0 \tag{61}$$

$$\epsilon^*(\{s(t)\}) = -\frac{\beta_2}{\beta_1\,\lambda_{R0}\,P_{run}} + \frac{1}{\beta_1\,\lambda_{R0}} v_0 \frac{(1-\alpha)\,\lambda_{R0}}{(1-\alpha)\,\lambda_{R0} + 2\,D_r} \cos\big(\theta(t)\big) \tag{62}$$

$$= -\frac{\beta_2}{\beta_1\,\lambda_{R0}\,P_{run}} + \frac{1}{\beta_1} \frac{(1-\alpha)}{(1-\alpha)\,\lambda_{R0} + 2\,D_r} \frac{1}{g}\; s(t) \tag{63}$$

$$= A + B\,s(t), \tag{64}$$

with $B = \frac{1}{\beta_1}\frac{1}{g}\frac{(1-\alpha)}{(1-\alpha)\,\lambda_{R0}+2\,D_r}$. Enforcing that $\langle \epsilon(\{s(t)\}) \rangle = 0$ gives that $A = 0,\ \beta_2 = 0$.

Comparing to Eqn. (14) of the drift speed derivation, $\epsilon(\{s(t)\}) = \int_{-\infty}^{t} K_b(t-t')\,s(t')\,dt'$, we see that the optimal strategy is achieved by a behavioral kernel that is proportional to a delta function at the current time:

$$K_b^*(T) = \epsilon_0\,\delta(T), \tag{65}$$

and $B = \epsilon_0$. Putting this together, we have:

$$\epsilon^*(\{s(t)\}) = \epsilon_0\,s(t). \tag{66}$$

The strategy that achieves the highest drift at a given information rate, in the regime of shallow gradients, is to modulate the tumble rate proportionally to the current signal $s(t)$. For this strategy, *all* signal-induced variations in the tumble rate contribute the gradient climbing—no signal-induced variations go unused. This optimal strategy can't be achieved by a real cell, which measures the signal by comparing concentrations at different times [33,34]. Still, no cell or agent can outperform this optimal strategy.

At the optimum:

$$v_d^* = \frac{v_0}{3}\,\frac{(1-\alpha)\,\lambda_{R0}}{(1-\alpha)\,\lambda_{R0} + 2\,D_r}\,(v_0\,g\,\epsilon_0)\,P_{run}, \tag{67}$$

$$\dot{I}^*_{s\to m} = \frac{1}{2}\lambda_{R0}\,\frac{(v_0\,g\,\epsilon_0)^2}{3}\,P_{run}. \tag{68}$$

Expressing the drift speed in terms of the information rate, one finds that the drift speed cannot exceed:

$$\frac{v_d}{v_0} \le \frac{(1-\alpha)\,\lambda_{R0}}{(1-\alpha)\,\lambda_{R0} + 2\,D_r}\,P_{run}\,\left(\frac{2}{3}\,\frac{\dot{I}_{s\to m}}{\lambda_{R0}\,P_{run}}\right)^{\frac{1}{2}}. \tag{69}$$

The term $\frac{\dot{I}^*_{s\to m}}{\lambda_{R0}\,P_{run}}$ inside the square root can be interpreted as the information the cell gets about the signal trajectory per run: higher information per run is necessary, but not sufficient, for higher drift. The factor outside of the square root determines how well the information is translated into drift in the presence of rotational diffusion. Therefore, in addition to information acquisition, navigation depends on the matching between the cell's behavioral parameters and the properties of the physical environment. Comparing to Eqn. 1 of the main text, we have

$$f(\boldsymbol{\theta}) = \frac{(1-\alpha)\,\lambda_{R0}}{(1-\alpha)\,\lambda_{R0} + 2\,D_r}\left(8\,\frac{D_r}{\lambda_{R0}}\,P_{run}\right)^{\frac{1}{2}}. \tag{70}$$

Section 5: Relevant bits for bacterial chemotaxis

$\dot{I}_{s\to m}$ measures the rate of information transfer from the full history of signal to motor behavior. But the response strategy that maximizes drift speed given some information rate $\dot{I}_{s\to m}$ is one that only responds to the current signal. This suggests that the information rate can be partitioned into two parts: information about current signal, which drives chemotaxis, and information about the rest of the signal trajectory. We will now show that the drift speed is proportional to the square root of the former, *for any behavioral response*. Among strategies with *the same information rate* $\dot{I}_{s\to m}$, the optimal response achieves the highest drift speed by only transferring information about current signal $s(t)$. We will make this more precise below.

$\dot{I}_{s\to m}$ can be partitioned as follows:

$$\dot{I}_{s\to m} = \frac{1}{dt}\left\langle \log\left(\frac{P(m(t+dt)|\{m(t)\},\{s(t)\})}{P(m(t+dt)|\{m(t)\})}\right)\right\rangle_{\{m(t)\},\{s(t)\},m(t+dt)} \tag{71}$$

(72)

$$= \frac{1}{dt}\left\langle \log\left(\frac{P\big(m(t+dt)\big|\{m(t)\},s(t)\big)}{P(m(t+dt)|\{m(t)\})}\right)\right\rangle_{\{m(t)\},\{s(t)\},m(t+dt)}$$
$$+ \frac{1}{dt}\left\langle \log\left(\frac{P(m(t+dt)|\{m(t)\},s(t),\{s(t-dt)\})}{P\big(m(t+dt)|\{m(t)\},s(t)\big)}\right)\right\rangle_{\{m(t)\},\{s(t)\},m(t+dt)}$$

$$= \dot{I}^*_{s\to m} + \dot{I}_{\{s(t-dt)\}\to m|s(t)}. \tag{73}$$

The first term measures the information transferred about current signal $s(t)$ to behavior. The second term measures how much information is transferred about past signals that aren't correlated with $s(t)$. In general, $\dot{I}^*_{s\to m} \leq \dot{I}_{s\to m}$, because $\dot{I}_{s\to m}$ contains information about $s(t)$ as well as the full history of s. For the optimal response, $\dot{I}_{s\to m} = \dot{I}^*_{s\to m}$ and $\dot{I}_{\{s(t)-dt\}\to m|s(t)} = 0$—no information is transferred about past signals that aren't correlated with the current signal. Note that there is a distinction between responding to a signal in the past (i.e. by making $\lambda(\{s\})$ depend on that past signal) and transferring information about a signal: responding to a signal at some time transfers information about all signals at other times that are correlated with it.

Next, we will show that the drift speed of any behavioral response is proportional to $\dot{I}^*_{s\to m}$. The derivation of $\dot{I}^*_{s\to m}$ is nearly identical to that of $\dot{I}_{s\to m}$, but with $\epsilon(\{s(t)\})$ replaced with $\epsilon\big(s(t)\big)$, which is the deviation of the tumble rate in response to the *current* signal, after averaged over histories of signal that end with signal $s(t)$ at the current time t. Carrying out the same steps gives:

$$\dot{I}^*_{s\to m} = \frac{1}{2}\lambda_{R0}\, P_{run}\, \big\langle \epsilon^2\big(s(t)\big)\big\rangle. \tag{74}$$

The arguments in the drift speed derivation let us derive $\epsilon\big(s(t)\big)$ in the linear regime as

$$\begin{aligned}\epsilon\big(s(t)\big) &= \langle \epsilon(\{s(t)\})|s(t), m(t) = R\rangle \\ &= g\, v_0\, \cos\big(\theta(t)\big)\int_0^\infty K_b(T)\, e^{-\left((1-\alpha)\lambda_{R0}+2\,D_r\right)T}\, dT.\end{aligned} \tag{75}$$

Comparing to the drift speed section, the right-hand side of the equality above can be written as:

$$\big\langle \epsilon^2\big(s(t)\big)\big\rangle = 3\, \big\langle \epsilon(\{s(t)\})\cos\big(\theta(t)\big)\big\rangle^2. \tag{76}$$

With this, the relevant information rate to the motors is

$$\dot{I}^*_{s\to m} = \frac{1}{2}\lambda_{R0}\, P_{run}\, \left(v_0\, g\int_0^\infty K_b(T)\, e^{-\left((1-\alpha)\lambda_{R0}+2\,D_r\right)T}\, dT\right)^2. \tag{77}$$

Therefore, we have for the drift speed of *any* behavioral response in the linear regime:

$$v_d = v_0 \frac{(1-\alpha)\,\lambda_{R0}}{(1-\alpha)\,\lambda_{R0} + 2\,D_r} P_{run} \left(\frac{1}{3}\left\langle \epsilon^2\big(s(t)\big)\right\rangle\right)^{\frac{1}{2}} \tag{78}$$

$$= v_0 \frac{(1-\alpha)\,\lambda_{R0}}{(1-\alpha)\,\lambda_{R0} + 2\,D_r} P_{run} \left(\frac{2}{3}\frac{\dot{I}^*_{s\to m}}{\lambda_{R0}\,P_{run}}\right)^{1/2}. \tag{79}$$

The drift speed therefore is set by how much information about the current signal a cell communicates to its behavior. $\dot{I}^*_{s\to m}$ is a kind of predictive information rate [35–37]: it measures the mutual information between the current signal and the cell's behavior state in the next dt, given the past of behavioral states, or $\dot{I}^*_{s\to m} = \frac{1}{dt} MI(s(t); m(t+dt)|\{m(t)\})$. The optimal response strategy makes $\dot{I}_{s\to m} = \dot{I}^*_{s\to m}$.

Eqn. (77) indicates that behavioral responses to signals that occurred within approximately one velocity correlation time, $\tau_v^{-1} = (1-\alpha)\lambda_{R0} + 2\,D_r$, into the past transmit relevant information because those signals are most correlated with the current signal. Responses to signals that occurred further in the past still transmit information, but less of it is relevant.

In static gradients, run-tumble transitions and rotational diffusion determine the signal statistics. We assumed that run-tumble transitions were Poisson processes, so to leading order in the gradient steepness g, the signal is Markovian. Therefore, signals farther in the past are less correlated with the current signal. However, if the signal were non-Markovian, for example if it were oscillatory, responding to past signals that are highly correlated with the current signal could be an effective way of transferring information about the current signal (for example, see Becker et al *PRL* 2015 [38], who solved a different but related optimization problem).

Section 6: Information-performance bound with optimal behavioral parameters

The bound derived in the previous section is valid for any behavioral parameters α, λ_{R0}, and P_{run}. In the main text, we measured how efficiently a typical RP437 *E. coli* uses information to climb gradients by comparing their performance to the maximum they could possibly achieve, which was set by measuring the behavioral parameters of the median phenotype P_{run} and plugging them into the bound (right hand side of supplemental Eqn. (69), main text Eqn. 1). But individual cells in an isogenic population or different strains of *E. coli* can have different behavioral parameters. The performance of these cells is bounded by a different curve from the one in Fig. 3 of the main text, given by Eqn. (69) with those cells' behavioral parameter values. In principle, the Lagrangian $\mathcal{L}$ can be further optimized with respect to the behavioral parameters to find the bound that no individual cell can exceed.

First setting the derivative of the Lagrangian with respect to λ_{R0} equal to zero, we get:

$$\frac{\partial \mathcal{L}}{\partial \lambda_{R0}} = \frac{v_0}{3} \frac{2\,D_r\,(1-\alpha)}{\left((1-\alpha)\,\lambda^*_{R0} + 2\,D_r\right)^2} (v_0\,g\,\epsilon_0)\,P_{run} - \frac{1}{6}\,\beta_1\,(v_0\,g\,\epsilon_0)^2\,P_{run} = 0 \tag{80}$$

Using $\beta_1 = \frac{1}{g\,\epsilon_0}\,\frac{(1-\alpha)}{(1-\alpha)\,\lambda_{R0}+2\,D_r}$ from the expressions above, this reduces to:

$$(1-\alpha)\,\lambda_{R0}^{*} = 2\,D_r. \tag{81}$$

This indicates that at the optimum, the rate at which the ideal cell actively changes direction by tumbling equals the rate at which it passively changes direction by rotational diffusion. The optimal value of this behavioral parameter depends on the physical environment through D_r. Interestingly, others have arrived at the same optimal mean tumble rate, but by optimizing a different objective[10].

For the optimal value of λ_{R0}, the drift speed is:

$$\frac{v_d}{v_0} \leq \left(\frac{(1-\alpha)\,\dot{I}_{s\to m}}{12\,D_r}\;P_{run}\right)^{1/2}, \tag{82}$$

and $f(\boldsymbol{\theta})$ reduces to $f(\boldsymbol{\theta}) = \left((1-\alpha)\,P_{run}\right)^{1/2}$.

Finally, the right-hand side of the inequality above is maximized when the cell's tumbles completely reorient its swimming direction, i.e. $\alpha = 0$, and tumbles are instantaneous, i.e. $P_{run} = 1$:

$$\frac{v_d}{v_0} \leq \left(\frac{\dot{I}_{s\to m}}{12\,D_r}\right)^{1.2}. \tag{83}$$

This is the absolute maximum speed at which a cell could possibly climb a gradient given its information rate. With these optimal behavioral parameters, $f(\boldsymbol{\theta}) = 1$, its maximal value. But, in addition to responding instantaneously, achieving $\alpha = 0$ and $P_{run} = 1$ might not be physically realizable for *E. coli* cells. If reorienting completely during tumbles takes finite time, then having $\alpha = 0$ might require that $P_{run} < 1$.

With this, we have series of bounds:

$$\frac{v_d}{v_0} \leq \frac{(1-\alpha)\,\lambda_{R0}}{(1-\alpha)\,\lambda_{R0}+2\,D_r}\;P_{run}\left(\frac{2}{3}\frac{\dot{I}_{s\to m}}{\lambda_{R0}\,P_{run}}\right)^{\frac{1}{2}} \leq \left(\frac{(1-\alpha)\,\dot{I}_{s\to m}}{12\,D_r}\;P_{run}\right)^{\frac{1}{2}} \leq \left(\frac{\dot{I}_{s\to m}}{12\,D_r}\right)^{\frac{1}{2}}. \tag{84}$$

The first inequality comes from optimizing the cell's behavioral response to signal; the second one comes from additionally optimizing the mean tumble rate λ_{R0}; and the third one comes from further optimizing α and P_{run}. These bounds are plotted in Supplementary Fig. S10.

Although we derived the optimal parameters assuming the optimal (instantaneous) behavioral response, these parameters should still be optimal for any given kernel. We focus on λ_{R0} in particular because, for the parameter values we measured in *E. coli*, it has the biggest effect on the bound. First, we found that the highest drift speed at fixed information rate is achieve by responding to the current signal only. However, if the cell responds to signals that are highly correlated with the current one, it can approach the bound. Longer runs make the signal correlation time longer, making the efficiency of any given kernel get closer to that of the instantaneous one. Second, runs longer than $1/(2\,D_r)$ waste information because they lose direction before the signal elicits a change in run duration. As a result, the optimal λ_{R0} should still be about $2\,D_r$ for any fixed kernel.

Section 7: Imperfect adaptation

To understand how imperfect adaptation affects our results, we need to revisit the mapping between response kernels that act on concentration and those that act on the rate of change of concentration.

Celani and Vergassola [10] derived a cell's drift speed in terms of its behavioral response kernel that acts on absolute concentration (see also Wong-Ng et al. [12]), even for imperfectly-adapting responses. As above, concentrations experienced along the cell's swimming trajectory modulate the tumble rate according to:

$$\begin{gathered}\lambda_R(t) = \lambda_{R0}\big(1-\epsilon(t)\big) \\ \epsilon(t) = \int_{-\infty}^{t} R_b(t-t')\, c(t')\, dt'\end{gathered} \tag{85}$$

The response function comes from linearizing the cell's biochemical signaling pathway and motor responses. As noted before [12], this response is only locally valid around a certain background concentration c_0. Although we will model the responses to $c(t)$ and dc/dt in this section, the approach is the same for responses to $\log\big(c(t)\big)$ and $\frac{d}{dt}\log\big(c(t)\big)$. If adaptation is imperfect, the kernel $R_b(T)$ integrates to value different from zero, $A = \int_0^\infty R_b(T)\, dT \neq 0$.

The drift speed is given by Eqn. (16) above, but there is some ambiguity in the literature about which value of the average tumble rate and which kernel $R_b(T)$ to use. We will clarify this ambiguity here, and in the process demonstrate that a kernel acting on *deviations* in concentration from the background, $c(t) - c_0$, is equivalent to one acting on the rate of change of concentration $s(t) = dc/dt$, even when the response does not adapt perfectly.

To start, we note that if the response does not adapt perfectly, then the baseline tumble rate is a function of the background concentration c_0. In terms of the expressions above,

$$\begin{aligned}\lambda_R(t) &= \lambda_{R0}\left(1 - \int_{-\infty}^{t} R_b(t-t')\, c(t')\, dt'\right) \\ &= \lambda_{R0}\left(1 - \int_{-\infty}^{t} R_b(t-t')(c(t') - c_0 + c_0)\, dt'\right) \\ &= \lambda_{R0}\left(1 - A\,c_0 - \int_{-\infty}^{t} R_b(t-t')(c(t') - c_0)\, dt'\right) \\ &= \lambda_{R0}(1 - A\,c_0)\left(1 - \frac{1}{1 - A\,c_0}\int_{-\infty}^{t} R_b(t-t')(c(t') - c_0)\, dt'\right)\end{aligned} \tag{86}$$

$$\lambda_R(t) = \lambda_{R1}\left(1 - \int_{-\infty}^{t} R_b'(t-t')(c(t') - c_0)\, dt'\right), \tag{87}$$

where

$$\begin{aligned} \lambda_{R1} &= \lambda_{R0}(1 - A\, c_0) \\ R_b'(T) &= \frac{R_b(T)}{1 - A\, c_0}. \end{aligned} \tag{88}$$

Here we see that the effect of the background c_0 is to multiply the average tumble rate λ_{R0} and divide the response function $R_b(T)$ by a factor of $(1 - A\, c_0) < 1$. If $A > 0$, this reduces the average tumble rate and increases the effective gain of the response. The gain increases because the same absolute response amplitude causes a larger relative change in $\lambda_R(t)$ when the baseline rate λ_{R1} is lower. Note that when $A = 0$, $\lambda_{R0} = \lambda_{R1}$, $R_b(T) = R_b'(T)$, and the c_0 term drops after integration.

Using simulations (details below), we verified that the equation for the drift speed is indeed given by Eqn. (16), however the average tumble rate and response function should be those in Eqn. (88), i.e.:

$$v_d = \frac{v_0^2}{3}\frac{(1-\alpha)\,\lambda_{R1}}{\left((1-\alpha)\,\lambda_{R1} + 2\,D_r\right)^2}\, P_{run}\, g \int_0^{\infty} \exp\left(-\left((1-\alpha)\,\lambda_{R1} + 2\,D_r\right) T\right)\, R_b'(T)\, dT. \tag{89}$$

Next, we show that these expressions can equivalently be written in terms of a response function $K_b(T) = \int_0^T R_b'(T')\, dT'$ acting on $s(t) = dc/dt$, with

$$\lambda_R(t) = \lambda_{R1}\left(1 - \int_{-\infty}^{t} K_b(t-t')\, s(t')\, dt'\right) \tag{90}$$

and

$$v_d = \frac{v_0^2}{3}\frac{(1-\alpha)\,\lambda_{R1}}{(1-\alpha)\,\lambda_{R1} + 2\,D_r}\, P_{run}\, g \int_0^{\infty} \exp\left(-\left((1-\alpha)\,\lambda_{R1} + 2\,D_r\right) T\right)\, K_b(T)\, dT. \tag{91}$$

To show this, we start from Eqn. (90). Since the system is stationary, we change variables to $T = t - t'$ and t:

$$\lambda_R(t) = \lambda_{R1}\left(1 - \int_0^{\infty} K_b(T)\, s(t - T\ )\, dT\right). \tag{92}$$

Plugging in proposed expression for $K(T)$ gives

$$\lambda_R(t) = \lambda_{R1}\left(1 - \int_0^\infty \left(\int_0^T R_b'(T')\, dT'\right)\, s(t - T\,)\, dT\right). \tag{93}$$

Then we integrate by parts:

$$\int_0^\infty \left(\int_0^T R_b'(T')\, dT'\right)\, s(t - T\,)\, dT = -\left(\int_0^\infty R_b'(T)\, dT\right) c(-\infty) + \int_0^\infty R_b'(T)\, c(t - T\,)\, dT \tag{94}$$

Here, we identify $c(-\infty)$ as the background concentration c_0 around which the system was linearized. Then the above line is equal to

$$\int_0^\infty \left(\int_0^T R_b'(T')\, dT'\right)\, s(t - T\,)\, dT = \int_0^\infty R_b'(T)\, \left(c(t - T\,) - c_0\right)\, dT. \tag{95}$$

The tumble rate is then

$$\lambda_R(t) = \lambda_{R1}\left(1 - \int_0^\infty R_b'(T)\, \left(c(t - T\,) - c_0\right)\, dT\right), \tag{96}$$

Which is equal to Eqn. (87) above. This equivalence indicates that a kernel responding to dc/dt captures the same information as one that responds to $c(t) - c_0$, i.e. changes in concentration relative to the background around which the system was linearized.

To show that the drift speed in Eqn. (91) is also correct, this time we start from the drift speed in Eqn. (89). First, we integrate by parts to get:

$$\begin{aligned} v_d = \frac{v_0^2}{3}\frac{(1-\alpha)\,\lambda_{R1}}{\left((1-\alpha)\,\lambda_{R1} + 2\,D_r\right)^2}\, P_{run}\, g\, \Bigg(&-\left(\int R_b'(T)\, dT\right)_{T=0} \\ &+ \left((1-\alpha)\,\lambda_{R1} + 2\,D_r\right)\int_0^\infty \exp\left(-\left((1-\alpha)\,\lambda_{R1} + 2\,D_r\right)T\right)\,\left(\int R_b'(T')\, dT'\right) dT\Bigg) \end{aligned} \tag{97}$$

$$\begin{aligned} = \frac{v_0^2}{3}\frac{(1-\alpha)\,\lambda_{R1}}{\left((1-\alpha)\,\lambda_{R1} + 2\,D_r\right)}\, P_{run}\, g\, \Bigg(&-\frac{\left(\int R_b'(T)\, dT\right)_{T=0}}{\left((1-\alpha)\,\lambda_{R1} + 2\,D_r\right)} \\ &+ \int_0^\infty \exp\left(-\left((1-\alpha)\,\lambda_{R1} + 2\,D_r\right)T\right)\,\left(\int R_b'(T')\, dT'\right) dT\Bigg) \end{aligned} \tag{98}$$

$$= \frac{v_0^2}{3}\frac{(1-\alpha)\,\lambda_{R1}}{\left((1-\alpha)\,\lambda_{R1} + 2\,D_r\right)}\, P_{run}\, g\, \left(\int_0^\infty \exp\left(-\left((1-\alpha)\,\lambda_{R1} + 2\,D_r\right)T\right)\,\left(\int_0^T R_b'(T')\, dT'\right)\, dT\right) \tag{99}$$

$$= \frac{v_0^2}{3}\frac{(1-\alpha)\,\lambda_{R1}}{\left((1-\alpha)\,\lambda_{R1} + 2\,D_r\right)}\, P_{run}\, g\, \left(\int_0^\infty \exp\left(-\left((1-\alpha)\,\lambda_{R1} + 2\,D_r\right)T\right)\, K_b(T)\, dT\right) \tag{100}$$

which agrees with Eqn. (91) above.

For completeness, we also derive the reverse mapping, from $K_b(T)$ to $R_b'(T)$. We start from Eqn. (92) for the tumble rate $\lambda_R(t)$. We then do the integral by parts to get

$$\int_0^\infty K_b(T)\, s(t-T\,)\, dT = c(t)\, K_b(0) - c(-\infty)\, K_b(\infty) + \int_0^\infty \frac{dK_b(T)}{dT}\, c(t-T\,)\, dT. \tag{101}$$

We take $c(-\infty) = c_0$, the background concentration, and recognize that $K_b(\infty) = \int_0^\infty R_b'(T)\, dT$. Moving the $c(t)$ term into the integral, we get

$$\begin{aligned}&\int_0^\infty K_b(T)\, s(t-T\,)\, dT \\ &\qquad = \int_0^\infty \left(K_b(0)\, \delta(T) + \frac{dK_b(T)}{dT}\right) c(t-T\,)\, dT - c_0 \int_0^\infty R_b'(T)\, dT.\end{aligned} \tag{102}$$

Comparing to Eqn. (87), this means

$$R_b'(T) = K_b(0)\, \delta(T) + \frac{dK_b(T)}{dT}. \tag{103}$$

When taking integrals, the Dirac delta function is treated as being entirely within the domain of integration. Therefore,

$$\begin{aligned}\int_0^T R_b'(T')\, dT' &= \int_0^T K_b(0)\, \delta(T') + \frac{dK_b(T')}{dT'}\, dT' \\ &= K_b(0) + K_b(T) - K_b(0) \\ &= K_b(T),\end{aligned} \tag{104}$$

for all T, even as $T \to 0$.

Taken together, these expressions indicate that the typical approach of writing down a response of the form in Eqn. (85) is actually

$$\lambda_R(t) = \lambda_{R0}(c_0 = 0)\left(1 - \int_{-\infty}^t R_b(t-t'|c_0 = 0)\, c(t')\, dt'\right), \tag{105}$$

i.e. it linearizes the system around $c_0 = 0$, *not* around the background concentration c_0. The steps that transform this response into Eqn. (87) (with λ_{R1} and $R_b'(T)$) essentially extrapolate the linearization around $c_0 = 0$ to a nonzero background concentration. But the dependence of the average tumble rate and the response gain on background concentration are most likely *not* linear, and even A can change with c_0 [23], making this approach give a less accurate estimate of the drift speed. Thus, Eqn. (88) is only valid near a background of $c_0 \sim 0$. The more accurate approach is to linearize the system's response to $c(t) - c_0$, around a particular c_0, to begin with:

$$\lambda_R(t) = \lambda_{R0}(c_0)\left(1 - \int_{-\infty}^{t} R_b(t - t'|c_0)\,(c(t') - c_0)\,dt'\right). \tag{106}$$

Here, λ_{R1} is replaced by $\lambda_{R0}(c_0)$, the average tumble rate in background c_0, and $R_b'(T)$ is replaced by $R_b(T|c_0)$, the behavioral response function in background c_0. Done this way, $\lambda_{R0}(c_0)$ and $R_b(T|c_0)$ already account for the fact that the cell is adapted to a background c_0, and no rescaling of the parameters is needed. As mentioned earlier, this formulation also applies to perfectly-adapting responses. This clarifies that linear theories of chemotaxis linearize the cell's responses to changes in concentration around a particular background c_0, and perfect adaptation is a special case. Even for perfectly-adapting responses, c_0 can enter the gain of $R_b(T)$ or $K_b(T)$ if it is close to the receptor dissociation constants for ligand when in the active or inactive states, K_a and K_i, respectively. This approach requires the gradient to be shallow, in part so that the cell's responses are small and approximately linear in past signals, but also so that the cell reaches a steady state drift speed before the background concentration changes appreciably. This second condition makes the shallow gradient regime different from the small gain regime in the case of imperfect adaptation.

To see how imperfect adaptation is encoded in $K_b(T)$, the response to the rate of change of concentration, we consider an example. Take as the kernel acting on concentration:

$$R_b(T|c_0) = R_0 \exp(-\lambda\, T). \tag{107}$$

This kernel response does not adapt perfectly: $A(c_0) = \int_0^\infty R_b(T|c_0)\,dT = R_0/\lambda$. Here, $R_b(T|c_0)$ is understood to be linearized around the correct c_0, as in Eqn. (106). The corresponding $K_b(T)$ is

$$K_b(T) = \int_0^T R_b(T'|c_0)\,dT' = \frac{R_0}{\lambda}(1 - \exp(-\lambda\, T)). \tag{108}$$

Imperfect adaptation is encoded by the property that $K_b(\infty) = \int_0^\infty R_b(T|c_0)\,dT = A(c_0) = R_0/\lambda \neq 0$.

These results show that the equations we used for drift speed and information rate earlier are valid even when the cells' responses don't adapt perfectly, with the caveat that they are locally valid around a particular background concentration c_0. Therefore, our analysis did not preclude an imperfectly-adapting kernel as a possible solution. However, we found that the optimal behavioral response kernel, i.e. $K_b^*(T) = \epsilon_0\, \delta(T)$, does adapt perfectly because $K_b^*(\infty) = 0$. This can be understood from the perspective of relevant versus irrelevant information. Signals far in the past are uncorrelated with the current signal, but imperfect adaptation implies that the cell responds to a signal (a change in concentration) that occurred infinitely far in the past, $K_b(\infty) > 0$. Therefore, responding to them only transmits irrelevant information and lowers efficiency of information usage.

Simulations to confirm Eqn. (89) were performed using the kernel in Eqn. (107) and taking the common approach of linearizing around $c_0 = 0$. In the simulation, a cell swims in 3D and climbs a gradient of concentration that increases in the x direction. Like in ref [12], we took $v_0 = 20\ \mu\mathrm{m/s}$, $D_r = 0$ for simplicity, $\lambda_{R0} = 1\ \mathrm{s}^{-1}$, $c_0 = 1$ mM, and the linear gradient steepness was $g = 1\ \mu\mathrm{M}/\mu\mathrm{m}$. Some differences were that tumbles were taken to fully reorient the cell, so $\alpha = 0$, the gain was set higher to $R_0 = 1.5\ x\ 10^{-4}$ $(\mu\mathrm{M\ s})^{-1}$, and the kernel rate parameter $\lambda = 0.5\ \mathrm{s}^{-1}$. Equations of motion were integrated by a forward Euler method with time step $\Delta t = \tau_v/50$, where $\tau_v^{-1} = \lambda_{R1} = \lambda_{R0}(1 - c_0\ R_0/\lambda) \sim 0.7\ \lambda_{R0}$ in this case. Cells were initialized as (imperfectly) adapted to the background. 10^5 cells were simulated for 120 seconds. The first $20\ \tau_v$ seconds of the simulation were thrown out, during which the cells forgot the initial condition and reached steady state. The drift speed was computed as the average velocity in the gradient direction, over all cells and all time. The resulting value of drift speed agreed well with the prediction of Eqn. (89), consistent with the changes in gain and average tumble rate caused by imperfect adaptation.

Section 8: Information rate from signal to kinase activity

Measuring information transfer between two continuous, time-dependent variables is highly nontrivial because of the need to infer high-dimensional probability distributions. However, a simple analytical expression has been derived in the special case of linear, Gaussian systems. If $S(\omega)$ is the power spectrum of the input signal, $K(\omega)$ is the linear response function mapping input to output, and $N(\omega)$ is the power spectrum of spontaneous noise in the output, then the mutual information rate $\dot{I}(X;Y)$ between input X and output Y is [39,40]:

$$\dot{MI}(X;Y) = \frac{1}{4\,\pi}\int_{-\infty}^{\infty} \log\left(1 + \frac{S(\omega)\ |K(\omega)|^2}{N(\omega)}\right) d\omega. \tag{109}$$

In our case, the input signal is $s(t) = \frac{d}{dt}\log\big(c(t)\big)$ and the output is $a(t)$, the activity of CheA kinases.

The mutual information rate between signal and kinase activity $\dot{MI}(s;a)$ is not generally equal to the transfer entropy rate from signal to kinase activity $\dot{I}_{s\to a}$, because kinase activity affects the cell's behavior, which feeds back onto the signal. Instead, $\dot{MI}(s;a) = \dot{I}_{s\to a} + \dot{I}_{a\to s}$. However, by breaking the feedback of kinase onto signals, we can quantify $\dot{I}_{s\to a}$ from $\dot{MI}(s;a)$. To do this, we measured kinase responses and noise in immobilized cells. Then, we separately measured the signal statistics from freely swimming cells (see below). In this setup, $\dot{I}_{a\to s} = 0$, so the mutual information rate and transfer entropy rate are equal: $\dot{MI}(s;a) = \dot{I}_{s\to a}$. Breaking this feedback excludes some correlations between s and a, making our estimate of $\dot{I}_{s\to a}$ slightly different from its value in a swimming cell. In particular, when a fluctuation lowers a, the run duration and signal correlation time get longer, which could affect $\dot{I}_{s\to a}$. But these variations in run duration are not correlated with the *sign* of the signal: the fluctuation makes the signal longer-lived, regardless of whether the cell is going up or down the gradient. Therefore, they don't change the covariance between signal and kinase activity, only higher-order correlations. As a result, this effect will only enter as a correction term multiplied by g^4, and it can be neglected in the shallow gradients we considered here.

Eqn. (109) is valid for Gaussian inputs and outputs related by a linear mapping. The response of a to s is approximately linear for the range of signals we used (see Supplementary Fig. S5 and previous work [15,41]), and the fluctuations in a are well-described by a Gaussian process (see noise autocorrelation functions in Fig. 2G, Supplementary Fig. S6B, and previous work [26,27]), consistent with the assumptions of Eqn. (109) above. But the statistics of the signal s are not Gaussian. Still, the expression above for the mutual information rate is a good approximation in shallow gradients. In static, one-directional, shallow exponential gradients, the signal is directly proportional to the component of the cell's velocity projected onto the gradient direction, $v_x(t)$: $s(t) = \frac{d}{dt}\log\left(c\left(x(t)\right)\right) = g\ v_x(t) = g\ v_0 \cos\left(\theta(t)\right)$, where θ is the angle between the cell's swimming direction and gradient direction. In the long term and to leading order in g, the distribution of $\cos\left(\theta(t)\right)$ (in 3D space) is uniform between -1 and 1, therefore s is uniform between $[-v_0\ g, v_0\ g]$. Eqn. (109) only keeps the second moment of the signal statistics; higher moments are zero for a Gaussian process. However, higher moments of s are multiplied by higher powers of g (i.e. g^4), and therefore they only add small corrections to the Gaussian information rate in shallow gradients. (The next power of g is g^4, not g^3, because the mutual information can only depend on even powers of g; otherwise, it would depend on whether the gradient was in the positive or negative x direction). This makes Eqn. (109) a good approximation of the information rate, even though the signal s is not Gaussian.

We can simplify the expression in Eqn. (109). As noted in the paragraph above, in static, exponential concentration profiles that vary in one direction the signal is $s(t) = g\ v_x(t)$. Therefore, the power spectrum of the signal is just $S(\omega) = g^2\ V(\omega)$, where $V(\omega)$ is the power spectrum of $v_x(t)$. Plugging this in, we have:

$$\dot{I}_{s\to a} \sim \dot{M}I(s;a) = \frac{1}{4\,\pi}\int_{-\infty}^{\infty} \log\left(1 + g^2 \frac{V(\omega)\ |K(\omega)|^2}{N(\omega)}\right) d\omega. \tag{110}$$

In shallow gradients, the term inside the log multiplied by g^2 is small for all frequencies ω. Therefore, we can further simplify this by linearizing the log:

$$\dot{I}_{s\to a} \sim g^2\ \frac{1}{4\,\pi}\int_{-\infty}^{\infty} \frac{V(\omega)\ |K(\omega)|^2}{N(\omega)} d\omega, \tag{111}$$

from which we see that the information rate is proportional to g^2 in shallow gradients. Calculating the information rate, as described in the next section, using Eqn. (111) overestimates the exact integral in Eqn. (110) by about 1% in the steepest gradients we used experimentally.

These expressions are analogous to the mutual information between two univariate Gaussian random variables related by a linear mapping. Say we have X and $Y = k\ X + \xi$, where the signal $X \sim N(0, \sigma_x^2)$ and the noise $\xi \sim N(0, \sigma_n^2)$. Since noise is assumed to be independent of X, the variance of Y is $\sigma_y^2 = k^2\sigma_x^2 + \sigma_n^2$. Then the mutual information between X and Y can be written in terms of the Pearson correlation r between them:

$$MI(X;Y) = -\frac{1}{2}\log(1 - r^2). \tag{112}$$

By definition,

$$r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2\,\sigma_y^2} = \frac{\langle x\,(k\,x\,+\xi)\rangle^2}{\sigma_x^2\,\left(k^2\sigma_x^2 + \sigma_n^2\right)} = \frac{\left(k\sigma_x^2\right)^2}{\sigma_x^2\,\left(k^2\sigma_x^2 + \sigma_n^2\right)} = \frac{k^2\sigma_x^2}{k^2\sigma_x^2 + \sigma_n^2}. \tag{113}$$

That is, r^2 measures the fraction of total variance in Y that comes from X. Plugging this into the mutual information and rearranging,

$$MI(X;Y) = \frac{1}{2}\log\left(\frac{1}{1 - r^2}\right) \tag{114}$$

$$= \frac{1}{2}\log\left(1 + \frac{k^2\sigma_x^2}{\sigma_n^2}\right) \tag{115}$$

For small signal variance σ_x^2, this can be simplified by linearizing the log:

$$MI(X;Y) \sim \frac{1}{2}\frac{k^2\sigma_x^2}{\sigma_n^2}. \tag{116}$$

Comparing this univariate mutual information to the mutual information rate in Eqns. (109)-(111), we see that the variances of the signal and the noise become their power spectra, $\sigma_x^2 \to S(\omega)$ and $\sigma_n^2 \to N(\omega)$, and the gain becomes the frequency response function, $k \to K(\omega)$. The univariate mutual information quantifies the ratio of signal-induced variance in Y to noise variance, $\frac{k^2\sigma_x^2}{\sigma_n^2}$. Analogously, the mutual information rate (Eqn. (111)) quantifies the ratio of signal-induced power in kinase activity $a(t)$ to noise power, $\frac{S(\omega)|K(\omega)|^2}{N(\omega)}$, integrated over frequency components. Since the chemotactic signal power $S(\omega)$ is proportional to the gradient steepness squared, $S(\omega) = g^2\,V(\omega)$, the signal-induced power in kinase activity and the information rate are both also proportional to g^2 when the gradient is shallow.

Returning to Eqn. (111), even with the dramatic simplification that this expression provides for estimating the mutual information rate, inferring the spectra $V(\omega)$, $K(\omega)$, and $N(\omega)$ nonparametrically from data still requires long trajectories. Instead, we used parameterized models for the signal and noise power spectra and for the signaling kernel that we constrained with data. These phenomenological models are described below.

Section 9: Models of signal statistics and kinase activity noise and response

In this section we explain the models used for the signal statistics, the noise statistics, and the linear response function. We start with the swimming statistics. As described above, in static gradients, the signals a cell experiences are proportional to its up-gradient velocity $v_x(t)$. The statistics of $v_x(t)$ are characterized by their power spectrum $V(\omega)$, which can be computed from the Fourier transform of the autocorrelation function $V(t)$. Importantly, in shallow gradients and to leading order in gradient steepness g, the statistics of $v_x(t)$ are identical to those of a cell swimming in the absence of a gradient. Therefore:

$$V(t) = \langle v_x(t) v_x(0) \rangle = \frac{v_0^2}{3} \, P_{run} \, e^{-\left((1-\alpha)\lambda_{R0} + 2\, D_r\right) |t|} = a_v \, e^{-\lambda_{tot} \, |t|}, \tag{117}$$

where the parameters have the same meaning as in earlier sections. λ_{tot} is the total rate at which the cell loses its swimming direction. The factor of 1/3 results from measuring variations of $v_x(t)$ in 3D space. We define the Fourier transform as

$$\mathcal{F}[f(t)] = \int_{-\infty}^{\infty} f(t) \, e^{-i\omega t} \, dt. \tag{118}$$

Therefore, the power spectrum of $v_x(t)$ is:

$$V(\omega) = \mathcal{F}[V(t)] = 2 \frac{v_0^2}{3} P_{run} \frac{(1-\alpha)\lambda_{R0} + 2\, D_r}{\left((1-\alpha)\lambda_{R0} + 2\, D_r\right)^2 + \omega^2} \tag{119}$$

$$= a_v \frac{2\, \lambda_{tot}}{\lambda_{tot}{}^2 + \omega^2}. \tag{120}$$

We explain how we compute the velocity autocorrelation function of a typical cell, as well as how we measure the behavioral parameters of the model, in **Section 21: Estimating behavioral parameters, run speed, and rotational diffusion**.

Next, we consider the linear response function, which can be inferred directly from the cells' kinase responses to an impulse (delta function) of signal. The cells' impulse responses have a stereotypical shape consisting of a fast change in kinase output, followed by slow relaxation to baseline due to adaptation. A simple phenomenological model that captures these features of the response to a unit impulse of stimulus, $\delta(t)$, is:

$$K(t) = G \left(1 - e^{-\frac{t}{\tau_1}}\right) e^{-\frac{t}{\tau_2}} H(t), \tag{121}$$

where G is the gain, τ_2 is the adaptation time, and τ_1 is the fast response time, and $H(t)$ is the Heaviside step function.

We infer CheA kinase output from FRET between fluorescently-labeled and overexpressed CheY and CheZ (see below), as has been done before [15,18,26,27,41]. Therefore, the FRET signal we measure has a relaxation time arising from CheY-CheZ binding and CheY-p dephosphorylation, and the value of τ_1 we infer includes these relaxation dynamics. The CheA autophosphorylation time has been measured before to be about $\tau_1 \sim 1/60$ s [5,6]; we measure τ_1 with CheY/CheZ relaxation to be about 0.22 s (Supplementary Table S1; **Section 16: Estimating linear response function parameters**). The information rate has a weak dependence on τ_1, and decreasing the value of τ_1 monotonically increases the information rate. We use the value of τ_1 from the literature to compute $\dot{I}_{s\to a}$, which results in a lower information efficiency. Using $\tau_1 = 0.22$ s decreases the information rate by 16% relative to the literature value; using $\tau_1 = 0$ increases it by 6%. The square root dependence of the bound in Eqn. (69) (Eqn. 1 of the main text) makes the effects of these differences on the efficiency even smaller. In Supplementary Fig. S8, we plot the dependence of the information rate on τ_1. Our fit to τ_2, the adaptation time, should not be affected by CheYp-CheZ binding because it is much longer than the measured τ_1. The gain G that we inferred should also not be affected by CheYp-CheZ binding.

The frequency response $K(\omega)$ is the Fourier transform of $K(t)$:

$$K(\omega) = \frac{G}{\tau_1} \frac{1}{\left(\frac{1}{\tau_2} + i\,\omega\right)\left(\frac{1}{\tau_1} + \frac{1}{\tau_2} + i\,\omega\right)}, \tag{122}$$

where $i^2 = -1$. $K(\omega)$ appears in Eqn. (111) for the information rate as $|K(\omega)|^2 = K(\omega)K^*(\omega)$, with $K^*(\omega)$ the complex conjugate of $K(\omega)$. This is:

$$|K(\omega)|^2 = \frac{G^2}{\tau_1^2} \frac{1}{\left(\left(\frac{1}{\tau_2}\right)^2 + \omega^2\right)\left(\left(\frac{1}{\tau_1} + \frac{1}{\tau_2}\right)^2 + \omega^2\right)}. \tag{123}$$

The noise in kinase output $a(t)$ is well-described by an Ornstein-Uhlenbeck process [26]. This process is characterized by two parameters, a diffusivity D_n and a relaxation time scale τ_n. The long-term distribution of the process is Gaussian with variance $\sigma_n^2 = D_n\,\tau_n$. The autocorrelation of this process, in the absence of signal, has the form:

$$\langle a(t)a(0)\rangle = \sigma_n^2\, e^{-\frac{|t|}{\tau_n}} \tag{124}$$

and the power spectrum, $N(\omega)$ is:

$$N(\omega) = \frac{2\,D_n}{\left(\frac{1}{\tau_n}\right)^2 + \omega^2}. \tag{125}$$

Using data from single-cell FRET experiments described in the following section, we constrained the parameters of the response function $K(\omega)$ and the noise spectrum $N(\omega)$: G, τ_1, τ_2, τ_n, and D_n. The parameter inference procedure is described in **Section 16: Estimating linear response function parameters** and **Section 17: Estimating noise statistics parameters**.

Error bars for the parameterized models above that are plotted in Fig. 2 of the main text come from propagating errors in parameter estimation. To compute how errors in parameter estimation propagate to errors in a function $f(x, \boldsymbol{\theta})$ with argument x (such as time t or radial frequency ω) and parameters $\boldsymbol{\theta}$, we used:

$$\sigma_f^2(x) = \sum_i \left|\frac{\partial f(x)}{\partial \boldsymbol{\theta}_i}\right|^2 \sigma_{\boldsymbol{\theta}_i}^2, \tag{126}$$

where $\sigma_f^2(x)$ is the variance in the function f at argument value x, $\boldsymbol{\theta}_i$ is the ith parameter, and $\sigma_{\boldsymbol{\theta}_i}^2$ is the variance of the ith parameter (standard error squared).

At this point, we have models for all of the expressions needed to calculate the transfer entropy rate from signal to kinase output. For the models above, the integral in Eqn. (111) can be solved analytically, giving the following expression for the information rate:

$$\dot{I}_{s\to a} \sim \frac{1}{\log(2)}\frac{1}{4}\left(\frac{g}{\tau_1}G\right)^2 \frac{a_v}{D_n} \frac{\frac{\tau_2}{\tau_n^2}\left(\frac{1}{\tau_1}+\frac{2}{\tau_2}+\lambda_{tot}\right)+\lambda_{tot}\left(\frac{1}{\tau_1}+\frac{1}{\tau_2}\right)}{\left(\frac{1}{\tau_1}+\frac{1}{\tau_2}\right)\left(\frac{1}{\tau_1}+\frac{2}{\tau_2}\right)\left(\lambda_{tot}+\frac{1}{\tau_2}\right)\left(\frac{1}{\tau_1}+\frac{1}{\tau_2}+\lambda_{tot}\right)} = \beta\, g^2, \tag{127}$$

where $\lambda_{tot} = (1-\alpha)\lambda_{R0} + 2\, D_r$ and $a_v = \frac{v_0^2}{3} P_{run}$. The factor of $1/\log(2)$ makes the units bits/s. We used the above expression to compute the information rate in the main text. By inferring model parameters from data, we inferred the prefactor, β.

Error bars for the information rate in Fig. 3 come from propagating uncertainties in the parameters to uncertainty in the information rate using Eqn. (127) above, with $f(x)$ being $\dot{I}_{s\to a}(g)$ in this case.

We emphasize that this is the total information transferred from the trajectory of signal to kinase output. Not all of this information is relevant to gradient climbing—only the part that is informative of the current signal contributes to gradient climbing.

Section 10: Overview of E-FRET analysis method

The goal of our FRET analysis is to quantify the degree of interaction between the phosphorylated response regulator CheY-mRFP and its phosphatase CheZ-mYFP to infer the output of the histidine kinase CheA that phosphorylates CheY [41,42] (see below). For this purpose, we used a 3-filter cube FRET imaging method, E-FRET [43], rather than the FRET method previously used in bacterial chemotaxis studies [18,26,41,42]. The two methods aim to infer the same quantity as a measure of the degree of molecular interaction (see below), but with a different set of observables and assumptions. The key differences are: (i) E-FRET provides a principled photobleaching correction method, enabling us to estimate the degree of molecular interaction quantitatively even in presence of photobleaching, while the previous FRET method gives a biased estimate of the degree of molecular interaction if photobleaching is nonnegligible. (ii) E-FRET is a 3-cube FRET imaging method (i.e., observes three different fluorescent signals), while the previous FRET is a 2-cube method. This makes E-FRET require fewer assumptions. Crucially, unlike the previous method, E-FRET does not require measuring fluorescent signals in the absence of FRET interactions, which is generally hard to measure precisely.

To introduce some notation used below, we consider a general bimolecular FRET system with the donor fluorophore D and acceptor fluorophore A fused to two target molecules X and Y respectively. In the presence of photobleaching, the system contains eight chemical species: $\mathrm{D}^*, \mathrm{D}, \mathrm{A}^*, \mathrm{A}, \mathrm{D}^*\mathrm{A}^*, \mathrm{D}^*\mathrm{A}, \mathrm{DA}^*$, and DA, where fluorescent and non-fluorescent (i.e., photobleached) molecules are represented with and without the star respectively, and free and complexed molecules are also distinguished. We denote background-subtracted fluorescence signals from single cells by I_{DD}, I_{DA}, and I_{AA}, which are respectively the signals obtained through the donor channel (donor excitation and donor emission), the FRET channel (donor excitation and acceptor emission), and the acceptor channel (acceptor excitation and acceptor emission). As in a typical 3-cube imaging setup, we selected filter sets (see Methods) such that the donor is not excited by the acceptor excitation wavelengths and the acceptor emission is not transmitted through the donor emission filter. Under these conditions, the three observables are linked to the concentrations of the chemical species as follows [43]:

$$
\begin{aligned}
I_{DD} &= C_{DD}\,([D^*] + [D^*A] + (1 - E_{max})[D^*A^*]) + \xi_{DD}, \\
I_{AA} &= C_{AA}\,([A^*] + [D^*A^*] + [DA^*]) + \xi_{AA}, \\
I_{DA} &= d_E\,\overline{I_{DD}} + a_E\,\overline{I_{AA}} + C_{DD}\,G_E\,E_{max}\,[D^*A^*] + \xi_{DA},
\end{aligned}
\tag{128}
$$

where ξ_{DD}, ξ_{AA}, and ξ_{DA} are zero-mean shot noise in each channel, and $\overline{I_{DD}}$ and $\overline{I_{AA}}$ are, respectively, the noise-free I_{DD} and I_{AA} signals. E_{max} is the maximum FRET efficiency, where FRET efficiency is defined as the probability of energy transfer from the donor to the acceptor per donor excitation event [43,44], and the maximum FRET efficiency is achieved when only the chemical species $\mathrm{D}^*\mathrm{A}^*$ is present. The first and second term on the right-hand side of the last equation, respectively, represent the bleedthrough of the donor emission into the acceptor emission filter and the cross-excitation of the acceptor by the donor excitation wavelengths. C_{DD}, C_{AA}, a_E, d_E, and G_E are parameters dependent on the imaging system and the photophysical properties of the FRET pair, which are defined as:

$$
\begin{aligned}
C_{DD} &\equiv \nu_D \epsilon_{DD} Q_D L_D S_D t_{DD}, \\
C_{AA} &\equiv \nu_A \epsilon_{AA} Q_A L_A S_A t_{AA}, \\
a_E &\equiv \frac{\nu_D \epsilon_{DA} t_{DA}}{\nu_A \epsilon_{AA} t_{AA}}, \\
d_E &\equiv \frac{L_A S_A t_{DA}}{L_D S_D t_{DD}}, \\
G_E &\equiv \frac{Q_A L_A S_A t_{DA}}{Q_D L_D S_D t_{DD}},
\end{aligned}
\tag{129}
$$

where ν_D (ν_A) is the intensity of illumination reaching the sample through the donor (acceptor) excitation filter, ϵ_{DD} is the absorption coefficient of the donor, ϵ_{DA} (ϵ_{AA}) is the absorption coefficient of the acceptor at the donor-excitation (acceptor-excitation) wavelength, Q_D (Q_A) is the quantum yield of donor (acceptor), L_D (L_A) is the throughput of the donor (acceptor) emission light-path, S_D (S_A) is the quantum sensitivity of the camera for donor (acceptor) emission, and t_{DA}, t_{AA}, and t_{DD} are, respectively, the exposure time for the FRET, acceptor, and donor channels.

As a measure of the degree of molecular interaction, E-FRET, as well as the above-mentioned previous FRET method, aims to obtain the following quantity [43]:

$$\mathcal{E} = \frac{[XY]}{[X_{total}]} E_{max}, \tag{130}$$

where $[X_{total}]$ is the total concentration of the carrier molecule to which the donor is attached (i.e., $[X_{total}] = [X] + [XY]$). In the absence of photobleaching, $\mathcal{E}$ is the FRET efficiency. E-FRET maintains that, under some assumptions (see below), the quantity $\mathcal{E}$ can be estimated by computing the following FRET index from the observables:

$$E_{corr}(t|a_E, d_E, G_E) = \frac{F_c(t|a_E, d_E)}{F_c(t|a_E, d_E) + G_E\, I_{DD}(t)} \frac{I_{AA}(0)}{I_{AA}(t)}, \tag{131}$$

where the sensitized emission F_c is

$$F_c(t|a_E, d_E) = I_{DA}(t) - d_E\, I_{DD}(t) - a_E I_{AA}(t). \tag{132}$$

In practice, the value of $I_{AA}(t)$, which is independent of FRET and thus changes slowly due to photobleaching, is interpolated from sparsely-sampled data points over time, while the values of $I_{DD}(t)$ and $I_{DA}(t)$ are directly measured more frequently in time [43]. In the limit that the system-dependent parameters a_E, d_E, and G_E can be determined with infinite precision, and zero measurement noise of the observables, one can show that E_{corr} converges to $\mathcal{E}$ (see below for proof):

$$E_{corr}(t|a_E, d_E, G_E) \rightarrow \mathcal{E} \equiv \frac{[XY]}{[X_{total}]} E_{max}. \tag{133}$$

Realistically though, a_E, d_E, and G_E can only be estimated with finite uncertainty. In what follows, we will show that errors in estimating these parameters creates a bias in E_{corr} that grows quasi-exponentially over the course of a time-lapse FRET experiment. We also discuss some predictable properties of the bias, and how one can exploit these properties to correct the bias.

Section 11: Derivation of the E-FRET formulae

Based on the original paper [43], here we re-derive the E-FRET formula (Eq. (131)) to make the assumptions and their validation in our system clearer. We first assume that the total concentrations of the donor and acceptor molecules inside the cell are conserved during the experiment, which in our case lasts $\lesssim 40$ minutes. The assumption is satisfied in our experiment because RP437 *E. coli* and their derivatives are auxotrophic for several amino acids, which we do not provide in the experimental media. Therefore, new proteins cannot be synthesized after we wash the cells (see Methods). Furthermore, the interval (~2 hrs) between cell washing and the beginning of the FRET experiment ensures that essentially all fluorescent proteins have matured [45]. Therefore, we have:

$$\begin{aligned}
[D_{total}] &= [D^*](t) + [D](t) + [D^*A^*](t) + [D^*A](t) + [DA^*](t) + [DA](t) \\
[A_{\text{total}}] &= [A^*](t) + [A](t) + [D^*A^*](t) + [D^*A](t) + [DA^*](t) + [DA](t).
\end{aligned} \tag{134}$$

Again, D is the donor molecule (CheZ-mYFP), A is the acceptor molecule (CheY-mRFP), and species with a star are not yet photobleached. We also assume that all of the target molecules are labeled by the fluorescent proteins:

$$\begin{aligned}
[D_{total}] &= [X_{total}] \\
[A_{total}] &= [Y_{total}].
\end{aligned} \tag{135}$$

Next, we assume that the photobleaching rate is a first-order decay process. Combined with the assumption that there's no synthesis of new fluorescent proteins, this leads to:

$$\begin{aligned}
\frac{d([D^*] + [D^*A^*] + [D^*A])}{dt} &= -\delta(t)([D^*] + [D^*A^*] + [D^*A]) \\
\frac{d([A^*] + [D^*A^*] + [DA^*])}{dt} &= -\alpha(t)([A^*] + [D^*A^*] + [DA^*]),
\end{aligned} \tag{136}$$

where $\delta(t)$ and $\alpha(t)$ are the bleaching rates of the donor and acceptor at time t, which can depend on the degree of FRET and hence can depend on t. Solving these equations, we get:

$$\begin{aligned}
[D^*] + [D^*A^*] + [D^*A] &= [X_{total}]\, e^{-\int_0^t \delta(t')dt'} \\
[A^*] + [D^*A^*] + [DA^*] &= [Y_{total}]\, e^{-\int_0^t \alpha(t')dt'}.
\end{aligned} \tag{137}$$

Lastly, we assume that the system is in a quasi-steady state at each time point, i.e. the kinetics of the binding and unbinding of the target molecules X and Y and the diffusion timescales of the donor and acceptor over the enclosed compartment are sufficiently short compared to the time scale of photobleaching. In our system, the binding-unbinding kinetics between CheY-p and CheZ (< 0.3 s [4]) and the time it takes CheY to diffuse throughout the cytoplasm (~0.05 s [46]) are much shorter than the time scale of photobleaching ($\gtrsim 10^2$ s). With this assumption, the fraction of *each free or complexed species* that is not photobleached decays exponentially, analogous to Eqns. (137). From this, we get:

$$\begin{aligned}
\frac{[D^*]}{[D^*] + [D]} &= e^{-\int_0^t \delta(t')dt'}, \\
\frac{[A^*]}{[A^*] + [A]} &= e^{-\int_0^t \alpha(t')dt'},
\end{aligned} \tag{138}$$

Furthermore, defining γ as the binding affinity constant between X and Y:

$$[D^*A^*] = \gamma\,[D^*][A^*] \tag{139}$$

$$= \gamma\left(([D^*]+[D])e^{-\int_0^t \delta(t')dt'}\right)\left(([A^*]+[A])e^{-\int_0^t \alpha(t')dt'}\right) \tag{140}$$

$$= \gamma\left([X]e^{-\int_0^t \delta(t')dt'}\right)\left([Y]e^{-\int_0^t \alpha(t')dt'}\right) \tag{141}$$

$$= [XY]\;e^{-\int_0^t \alpha(t')+\delta(t')dt'}. \tag{142}$$

Under these assumptions, Eqns. (128) become

$$\begin{gathered} I_{DD} \simeq C_{DD}\left([X_{total}]\;e^{-\int_0^t \delta(t')dt'} - E_{max}\;[XY]\;e^{-\int_0^t \alpha(t')+\delta(t')dt'}\right) \\ I_{AA} \simeq C_{AA}\;[Y_{total}]\;e^{-\int_0^t \alpha(t')dt'} \\ I_{DA} \simeq d_E\,I_{DD} + a_E\,I_{AA} + C_{DD}\,G_E\,E_{max}\;[XY]\;e^{-\int_0^t \alpha(t')+\delta(t')dt'}, \end{gathered} \tag{143}$$

and Eqn. (132) becomes

$$F_c(t) = C_{DD}\,G_E\,E_{max}\;[XY]\;e^{-\int_0^t \alpha(t')+\delta(t')dt'}. \tag{144}$$

By plugging these expressions in the formula for E_{corr} in Eqn. (131), one gets

$$E_{corr}(t) = \frac{F_c(t)}{F_c(t)+G_E I_{DD}(t)}\frac{I_{AA}(0)}{I_{AA}(t)} \tag{145}$$

$$\simeq \frac{C_{DD}G_E E_{max}[XY]e^{-\int_0^t \alpha(t')+\delta(t')dt'}}{C_{DD}G_E E_{max}\;[XY]e^{-\int_0^t \alpha(t')+\delta(t')dt'} + C_{DD}G_E\left([X_{total}]\;e^{-\int_0^t \delta(t')dt'} - E_{max}\;[XY]e^{-\int_0^t \alpha(t')+\delta(t')dt'}\right)}\frac{1}{e^{-\int_0^t \alpha(t')dt'}} \tag{146}$$

$$= \frac{E_{max}[XY]}{[X_{total}]} = \mathcal{E}, \tag{147}$$

which is the degree of molecular interaction defined above.

Section 12: Measurements of imaging system parameters

The imaging-system parameters a_E, d_E and G_E were determined in the following way. The cross-talk coefficients a_E and d_E can be estimated by observing the fluorescent signals from strains that express only the acceptor or the donor because [43]

$$\begin{gathered} a_E \equiv \frac{\nu_D \epsilon_{DA} t_{DA}}{\nu_A \epsilon_{AA} t_{AA}} \simeq \frac{I_{DA(A)}}{I_{AA(A)}}, \\ d_E \equiv \frac{L_A S_A t_{DA}}{L_D S_D t_{DD}} \simeq \frac{I_{DA(D)}}{I_{DD(D)}}, \end{gathered} \tag{148}$$

where A (D) in the parentheses in the lower index indicates that the corresponding fluorescent signals are obtained from the strain that only expresses the acceptor (the donor). The approximations above are equalities in the limit of zero measurement noise. The equations can be shown by noting that

$$\begin{aligned} I_{DA(A)} &= [A^*]\nu_D\epsilon_{DA}Q_AL_AS_At_{DA} + \xi_{DA(A)} \\ I_{AA(A)} &= [A^*]\nu_A\epsilon_{AA}Q_AL_AS_At_{AA} + \xi_{AA(A)} \\ I_{DA(D)} &= [D^*]\nu_D\epsilon_{DD}Q_DL_AS_At_{DA} + \xi_{DA(D)} \\ I_{DD(D)} &= [D^*]\nu_D\epsilon_{DD}Q_DL_DS_Dt_{DD} + \xi_{DD(D)}, \end{aligned} \tag{149}$$

where $\xi_{DA(A)}$, $\xi_{AA(A)}$, $\xi_{DA(D)}$ and $\xi_{DD(D)}$ represent shot noise. We obtained estimates for the parameters $a_{E,est}$ and $d_{E,est}$ by linear least-squares fitting the background-subtracted fluorescence signals from hundreds of cells (Supplementary Fig. S4AB):

$$\begin{aligned} a_{E,est} &= \arg\min_{a_E} \sum_i \left(a_E\, I_{AA(A),i} - I_{DA(A),i}\right)^2, \\ d_{E,est} &= \arg\min_{d_E} \sum_i \left(d_E\, I_{DD(D),i} - I_{DA(D),i}\right)^2, \end{aligned} \tag{150}$$

where subscript i indicates different cells. The values we obtained were $a_{E,est} =$ $0.3369\ (\pm 0.0006)$, and $d_{E,est} = 0.0891\ (\pm 0.0001)$.

The parameter G_E quantifies the change in sensitized emission F_c per change in I_{DD} due to FRET, $G_E = \left|\frac{dF_c}{dI_{DD}}\right|$, which in principle can be measured by using a FRET strain expressing both donor and acceptor as [43]

$$G_E \equiv \frac{Q_AL_AS_At_{DA}}{Q_DL_DS_Dt_{DD}} \simeq \frac{F_c'}{I_{DD}^{post} - I_{DD}'}, \tag{151}$$

where I_{DD}^{post} is the intensity of donor fluorescence in the condition where the acceptor is completely photobleached while the donor remains intact, and F_c' and I_{DD}' correspond to F_c and I_{DD}, respectively, in the absence of photobleaching. Again, the equality is exact in the limit of zero measurement noise. The relation can be shown by noting

$$\begin{aligned} F_c' &= [XY]\nu_D\epsilon_{DD}E_{max}Q_AL_AS_At_{DA} + \xi_{F_c} \\ I_{DD}^{post} - I_{DD}' &= [XY]\nu_D\epsilon_{DD}E_{max}Q_DL_DS_Dt_{DD} + \xi_{DD}', \end{aligned} \tag{152}$$

where ξ_{F_c} and ξ_{DD}' represent the effects of shot noise. To avoid issues associated with acceptor photobleaching [44], we can take advantage of the fact that, in our system and microfluidic device, FRET changes can be induced rapidly by applying a step change of chemoattractant stimulus. Doing so lets us measure changes in F_c and I_{DD} before substantial photobleaching occurs, and we can estimate G from:

$$G_E \simeq \frac{|\Delta F_c|}{|\Delta I_{DD}|}. \tag{153}$$

We estimated the value of G_E by least-squares fitting the fluorescence signals from multiple cells (Supplementary Fig. S4C), i.e.,

$$G_{E,est} = \arg\min_{G_E} \sum_i \left(G_E\,|\Delta I_{DD,i}| - |\Delta F_{c,i}|\right)^2, \tag{154}$$

where $|\Delta I_{DD,i}|$ and $|\Delta F_{c,i}|$ for the i-th cell were obtained by applying a saturating stimulus (1 mM MeAsp and 1 µM serine) on top of a background stimulus 100 µM MeAsp, and then removing the stimulus and the background (i.e., 0 M chemoattractants) in the microfluidic chamber because it induces the maximal FRET change [26,41]. The obtained value was $G_{E,est} = 0.3497\ (\pm 0.0018)$.

Section 13: Effects of parameter-estimation error on the FRET signal

The FRET index E_{corr} computed from observables (Eqn. (131); see Supplementary Fig. S4D) provides an unbiased estimator of $\mathcal{E}$ (Eqn. (130)) in the presence of photobleaching, given the true values of the system-dependent parameters a_E, d_E and G_E. However, the parameters are always estimated with finite precision. Furthermore, although it is often assumed that the values are invariant given a system, they may not be constant over the course of measurements. It has been reported that some fluorescent proteins change their fluorescent properties upon photobleaching [44]. Zal and Gascoigne only explored how errors in the parameter estimates bias E_{corr} in the absence of photobleaching [43]. Here we study the effect of the parameter-estimation error in the presence of photobleaching and propose a method to correct for the effects.

The estimated values of the parameters can be written as

$$\begin{aligned} a_{E,est} &= a_E + \Delta a_E, \\ d_{E,est} &= d_E + \Delta d_E, \\ G_{E,est} &= G_E + \Delta G_E, \end{aligned} \tag{155}$$

where true values of the parameters are denoted by a_E, d_E, and G_E and the deviations from them by Δa_E, Δd_E, and ΔG_E. First, we note that E_{corr} can be approximated as

$$E_{corr} = \frac{\frac{F_c}{I_{DD}}}{\frac{F_c}{I_{DD}} + G_E} \frac{I_{AA}(0)}{I_{AA}} \tag{156}$$

$$\approx \frac{I_{AA}(0)}{G_E} \frac{F_c}{I_{DD} I_{AA}} \tag{157}$$

$$= \frac{I_{AA}(0)}{G_E} \frac{I_{DA} - d_E\, I_{DD} - a_E\, I_{AA}}{I_{DD} I_{AA}}. \tag{158}$$

In the second line, we used $\frac{F_c}{I_{DD}} \ll G_E$ to simplify the following calculation, but it is not essential. This assumption is valid in a typical low FRET-efficiency experiment where the value of F_c is sufficiently lower than I_{DD} $\left(\frac{F_c}{I_{DD}} \ll 1\right)$ and yet, to be able to detect FRET signals, the parameter G_E needs to be $\sim\mathcal{O}(1)$ [43]. In our setup, $\frac{F_c}{I_{DD}} \lesssim 0.05$ and $G_E \simeq 0.35$.

The error in E_{corr} due to the error in the estimated parameters Δa_E, Δd_E, and ΔG_E can be written as

$$\Delta E_{corr} = E_{corr}(a_E + \Delta a_E, d_E + \Delta d_E, G_E + \Delta G_E) - E_{corr}(a_E, d_E, G_E) \tag{159}$$

$$\simeq \frac{\partial E_{corr}(a_E, d_E, G_E)}{\partial a_E}\Delta a_E + \frac{\partial E_{corr}(a_E, d_E, G_E)}{\partial d_E}\Delta d_E + \frac{\partial E_{corr}(a_E, d_E, G_E)}{\partial G_E}\Delta G_E \tag{160}$$

$$\simeq -\frac{I_{AA}(0)}{G_E I_{DD}}\Delta a_E - \frac{I_{AA}(0)}{G_E I_{AA}}\Delta d_E - \frac{E_{corr}}{G_E}\Delta G_E. \tag{161}$$

Thus, the fraction of error in E_{corr} can be written as

$$\frac{\Delta E_{corr}}{E_{corr}} = -\frac{I_{AA}\Delta a_E}{F_c} - \frac{I_{DD}\Delta d_E}{F_c} - \frac{\Delta G_E}{G_E}. \tag{162}$$

From the assumptions about the molecular interactions and photobleaching that underly the E-FRET method, the observables I_{DD} and I_{AA} and the sensitized emission F_c can be written as

$$\frac{\Delta E_{corr}}{E_{corr}} = -\frac{I_{AA}\Delta a_E}{F_c} - \frac{I_{DD}\Delta d_E}{F_c} - \frac{\Delta G_E}{G_E}. \tag{163}$$

$$I_{DD} \simeq C_{DD}\left([X_{total}]\, e^{-\int_0^t \delta(t')dt'} - E_{max}\,[XY]\, e^{-\int_0^t \alpha(t')+\delta(t')dt'}\right) \sim C_{DD}\,[X_{total}]\, e^{-\int_0^t \delta(t')dt'} \tag{164}$$

$$I_{AA} \simeq C_{AA}\,[Y_{total}]\, e^{-\int_0^t \alpha(t')dt'} \tag{165}$$

$$F_c(t) \simeq C_{DD}\, G_E\, E_{max}\,[XY]\, e^{-\int_0^t \alpha(t')+\delta(t')dt'}, \tag{166}$$

where $\delta(t) > 0$ and $\alpha(t) > 0$ are, respectively, the (time-dependent) rates of photobleaching of the donor and acceptor, and the final approximation for I_{DD} is valid under the assumption $\frac{F_c}{I_{DD}} \ll G_E$.

Using these expressions, we get

$$\frac{\Delta E_{corr}}{E_{corr}} \sim A\, e^{\int_0^t \delta(t')dt'}\Delta a_E + D\, e^{\int_0^t \alpha(t')dt'}\Delta d_E - \frac{\Delta G_E}{G_E}, \tag{167}$$

where $A = \frac{C_{AA}\,[Y_{total}]}{C_{DD}\,G_E\,E_{max}\,[XY]} > 0$ and $D = \frac{[X_{total}]}{G_E E_{max}\,[XY]} > 0$. The first and the second terms grow quasi-exponentially as the fluorescent proteins photobleach; thus, the measured value of E_{corr}, at baseline levels of molecular interaction, changes over time. Note that the time scale of this change is governed by the time scale of photobleaching. We discuss how we corrected the baseline of E_{corr} in the next section below.

The remaining question is how uncertainty in the parameters a_E, d_E, and G_E, in the presence of photobleaching, affects the mapping between *changes* in molecular interactions and the corresponding *change* in E_{corr}. If this mapping is sensitive to photobleaching, it severely limits the reliability of the E_{corr} measurements because the same changes in molecular interactions would lead to different changes in E_{corr} at different times in the experiment. To address this, we analyze the sensitivity of E_{corr} to the change in the degree of molecular interaction and its dependence on photobleaching.

The degree of molecular interaction is dictated by the time-dependent binding affinity $\gamma(t)$ between the two target molecules X and Y. Therefore, the sensitivity of E_{corr} to changes in γ at a given time can be quantified by $\frac{\partial E_{corr}(\gamma|a_E,d_E,G_E)}{\partial\gamma}$. With errors in the parameters, this quantity can be written as

$$\frac{\partial E_{corr}(\gamma|a_E+\Delta a_E, d_E+\Delta d_E, G_E+\Delta G_E)}{\partial\gamma} = \frac{\partial E_{corr}(\gamma|a_E,d_E,G_E)}{\partial\gamma} + \frac{\partial\Delta E_{corr}(\gamma)}{\partial\gamma} \tag{168}$$

$$= \frac{\partial E_{corr}(\gamma|a_E,d_E,G_E)}{\partial\gamma}\left(1 + \frac{\frac{\partial\Delta E_{corr}(\gamma)}{\partial\gamma}}{\frac{\partial E_{corr}(\gamma|a_E,d_E,G_E)}{\partial\gamma}}\right) \tag{169}$$

$$\equiv \frac{\partial E_{corr}(\gamma|a_E,d_E,G_E)}{\partial\gamma}(1+\Delta).$$

Thus, $\Delta \equiv \frac{\frac{\partial\Delta E_{corr}}{\partial\gamma}}{\frac{\partial E_{corr}(\gamma|a_E,d_E,G_E)}{\partial\gamma}}$ characterizes the bias error, and the question is how this quantity behaves with photobleaching. To compute this, we note

$$\Delta = \frac{\frac{\partial}{\partial\gamma}\left(-\frac{I_{AA}(0)}{G_E I_{DD}(\gamma)}\Delta a_E - \frac{I_{AA}(0)}{G_E I_{AA}}\Delta d_E - \frac{E_{corr}(\gamma)}{G_E}\Delta G_E\right)}{\frac{\partial E_{corr}(\gamma)}{\partial\gamma}} \tag{170}$$

$$= \frac{-\frac{I_{AA}(0)\Delta a_E}{G_E}\frac{\partial}{\partial\gamma}\left(\frac{1}{I_{DD}(\gamma)}\right) - \frac{\Delta G_E}{G_E}\frac{\partial E_{corr}}{\partial\gamma}}{\frac{\partial E_{corr}}{\partial\gamma}} \tag{171}$$

$$= \frac{-\frac{I_{AA}(0)\Delta a_E}{G_E}\frac{\partial}{\partial\gamma}\left(\frac{1}{I_{DD}(\gamma)}\right)}{\frac{\partial E_{corr}}{\partial\gamma}} - \frac{\Delta G_E}{G_E}, \tag{172}$$

where we used the fact that I_{AA} is independent of γ, i.e., $\frac{\partial I_{AA}}{\partial\gamma} = 0$. We note

$$\frac{\partial E_{corr}}{\partial\gamma} \simeq \frac{\partial}{\partial\gamma}\left(\frac{1}{G_E}\frac{F_c(\gamma)}{I_{DD}(\gamma)}\frac{I_{AA}(0)}{I_{AA}}\right) \tag{173}$$

$$= \frac{I_{AA}(0)}{G_E I_{AA}}\frac{\partial}{\partial\gamma}\left(\frac{F_c}{I_{DD}}\right) \tag{174}$$

$$= \frac{I_{AA}(0)}{G_E I_{AA}}\left(\frac{1}{I_{DD}}\frac{\partial F_c}{\partial\gamma} - \frac{F_c}{I_{DD}^2}\frac{\partial I_{DD}}{\partial\gamma}\right) \tag{175}$$

$$= -\frac{I_{AA}(0)}{G_E I_{AA}}\frac{1}{I_{DD}}\frac{\partial I_{DD}}{\partial\gamma}\left(-\frac{\frac{\partial F_c}{\partial\gamma}}{\frac{\partial I_{DD}}{\partial\gamma}} + \frac{F_c}{I_{DD}}\right) \tag{176}$$

$$= -\frac{I_{AA}(0)}{G_E I_{AA}}\frac{1}{I_{DD}}\frac{\partial I_{DD}}{\partial\gamma}\left(G_E + \frac{F_c}{I_{DD}}\right), \tag{177}$$

where at the final step we used $-\frac{\partial F_c}{\partial\gamma}/\frac{\partial I_{DD}}{\partial\gamma} = |\Delta F_c|/|\Delta I_{DD}| = G_E$. By plugging this to the expression for Δ, we get

$$\Delta = \frac{\frac{I_{AA}(0)\Delta a_E}{G_E}\frac{1}{I_{DD}^2}\frac{\partial I_{DD}}{\partial\gamma}}{-\frac{I_{AA}(0)}{G_E I_{AA}}\frac{1}{I_{DD}}\frac{\partial I_{DD}}{\partial\gamma}\left(G_E + \frac{F_c}{I_{DD}}\right)} - \frac{\Delta G_E}{G_E} \tag{178}$$

$$= \frac{-I_{AA}\Delta a_E}{I_{DD}\left(G_E + \frac{F_c}{I_{DD}}\right)} - \frac{\Delta G_E}{G_E} \tag{179}$$

$$\simeq -\frac{I_{AA}\,\Delta a_E}{I_{DD}\,G_E} - \frac{\Delta G_E}{G_E} \tag{180}$$

$$= H\,e^{\int_0^t \delta(t')-\alpha(t')dt'}\Delta a_E - \frac{\Delta G_E}{G_E}, \tag{181}$$

where $H = \frac{C_{AA}\,[Y_{total}]}{C_{DD}\,[X_{total}]G_E} > 0$ and in the third line we used $\frac{F_c}{I_{DD}} \ll G_E$. This expression tells us that the relative error in the mapping from molecular interaction to E_{corr}, Δ, is small if Δa_E and ΔG_E are small. Furthermore, this relative error grows slower than the relative error in the baseline of E_{corr}, $\frac{\Delta E_{corr}}{E_{corr}}$, because only the difference of the donor and acceptor photobleaching rates appears in the exponential. Additionally, the coefficient H is typically smaller than the coefficients in

$\frac{\Delta E_{corr}}{E_{corr}}$, A and D. In fact, assuming $C_{AA} \approx C_{DD}$ and $[Y_{total}] \approx [X_{total}]$, one can show that both H/A and H/D are bounded by $\frac{E_{max}\,[XY]}{[X_{total}]} < 1$.

In summary, the estimator of the degree of molecular interaction E_{corr} computed from the observables has the following properties due to uncertainties in estimating the system parameters: (i) the baseline level of E_{corr} changes on the timescale of fluorescence photobleaching, even if there is no change in the level of molecular interactions; and (ii) biologically-induced changes in molecular interactions are relatively well-preserved in changes in E_{corr}. This means that although the baseline level of E_{corr} changes over time, deviations from that baseline faithfully represent changes in molecular interactions.

We tested whether these properties are observed in actual FRET data by investigating both the absolute levels of and changes in E_{corr} before and after photobleaching (Supplementary Fig. S4D-G). Under the assumption that each cell retains approximately identical FRET-response properties over the course of an experiment, our analyses predict that the baseline level of E_{corr} monotonically changes over time, while the changes in E_{corr} upon stimuli remain approximately invariant over the course of a measurement. Consistent with these predictions, we observed: (i) the absolute level of E_{corr} slowly increased over time (Supplementary Fig. S4E) and the values after photobleaching were higher than those before photobleaching (Supplementary Fig. S4F), although the degree of the change was moderate (Supplementary Fig. S4F; roughly 15% increase after >1200 frames of image acquisition); (ii) the changes in E_{corr} induced by fast-switching, identical stimuli showed essentially no bias after photobleaching (Supplementary Fig. S4G). Thus, the behavior of our FRET data due to photobleaching can be consistently explained by the properties of E_{corr}.

Section 14: Correcting baseline FRET signal

The analyses above have established that the baseline level of E_{corr} monotonically changes on the timescale of photobleaching because of parameter uncertainties, but changes in E_{corr} due to fast biological FRET interactions are more reliable. This suggests that the slow artifact in the baseline level E_{corr} can be corrected for by estimating and subtracting the slowly-varying trend. We obtained a corrected signal $\overline{E_{corr}}$ in the following way. First, since the drift in baseline E_{corr} is slow compared to the durations of our experiments (Supplementary Fig. S4EF), we assumed that the error in E_{corr} changes approximately linearly with frame number. The slope of the linear function was estimated by measuring the minimum values of E_{corr} at the beginning (frame number i_1) and the end (frame number i_2) of each measurement by applying a saturating stimulus, and by fitting a linear function to $\{(i_1, E_{corr}(i_1)), (i_2, E_{corr}(i_2))\}$. That is, the corrected E_{corr} is written as

$$\overline{E_{corr}}(i) = E_{corr}(i) - \frac{E_{corr}(i_2) - E_{corr}(i_1)}{i_2 - i_1} i, \tag{182}$$

where $E_{corr}(i_1)$ and $E_{corr}(i_2)$ were estimated as medians of 18 consecutive frames to mitigate the effect of measurement noise. With this correction, we observed population-average FRET signals that show essentially no trends, only stimulus-induced changes (see, e.g., Supplementary Figs. S5A and S6A), validating the assumption of the linearity of the error-induced trend.

Section 15: Converting FRET signal to kinase activity

Using the corrected FRET signal $\overline{E_{corr}}$, we defined the kinase activity as

$$a(t) = \frac{\overline{E_{corr}}(t) - \overline{E_{corr}^{min}}}{\overline{E_{corr}^{max}} - \overline{E_{corr}^{min}}}, \tag{183}$$

where $\overline{E_{corr}^{min}}$ and $\overline{E_{corr}^{max}}$ are respectively the minimum and maximum values of $\overline{E_{corr}}$ obtained from responses to a saturating stimulus and following removal of background at the beginning of each measurement (see Methods). Note that the mutual information rate between chemoattractant signal $s(t)$ and the kinase output of the chemotaxis signaling pathway, which is the ultimate goal of our FRET analysis, is invariant to this linear conversion—we could have computed it from, e.g., $\overline{E_{corr}}(t)$. However, we computed $a(t)$ to facilitate comparison with preceding works where the same normalized measure was used [15,26]. The distributions of the steady-state values of a, a_0, across cells in isogenic populations were evaluated under two different stimulus and illumination conditions, giving: $a_0 = 0.29 \pm 0.07$ (mean and standard deviation, estimated from the experiments for response-function extraction; Supplementary Fig. S5C), and $a_0 = 0.30 \pm 0.08$ (mean and standard deviation, estimated from the experiments for signaling-noise characterization; Supplementary Fig. S6C). These values are comparable to previously reported values: $a_0 \approx 1/3$ from population-averaged FRET measurements [15] and $a_0 = 0.30 \pm 0.07$ (mean and standard deviation) from single-cell FRET measurements [26].

Biochemical interpretations of the kinase activity $a(t)$ are the following. First, from Eqn. (130), we write [41,42]

$$\overline{E_{corr}}(t) \approx \frac{E_{max}\,[\text{CheYp} \cdot \text{CheZ}]}{[\text{CheZ}_{\text{total}}]}, \tag{184}$$

where the equality holds in the limit of zero measurement noise. Using this expression, a can be written as

$$a(t) \approx \frac{[\text{CheYp} \cdot \text{CheZ}](t) - [\text{CheYp} \cdot \text{CheZ}]^{min}}{[\text{CheYp} \cdot \text{CheZ}]^{max} - [\text{CheYp} \cdot \text{CheZ}]^{min}}, \tag{185}$$

and therefore $a(t)$ is primarily the relative degree of CheYp-CheZ formation. On time scales longer than the time scale of CheY-p hydrolysis by CheZ (~0.3 s [4]), phosphorylation and dephosphorylation of CheY equilibrate [41,42], and therefore

$$[\text{CheYp} \cdot \text{CheZ}](t) = A(t)\frac{k_A}{k_Z}[\text{CheA}] \approx A(t)\frac{k_A}{k_Z}[\text{CheA}_{total}], \tag{186}$$

where k_A and k_Z are, respectively, the rate constants for autophosphorylation of CheA and for hydrolysis of CheY-p by CheZ, $A(t)$ $(0 < A(t) < 1)$ is the fraction of active CheA, and $[\text{CheA}_{total}]$ is the total concentration of CheA. Given the conservation equation $[\text{CheA}_{total}] = [\text{CheA}] + [\text{CheAp}]$, the last step of the above equations holds when $[\text{CheAp}] \ll [\text{CheA}]$. This is achieved when the total concentrations of CheY-mRFP and CheZ-mYFP in the cell are large, which we achieve by overexpressing them in our experimental conditions (see Methods), as done before [26]. Assuming this, a can be written as

$$a(t) \approx \frac{A(t) - A^{min}}{A^{max} - A^{min}}, \tag{187}$$

and therefore $a(t)$ can also be interpreted as the relative autophosphorylation activity of the kinase CheA, which we have referred to as kinase activity.

Section 16: Estimating linear response function parameters

A kinase activity time series from a single cell (labeled by k) consisted of 10 step-up responses $a_{k,+,i}(t)$ $(i = 1,2 \ldots 10)$, where the concentration of MeAsp was changed from $c_0 = 100$ µM to $c_+ = 110$ µM, and 10 step-down responses $a_{k,-,i}(t)$, where the concentration was changed from c_0 to $c_- = 90$ µM (Fig. 2D, and Supplementary Fig. S5A). Each response $a_{k,s,i}(t)$ $(s = \{+,-\}$ for shorthand) consisted of a pre-stimulus measurement (7.5 s; 10 time points) and post-stimulus measurement (30 s; 40 time points) of kinase output, followed by a 60 second interval before the next step-response measurement (Fig. 2D, and Supplementary Fig. S5A; Methods). The kinase response induced by an impulse of stimulus (step change in concentration) was thus defined as

$$\Delta a_{k,s,i}(t) = a_{k,s,i}(t) - \langle a_{k,s,i}(t) \rangle_{prestim}, \tag{188}$$

where $\langle a_{k,s,i}(t) \rangle_{prestep}$ is the time-averaged kinase activity in the moments before the step change of concentration was delivered. The steady-state kinase activity of the k-th cell $a_{0,k}$ was estimated by the average of the pre-stimulus kinase activities $\langle a_{k,s,i}(t) \rangle_{prestep}$, i.e.,

$$a_{0,k} = \frac{1}{2\, N_{stim}} \sum_{s=\{+,-\}} \sum_{i=1}^{N_{stim}} \langle a_{k,s,i}(t) \rangle_{prestim}, \tag{189}$$

where $N_{stim} = 10$ is the number of stimuli of each sign delivered to the cells. The average and the standard error of the kinase activity responses in each cell were computed as

$$\langle \Delta a_{k,s} \rangle(t) = \frac{1}{N_{stim}} \sum_{i=1}^{N_{stim}} \Delta a_{k,s,i}(t)\,, \quad SE_{k,s}(t) = \frac{1}{\sqrt{N_{stim}}} \sqrt{\frac{1}{N_{stim}-1} \sum_{i=1}^{N_{stim}} \left(\Delta a_{k,s,i}(t) - \langle \Delta a_{k,s} \rangle(t) \right)^2}\ . \tag{190}$$

To each single cell average response, we fitted the following function with 3 parameters, G, τ_1, and τ_2:

$$f_s\left(t \middle| G_k, \tau_{1,k}, \tau_{2,k}\right) = -\log\left(\frac{c_s}{c_0}\right) K(t - t_0), \tag{191}$$

where $K(t)$ is the parameterized impulse-response function discussed earlier and t_0 denotes the timing at which the stimulus levels were changed. $f_s\left(t \middle| G_k, \tau_{1,k}, \tau_{2,k}\right)$ is the result of convolving the response function $K(t)$ with a delta function of signal $s(t)$ with amplitude $\log\left(\frac{c_s}{c_0}\right)$ at time t_0, or $s(t) = \frac{d}{dt}\log(c) = \log\left(\frac{c_s}{c_0}\right)\delta(t_0)$. The minus sign is needed because positive signals lead to drops in kinase activity. The time of each stimulus t_0 was inferred from the data in the following way. For each signal, the first time point at which the population response was more than 3 standard deviations below baseline $a_{0,k}$ was found. Then, t_0 for that signal was defined as the time half-way between that time point and the previous one. This was repeated for each stimulus. The same values of t_0 were used for all cells.

We fit this function to both the step-up and step-down responses simultaneously, since in the linear-response regime the up and down responses are expected to be symmetric. The measurements of $\langle \Delta a \rangle(t)$ were not smoothed before fitting. The fitting was done using a Bayesian framework [47]. Log-uniform distributions were used as priors for each parameter:

$$\begin{gathered} P\left(\tau_{1,k}\right) = \mathcal{U}\left(\tau_{1,k}; 10^{-2}, 5\right) \\ P\left(\tau_{2,k}\right) = \mathcal{U}\left(\tau_{2,k}; 10^{-1}, 10^{2}\right) \\ P(G_k) = \mathcal{U}\left(G_k; 10^{-1}, \frac{a_{0,k}}{\log\left(\frac{c_+}{c_0}\right)}\right), \end{gathered} \tag{192}$$

where $\mathcal{U}(x; a, b) = \frac{1}{x \log b/a}$ for $a \leq x \leq b$, and $\mathcal{U}(x; a, b) = 0$ otherwise. Note the upper bound for $P(G_k)$ comes from the positivity of the kinase activity $a_{k,s,i}(t) > 0$. The log posterior distribution is defined as

$$\log P\left(G_k, \tau_{1,k}, \tau_{2,k} \middle| \mathcal{D}\right) = -\frac{1}{2}\sum_{n,s} \frac{\left(\langle \Delta a_{k,s}\rangle(t_n) - f_s\left(t_n \middle| G_k, \tau_{1,k}, \tau_{2,k}\right)\right)^2}{SE_{k,s}^2(t_n)} + \log\left(P(G_k)\right) + \log\left(P\left(\tau_{1,k}\right)\right) + \log\left(P\left(\tau_{2,k}\right)\right) + C, \tag{193}$$

where the first term on the right-hand side is the log-likelihood function and C comes from a normalization constant. A maximum a posteriori probability (MAP) estimate, defined as the mode of the posterior distribution was obtained using a MATLAB optimization function (*fminunc*) (Supplementary Fig. S5C). As a measure for the parameter-estimation uncertainty, we computed 25 and 75 percentiles of the marginalized posterior distribution of each parameter ($P(G_k|\mathcal{D})$, $P\left(\tau_{1,k}|\mathcal{D}\right)$, and $P\left(\tau_{2,k}|\mathcal{D}\right)$) obtained by a Markov chain Monte Carlo algorithm, slice sampling [47].

The population-level representative value of each parameter was defined as the median of the MAP estimates, and the uncertainty of the value was evaluated using the median absolute deviations (MAD) from the population's median parameter values:

$$\mathrm{SE} = \frac{1.4826 \times \mathrm{median}(|X_i - \mathrm{median}(X)|)}{\sqrt{N_{Data}}}, \tag{194}$$

where $X = \{X_1, X_2, \dots, X_{N_{Data}}\}$ is the MAP estimates of each parameter. The MAD in the numerator equals the standard deviation of X when it is Gaussian distributed, but is robust to outliers.

Section 17: Estimating noise statistics parameters

One of the first direct measurements of the fluctuation of the chemotaxis signaling pathway, or signaling noise, showed that the dynamics approximately follow an Ornstein-Uhlenbeck (OU) process. The OU process is specified by two parameters, τ_n and D_n, and obeys the following Langevin equation

$$\frac{dx}{dt} = -\frac{1}{\tau_n}x + \sqrt{2\,D_n}\,\xi(t) \tag{195}$$

where $\xi(t)$ is a Gaussian white noise with average zero and a delta correlation in time

$$\langle \xi(t) \rangle = 0, \qquad \langle \xi(t)\xi(t') \rangle = \delta(t - t'), \tag{196}$$

and $\delta(t)$ is the Dirac delta function.

Following preceding work [26], we modeled the signaling noise measured by FRET (Supplementary Fig. S6A) as an OU process, and estimated the process parameters (and their uncertainties) using a Bayesian filtering-based method [48]. To introduce some notation, the measurements were conducted at time points $t_1, t_2, \dots, t_T$, and we obtained a series of observables (i.e., FRET signals) $y_{1:T} = \{y_1, y_2, \dots, y_T\}$, which are noise-corrupted measurements of the true, hidden state of the system $x_{1:T} = \{x_1, x_2, \ldots, x_T\}$, which obeys Eqn. (195) with unknown parameters $\boldsymbol{\theta} = \{\tau_n, D_n\}$. Using the framework of a state-space model, the parameter-estimation problem can be written in the form

$$\begin{gathered} \boldsymbol{\theta} \sim p(\boldsymbol{\theta}) \\ x_k \sim p(x_k | x_{k-1}, \boldsymbol{\theta}) \\ y_k \sim p(y_k | x_k) \end{gathered} \tag{197}$$

where $p(\boldsymbol{\theta})$ is the prior distribution of the parameters $\boldsymbol{\theta}$, $p(x_k | x_{k-1}, \boldsymbol{\theta})$ the transition probability distribution specified by the dynamical model, and $p(y_k | x_k)$ is the measurement model.

Our goal is to evaluate the posterior distribution of the parameters $\boldsymbol{\theta}$ given the data $y_{1:T}$, $p(\boldsymbol{\theta} | y_{1:T})$. Using Bayes' rule, this can be written as

$$p(\boldsymbol{\theta} | y_{1:T}) \propto p(y_{1:T} | \boldsymbol{\theta})\, p(\boldsymbol{\theta}), \tag{198}$$

where $p(y_{1:T} | \boldsymbol{\theta})$ is the likelihood function. The likelihood function can be written as

$$p(y_{1:T} | \boldsymbol{\theta}) = \prod_{k=1}^{T} p(y_k | y_{1:k-1}, \boldsymbol{\theta}) = \prod_{k=1}^{T} \int p(y_k | x_k) p(x_k | y_{1:k-1}, \boldsymbol{\theta}) dx_k, \tag{199}$$

where we define $p(y_1 | y_{1:0}, \boldsymbol{\theta}) \equiv p(y_1 | \boldsymbol{\theta})$. The predictive distribution of the state $p(x_k | y_{1:k-1}, \boldsymbol{\theta})$ can be written as

$$p(x_k | y_{1:k-1}, \boldsymbol{\theta}) = \int p(x_k | x_{k-1}, \boldsymbol{\theta}) p(x_{k-1} | y_{1:k-1}, \boldsymbol{\theta}) dx_{k-1}. \tag{200}$$

The posterior distribution of the hidden state in the integral $p(x_{k-1} | y_{1:k-1}, \boldsymbol{\theta})$ can be written, using Bayes' rule, as

$$p(x_k | y_{1:k}, \boldsymbol{\theta}) = \frac{p(y_k | x_k) p(x_k | y_{1:k-1}, \boldsymbol{\theta})}{p(y_k | y_{1:k-1}, \boldsymbol{\theta})}, \tag{201}$$

where we use $p(y_k | x_k, y_{1:k-1}) = p(y_k | x_k)$. Given the predictive distribution of the state at the initial time point $p(x_1 | \boldsymbol{\theta})$, using the Eqns. (200) and (201) recursively, the predictive distributions at the following time points $\{t_k\}$ can be computed, which then gives the likelihood function using Eqn. (199).

Under the assumption that the sequential states $x_{1:T}$ follow an OU process with parameters $\boldsymbol{\theta}$, we can write the transition probability distribution as

$$p(x_k|x_{k-1},\boldsymbol{\theta}) = \mathcal{N}(x_k; x_{k-1}v, V_x), \tag{202}$$

where $v \equiv e^{-\frac{t_k-t_{k-1}}{\tau_n}}$, $V_x \equiv D_n\,\tau_n\,(1-v^2)$, and $\mathcal{N}(x;\mu,\sigma^2)$ is a Gaussian distribution with mean μ and variance σ^2. Also, assuming that the measurement noise follows the Gaussian with zero mean and variance σ_m^2, we can write the measurement model as

$$p(y_k|x_k) = \mathcal{N}(y_k; x_k, \sigma_m^2). \tag{203}$$

Under these assumptions, one can find closed-form expressions for Eqns. (199)-(201) [48]. Specifically, the likelihood function (Eqn. (199)) can be written as

$$\begin{aligned} p(y_{1:T}|\boldsymbol{\theta}) &= \prod_{k=1}^{T}\int p(y_k|x_k)p(x_k|y_{1:k-1},\boldsymbol{\theta})dx_k \\ &= \prod_{k=1}^{T}\mathcal{N}\big(y_k;\mu(x_k|\boldsymbol{\theta}),\sigma_m^2+V(x_k|\boldsymbol{\theta})\big)\,, \end{aligned} \tag{204}$$

where $\mu(x_k|\boldsymbol{\theta})$ and $V(x_k|\boldsymbol{\theta})$ are the mean and variance of the predictive distribution $p(x_k|y_{1:k-1},\boldsymbol{\theta})$. Evaluating Eqn. (200), the predictive distribution is written as

$$p(x_k|y_{1:k-1},\boldsymbol{\theta}) = \mathcal{N}\big(x_k;\mu(x_k|\boldsymbol{\theta}),V(x_k|\boldsymbol{\theta})\big) \tag{205}$$

$$= \int p(x_k|x_{k-1},\boldsymbol{\theta})p(x_{k-1}|y_{1:k-1},\boldsymbol{\theta})dx_{k-1} \tag{206}$$

$$= \mathcal{N}\big(x_k; v\mu'(x_{k-1}|\boldsymbol{\theta}), V_x + v^2V'(x_{k-1}|\boldsymbol{\theta})\big), \tag{207}$$

where $\mu'(x_{k-1}|\boldsymbol{\theta})$ and $V'(x_{k-1}|\boldsymbol{\theta})$ are the mean and variance of the posterior distribution of the state x_{k-1}, $p(x_{k-1}|y_{1:k-1},\boldsymbol{\theta})$. Evaluating Eqn. (201), the posterior distribution $p(x_k|y_{1:k},\boldsymbol{\theta})$ is written as

$$p(x_k|y_{1:k},\boldsymbol{\theta}) \equiv \mathcal{N}\big(x_k;\mu'(x_k|\boldsymbol{\theta}),V'(x_k|\boldsymbol{\theta})\big) \tag{208}$$

$$= \frac{p(y_k|x_k)p(x_k|y_{1:k-1},\boldsymbol{\theta})}{p(y_k|y_{1:k-1},\boldsymbol{\theta})} \tag{209}$$

$$= \;\mathcal{N}\left(x_k; v\frac{y_{k-1}V(x_k|\boldsymbol{\theta})+\mu(x_k|\boldsymbol{\theta})\sigma_m^2}{\sigma_m^2+V(x_k|\boldsymbol{\theta})},\frac{\sigma_m^2V(x_k|\boldsymbol{\theta})}{\sigma_m^2+V(x_k|\boldsymbol{\theta})}\right), \tag{210}$$

which is dependent on the mean and variance of the predictive distribution $p(x_k|y_{1:k-1},\boldsymbol{\theta})$. As the predictive distribution at the initial time point, $p(x_1|\boldsymbol{\theta})$, we chose

$$p(x_1|\boldsymbol{\theta}) = \delta(x_1 - y_1). \tag{211}$$

In the above formulation, we assumed that the variance of the measurement noise σ_m^2 was known. In our case, since the measurement noise is dominated by the shot noise of fluorescent signals, even after the photobleaching correction, σ_m^2 can slowly increase as more fluorescent proteins photobleach over the course of measurement, $\sigma_m^2 = \sigma_m^2(t_k)$. To estimate $\sigma_m^2(t_k)$, we first segmented the FRET time series into short 12 segments of identical length (100-sec interval with 1 sec sampling interval) and estimated the measurement noise in each segment $\sigma_{m,i}^2$ assuming that the noise level is approximately constant within the segment. The measurement noise at each segment was estimated by

$$\sigma_{m,i}^2 = C(\tau = 0) - A\, e^{-\frac{\tau}{B}}|_{\tau=0}, \tag{212}$$

where $C(\tau)$ is the autocorrelation function computed from the (mean-subtracted) raw FRET time series and $A\, e^{-\frac{\tau}{B}}$ is a fit to $C(\tau)$ over the range of $0 < \tau \leq 10$ s (note we exclude the zero lag time point $\tau = 0$). This gives an estimation of the variance of shot noise because $C(\tau = 0)$ estimates the sum of the shot noise and biological noise at lag time zero, while $A\, e^{-\frac{\tau}{B}}|_{\tau=0}$ estimates the biological noise at lag time zero. In fitting $A\, e^{-\frac{\tau}{B}}$ to $C(\tau)$, we only used the first 10 seconds because the statistical uncertainty of $C(\tau)$ is relatively large for $\tau > 10$ s. Once the measurement noise of each segment is obtained, we then estimated the measurement noise level at each time point $\sigma_m^2(t_k)$ by fitting a linear function $\sigma_m^2(t_k) = a\, t_k + b$ to $\left(\{\bar{t}_i\}, \{\sigma_{m,i}^2\}\right)$ (segments $i = 1, \dots, 12$), where $\bar{t}_i$ is the center of the time interval of segment i.

The log posterior distribution $\log p(\boldsymbol{\theta}|y_{1:T}) = \log p(y_{1:T}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) + \mathrm{Const.}$ was approximated by using a Markov chain Monte Carlo method, a slice sampling (*45*), and the mode (i.e., MAP estimate) and 25 and 75 percentiles of the posterior distribution of each parameter was estimated (Supplementary Fig. S6C). The prior distributions used were

$$\begin{gathered} P(\tau_n) = \mathcal{U}_{lin}(\tau_n; 0, 10^2) \\ P(D_n) = \mathcal{U}_{lin}(D_n; 0, 10^{-2}) \end{gathered} \tag{213}$$

where $\mathcal{U}_{lin}(x; a, b) = \frac{1}{b-a}$ for $a \leq x \leq b$ and $\mathcal{U}_{lin}(x; a, b) = 0$ otherwise. The autocorrelation function of the Ornstein-Uhlenbeck process $C_{OU}(\tau) = D_n\, \tau_n\, e^{-\tau/\tau_n}$ with the MAP estimates of the parameters τ_n and D_n closely match the autocorrelation function $C(\tau)$ directly computed from the FRET time series. The representative value of each parameter in the population and its uncertainty was evaluated in the same way as those for the signaling-response parameters.

Section 18: Cell detection

Movies of swimming cells were recorded to measure their behavioral parameters, run speed, rotational diffusion coefficient, and chemotactic drift in varying gradients. All but the chemotactic drift measurements were recorded with a 4X objective, whereas the chemotaxis measurements were recorded with a 10X objective (Methods). All movies were recorded at 20 frames per second by phase-contrast imaging for 1000 seconds. All of the steps and parameters used below were applied to trajectories recorded with both the 10X and 4X objectives, unless noted otherwise. In the chemotaxis experiments, each movie's frames were rotated slightly to align the x-axis with the direction of increasing concentration.

Movies were analyzed using custom MATLAB code. To detect cells in each frame of the movie, first, 5-second blocks of frames were loaded at a time, and the pixel-wise median of those frames was computed as a background image. The background image was then subtracted from each frame in the block. Most of the pixels did not contain a cell; therefore, the histogram of pixel intensities was fit in the vicinity of its peak with a Gaussian distribution to extract the mean and standard deviation of the noise pixel intensities. Since cells appear as dark spots in phase contrast imaging, each frame was subtracted from the background image, making cells become bright spots in the background-adjusted images. Each background-adjusted frame was then smoothed with a weak Gaussian filter, the standard deviation of which depended on the objective used for imaging. The pixel size of the 10X objective is ~0.65 μm, and the filter's standard deviation was set to 1 pixel; The pixel size of the 4X objective is ~1.62 μm, and the filter's standard deviation was set to 0.4 pixels. All pixel intensities in the block that were less than 2 noise standard deviations above the noise mean were set to zero. This left patches of non-zero pixels in each frame, most of which contained cells. Patches consisting of one non-zero pixel were set to zero intensity, because cells typically occupy several pixels. Patches in which the highest-intensity pixel was less than 25 (or 15) noise standard deviations above the noise mean at 10X (at 4X) were also set to zero, only leaving pixel patches that contained cells in focus. In case more than one cell fell within a pixel patch, we used MATLAB's *imregionalmax* function to find local intensity peaks after smoothing each frame again with a Gaussian filter of standard deviation 0.5 pixels, and each local intensity maximum was considered a different cell. Finally, the cells' positions were refined to sub-pixel resolution using the method in [49].

Section 19: Cell tracking

Linking cell detections into tracks was also done using custom MATLAB code. Before linking, detections were pruned for quality. If two detections occurred within $v_0 \, \Delta t$ of each other, both detections were removed. For tracking purposes, v_0 was set to 50 μm/s and $\Delta t = 50 \text{ ms}$ was the time between frames. In some movies, a cell became immobilized and spun in place, causing it to still be visible after background subtraction. Therefore, pixels in each movie with >15 times more detections than the median number of detections per pixel (among pixels that contained detections) were identified. Detections within 7 (or 3) pixels of these high-detection pixels at 10X (at 4X) were removed.

As a first pass, starting from the first frame, for each detection in the frame, the closest detection in the next frame within a distance of $r = v_0 \, \Delta t$ was found. Then, for each detection in frame i, the closest detection in frame $i + 1$ within distance r was considered the same cell. This process was repeated for each cell in the frame, and then for all frames, in chronological order. After this, tracks that lasted only one frame were considered false positive detections and were removed.

This assignment procedure could leave gaps in the track of a given cell if it disappeared from view for more than one frame, for example by swimming vertically out of the depth of field. Linking tracks across these gaps increases the average time we observe a given cell, allowing us to better estimate its behavioral parameters or drift speed. To close these gaps, we looped over tracks from longest to shortest for a given gap size k, in frames. For each track, we searched for tracks that started k frames after (before) the current one within a distance $r(k) = \min\,(v_0 \, \Delta t \, k, \sqrt{D \, \Delta t \, k})$ of the current track's end (start). For tracking, we set $D = 300 \text{ μm}^2/\text{s}$. If nearby tracks were found, the closest ones were linked to the current track. The loop then continues to the next track. Tracks for which links were made are then revisited. Once no more links were made at gap size k, the gap size was increased by 1 frame. This process was repeated up to a maximum of 10 frames (0.5 s). Following this procedure, tracks that lasted only two frames were considered false positives and were removed.

When two cells came close to each other, it was possible that we misassigned their identities after that event. To try to identify and "untangle" these crossings, we first found all events in which two trajectories came within 5 μm of each other. We then generated the hypothetical trajectories resulting from swapping the assignments of the trajectories for all time after the event. Splines were fit to the original two trajectories and the hypothetical trajectories within ± 0.2 s of the event. Then, the average acceleration magnitude along the splines of the original trajectories and along those of the hypothetical trajectory assignments were computed. If the average acceleration of the hypothetical trajectory assignments was less than 90% of the original ones, we performed the swap—doing so preferred trajectory assignments with fewer sharp turns, which could indicate a misassignment, but had a slight preference against making swaps. This process identified sharp turns when two trajectories came close, which could indicate two trajectories crossing each other almost perpendicularly but that were misassigned, and then it attempted to correct those misassignments. This process was repeated for all events when two trajectories came close, in chronological order. The same cell's trajectory could be swapped with other trajectories at multiple points in the movie.

Trajectories were smoothed by convolving their x- and y-positions, separately, with a Gaussian in time of width $\Delta t/2$. Velocities in the x and y directions were computed using first-order central differences: $v_x(t_i) = \big(x(t_i + \Delta t) - x(t_i - \Delta t)\big)/(2\,\Delta t)$ (likewise for $v_y(t_i)$). The swimming speeds projected onto the x-y plane (the z-direction being the objective point of view) at each time point were computed from the magnitudes of the velocity vectors at that time. Projected angular velocities were computed by calculating the angle between the vector incoming and outgoing at each position—i.e. the angle between $\big(x(t_i) - x(t_{i-1}), y(t_i) - y(t_{i-1})\big)$ and $\big(x(t_{i+1}) - x(t_i), y(t_{i+1}) - y(t_i)\big)$—and divided by Δt. These were then corrected so that the branch cuts at π and $-\pi$ did not lead to artificially large angular velocities.

Section 20: Tumble detection

Tumbles were detected using a variant of the algorithm introduced by Masson et al. [11]. We did not use a speed criterion to detect tumbles, only a criterion based on the projected angular speed. This is because we only see a projection of the cell's trajectory; therefore, if rotational diffusion turns the cell's heading towards or away from the viewing direction, artificial speed variations could be observed. Angular speed is still useful for detecting tumbles because the likelihood of a tumble not causing a change of heading in the x- or y-direction is very small. Briefly, we found each peak in the angular speed of a cell's trajectory and its surrounding troughs in angular speed. The peak angular speed of a tumble had to be greater than $25\ \mathrm{rad/s}$, and the troughs of angular speed surrounding a tumble had to be less than $15\ \mathrm{rad/s}$. For any points within this segment of time to be considered a tumble, the cumulative absolute change in angle from trough to trough had to be larger than $\beta\,\sqrt{D_r\,\Delta t}$, where $\beta = 4$, $D_r = 0.1\ \mathrm{rad}^2/\mathrm{s}$ (only for tumble detection), and Δt was the time between the troughs. The beginning and end of the tumble were chosen to be the times when the angular speed fell below the smaller of 1/2 the peak angular speed or $15\ \mathrm{rad/s}$, and all points in between were set to the tumble state. Runs that lasted one frame were set to tumbles.

Section 21: Estimating behavioral parameters, run speed, and rotational diffusion

Behavioral parameters $\theta = \{\lambda_{R0}, \alpha, P_{run}\}$ and run speed v_0 were extracted from trajectories of wild-type (RP437) *E. coli* swimming in the absence of a gradient and recorded with a 4X objective. Our goal was to estimate the parameters of a typical cell. Previous work has shown that behavioral parameters in *E. coli* are correlated with their tumble bias [25,50,51], $TB = 1 - P_{run}$, and therefore correlated with P_{run}. Therefore, we binned tracks by P_{run} (bin size 0.02), and then computed average behavioral parameters within each bin. The average trajectory duration among all cells was 4.7 seconds, and the total trajectory time was 6.3×10^5 seconds. Cells in bins $P_{run} = 0$ and $P_{run} = 1$ were excluded; these were trajectories for which we could not infer P_{run}.

To estimate the median "run bias" P_{run} in the population, we first computed the single-cell run biases from $P_{run} = \frac{t_{run}}{t_{run}+t_{tumble}}$. Here, t_{run} and t_{tumble} are the total time that the cell spent in the run and tumble states, respectively. Then, we computed the time-weighted distribution of P_{run}, or the fraction of total trajectory time corresponding to cells of phenotype P_{run}. This is necessary to get an unbiased phenotype distribution because cells with high diffusivity (large P_{run} for example) enter *and* leave the field of view more frequently that cells with low diffusivity, whereas cells with low diffusivity spend more time in the field of view but enter less frequently. As a result, the raw counts of each phenotype over the whole movie are biased samples of the population's phenotype distribution, over-counting high-diffusivity phenotypes. However, the distribution of phenotypes in a given frame is unbiased. Therefore, one could average the distribution of phenotypes seen in each frame over all frames of the movie, but this is equivalent to weighting each phenotype by the total time it was observed.

The swimming speed during runs v_0 was computed by taking the average of the speed during runs among cells with P_{run} near the population-median value, with each cell given weight proportional to its duration. This estimate alone would be biased by our projected viewpoint. Assuming that in shallow gradients the headings are uniformly distributed over all possible directions, v_0 is expected to be underestimated by a factor of $\pi/4$. Therefore, to correct for the projection bias, we multiplied the computed value of v_0 by $4/\pi$. This v_0 was used to compute v_d/v_0 in Fig. 3C of the main text. Mean speeds in the behavior experiments ($v_0 = 22.61 \pm 0.07\ \mu\text{m/s}$) were comparable to those in the gradient experiments ($v_0 = 21.9 \pm 0.2\ \mu\text{m/s}$).

The persistence of tumbles α was computed from the average of the cosine of the angle between cells' projected headings before and after tumbles. Since we observe projections of the true headings, this slightly overestimates the true value of α. We consider a model to correct for this bias. In the absence of a gradient, cells don't have a preference for one direction over another, so the distribution of new headings after a tumble must be centered around the heading before the tumble. This is encoded in a model by making the distribution of post-tumble headings $\hat{u}$, given the pre-tumble heading u, depend only on $u \cdot \hat{u}$. Celani and Vergassola[10] used the tumble angle distribution $P(\hat{u}|u) = \frac{1}{4\pi}(1 + u \cdot \hat{u})$, which is the simplest approximation of a general tumble angle distribution that could have more complex dependence on $u \cdot \hat{u}$. In this model, the true directional persistence is $\alpha = \langle u \cdot \hat{u} \rangle = 1/3$, but if only projected trajectories are observed, the observed value is $\alpha = \pi^2/32$ (~0.31). Therefore, we multiplied our computed value of α by $1/3 * 32/\pi^2$ to correct for projection biases.

We estimated the mean tumble rate λ_{R0} in two ways. First, assuming exponentially-distributed runs, the probability that a cell tumbles in time step dt is $p_{tumble,dt} = \lambda_{R0}\,\Delta t$, where $\Delta t = 50$ ms is the frame interval of the movie. We calculated p_{tumble} as the total number of tumbles observed divided by the total number of cell-frames in which runs were observed, again only among cells with P_{run} near the population-median. Then we compute the mean tumble rate from $\lambda_{R0} = p_{tumble,dt}/\Delta t$, giving λ_{R0}~$0.893 \pm 0.003\ \mathrm{s}^{-1}$. This was the value used to compute $F(\boldsymbol{\theta})$ in the efficiency and β in the information rate $\dot{I}_{s\rightarrow a}$. Another method was to compute the mean tumble rate from the median P_{run} and average tumble time τ_T over cells with P_{run} near the median, using $P_{run} \sim 1/(1 + \lambda_{r0}\,\tau_T)$, which gave λ_{R0}~$0.912 \pm 0.003\ s^{-1}$.

Behavioral parameters also appear in the up-gradient velocity autocorrelation function $V(t)$, which is used to estimate the information rate $\dot{I}_{s\rightarrow a}$. Although the cells were not swimming in a gradient, their velocity statistics in the absence of a gradient are approximately the same as in shallow gradients. Therefore, we estimated $V(t)$ by computing the autocorrelation of velocity along one dimension, chosen arbitrarily to be the x-axis of our field of view, which is the gradient direction in the gradient experiments, and averaged over cells within bins of P_{run}. The resulting correlation functions decayed roughly exponentially (Fig. 2C; Supplementary Fig. S2A), and we fit the correlation function of the median bin of P_{run} with the functional form $V(t) = a_v\, e^{-\lambda_{tot}\,|t|}$ by nonlinear least squares using MATLAB's *fit* function. In the fit, each time delay t was given weight proportional to the number of samples observed at that time delay. The first few time points of $V(t)$ were excluded from the fit because they include sharp drops in velocity correlation that result from tumbles inevitably having finite velocity that rapidly decorrelates. In the median bin of P_{run}, the average trajectory duration was 7.0 seconds, and the total trajectory time was 1.3×10^4 seconds. Comparing to the model for $V(t)$ earlier, in theory, $\lambda_{tot} = (1-\alpha)\lambda_{R0} + 2\,D_r$ and $a_v = \frac{v_0^2}{3} P_{run}$. The fit values were comparable to these theoretical expressions using estimates of the individual parameters above and an estimate of D_r (below): $\lambda_{tot} = 0.862 \pm 0.005\ \mathrm{s}^{-1}$, compared to $(1-\alpha)\lambda_{R0} + 2\,D_r = 0.93 \pm 0.01\ \mathrm{s}^{-1}$; $a_v = 157.1 \pm 0.5\ \left(\frac{\mu\mathrm{m}}{\mathrm{s}}\right)^2$, compared to $\frac{v_0^2}{3} P_{run} = 151 \pm 2\ \left(\frac{\mu\mathrm{m}}{\mathrm{s}}\right)^2$. We directly used the fit values a_v and λ_{tot} in $V(\omega)$ when computing the information rate $\dot{I}_{s\rightarrow a}$.

Rotational diffusion was measured from trajectories of cells lacking the gene for *cheY* and therefore shouldn't be able to tumble. For these cells, $\lambda_{R0} = 0$, so we extracted the rotational diffusion coefficient again by fitting their average velocity autocorrelation function. We filtered out trajectories that were shorter than 5 seconds or that appeared to tumble. After filtering, the average trajectory duration was 9.5 seconds, and the total trajectory time was 9.7×10^4 seconds. The resulting velocity autocorrelation function was well-fit by a single decaying exponential (Supplementary Fig. S2B), whose decay rate $\lambda_{tot,\Delta cheY} = 2\,D_r$ was consistent with a previously reported[23] value for D_r (here, $D_r = 0.0441 \pm 0.0001\ \mathrm{rad}^2/\mathrm{s}$; previously $D_r = 0.062\ \mathrm{rad}^2/\mathrm{s}$).

Uncertainties in v_0, α, and λ_{R0} and their dependence on P_{run} were computed by bootstrapping tracks from within each bin of P_{run}, excluding cells with $P_{run} = 0$ or $P_{run} = 1$ (bin size 0.02). The bootstrapped sample for each bin was the same size as the number of tracks in the bin, and samples were drawn with replacement 100 times. From each bootstrapped sample, the average of each parameter was computed. The standard deviation of the average of each parameter among the bootstrapped samples was taken to be the uncertainty. Uncertainties of the median P_{run} were smaller than the bin size and therefore were taken to be half the size of a bin. The uncertainties of a_v, λ_{tot}, and D_r were determined from the uncertainty of the exponential fit: the 68% confidence interval of each parameter was determined from MATLAB's *fit* function; dividing by 2 gave uncertainties equivalent to one standard deviation, assuming normally-distributed parameter uncertainties.

Section 22: Estimating population-average drift speeds

To compute the drift speeds, we first computed the average x-velocity among cells in each frame of a given movie, $\langle v_x(t)\rangle$. This time course of "ensemble average" $\langle v_x(t)\rangle$ resembled an OU process with correlation time of about 1.5 s, similar to the cells' typical reorientation time $1/\lambda_{tot}$ (Fig S9). The drift speed v_d was computed as the average of the time course $\langle v_x(t)\rangle$. This is equivalent to computing the drift speed of each cell and taking a weighted average, with weights given by the duration of each cell's trajectory. Time-weighted averaging like this is necessary to get an unbiased estimate of the population average drift speed, even for a population without diversity. This is because, with a finite depth of field, there should be disproportionately more short trajectories swimming vertically, with small up-gradient displacement, than long up-gradient trajectories. However, the distribution of swimming directions in a given frame is unbiased, and time-averaging takes advantage of this.

Since the time course of $\langle v_x(t)\rangle$ has a finite correlation time, the values of $\langle v_x(t)\rangle$ at consecutive time points are not independent. To compute the uncertainty of v_d, we first computed the autocorrelation function of $\langle v_x(t)\rangle$ in each movie and fit the result with a decaying exponential $a + b\, e^{-t/\tau_c}$ to get the correlation time of fluctuations τ_c. The effective number of independent frames was then $n_I = n\frac{\Delta t}{\Delta t+2\,\tau_c}$ [3], where n is the total number of frames in the movie (minus outliers). The uncertainty (standard error) of v_d in a given experiment i was then computed as $\sigma_i = \sigma_{v_x}/\sqrt{n_I}$, where again σ_{v_x} was the variance of $\langle v_x(t)\rangle$ in that experiment.

Then, we computed the weighted average of drift speeds $v_{d,i}$ in the same gradient steepness to determine $v_d(g)$. Each $v_{d,i}$ was weighted inversely proportional to its squared standard error $1/\sigma_i^2$. To compute the uncertainty of this average, we first note that there were experiment-to-experiment variations in the mean drift speeds $v_{d,i}$ that exceeded the typical within-experiment uncertainty σ_i. This suggested that there was another random, experiment-to-experiment source of variation. This could result from observing a different sample of phenotypes from the population in each experiment. To account for these effects in the uncertainty of the average $v_d(g)$, we assume that a random effect is added to the drift speed in each experiment, with variance σ_r^2. This random effect is assumed to be uncorrelated with our measurement errors, and we estimated σ_r^2 using the variance of $v_{d,i}$ among experiments with the same gradient steepness. Together, the uncertainty of the average $v_d(g)$ is then given by $\sigma^2 = 1/\left(\sum_i \frac{1}{\sigma_i^2}\right) + \frac{1}{(N-1)}\sigma_r^2$, where N is the number of experiments performed in that gradient condition. The first term results from weighting each $v_{d,i}$ in the mean by the inverse of its measurement variance. The second term results from averaging over N realizations of the random effect, which all have the same statistics.

The average trajectory duration was 1.52 seconds, the average trajectory time per experiment was 2×10^4 seconds, and the total trajectory time was 5.5×10^5 seconds. Multiple experiments were performed for each gradient condition ($N \geq 5$ each).

The gradient steepness g in each experiment was estimated from fluorescein fluorescence images as follows. The background fluorescence intensity was measured by taking an image of a device filled with water but no fluorescein. The average intensity of this image was similar to that of an image with no sample mounted on the microscope. The average intensity of this background image was subtracted from all of the fluorescein gradient images, making zero-intensity regions correspond to regions where the MeAsp attractant concentration was 100 μM. To correct the influence of relative depth variations across the width of the gradient region and for spatial variations in illumination, an image of a device filled with fluorescein solution was taken. We will refer to this as the "blank" image. The average background intensity was subtracted from this image, as well. Then, in each gradient image and the blank image, the average intensity in a horizontal strip was computed ($\pm$300 pixels around row 1024; image size 2048 x 2048 pixels). This produced fluorescence intensity profiles across the gradient region, $I(x,t)$. The fluorescence profile from the blank image, $I_b(x,t)$, was normalized to a maximum value of 1. The fluorescence profiles at each time point of each experiment were aligned to the blank profile, and divided point by point by the normalized blank profile: $i(x,t) = I(x,t)/I_b(x,t)$. We separately quantified the effects of photobleaching by imaging a device full of fluorescein using the same protocol as a gradient experiment. Changes in intensity due to photobleaching were negligible (<0.2%).

At this point, we had fluorescence profiles that were corrected for background intensity, illumination variations, and relative variations of device depth. What remained was to determine the linear transformation from intensity, $i(x,t)$, to MeAsp concentration, $c(x,t)$, which could be experiment-dependent due to variations in absolute dimensions of different devices. In addition to knowing that $i(x,t) = 0$ corresponds to $c(x,t) = 100$ μM, we also know that the maximum value of $i(x,t)$, i_{max}, corresponds to $c(x,t) = c_1$, where c_1 is the concentration of MeAsp in the high-concentration reservoir. To estimate the maximum $i(x,t)$ in each experiment, we used the earliest fluorescein image from a given experiment, and used the value of $i(x,t)$ at the location in the high-concentration reservoir that was furthest from the gradient region. The reasoning for this was that diffusion or flow between the reservoirs could transport fluorescein (and MeAsp) from the high-concentration reservoir to the low one. The location furthest from the gradient region at the first time point is least affected by diffusion or flow, providing the best estimate of the maximum $i(x,t)$ before any mixing between reservoirs occur. With this, we could estimate concentrations from intensities using $c(x,t) = c_0 + \frac{(c_1 - c_0)}{i_{max}} i(x,t)$, where $c_0 = 100$ μM.

Finally, we used this transformation to estimate the concentration of MeAsp in each reservoir in the image taken just before the tracks movie and the time point just after it. We estimated g in each image using $g_{est} = \frac{\log(c_{1,est}) - \log(c_{0,est})}{\Delta x}$, where $c_{1,est}$ and $c_{0,est}$ are the estimated MeAsp concentrations in the two reservoirs, and $\Delta x = 1$ mm is the width of the gradient channel (see Methods). The final estimate of the gradient during the movie was the average of these two estimates. We took the uncertainty of the estimate to be the difference between the estimates at the two time points divided by two.

In all experiments, $g_{est} < g$; the estimated g was less than the intended one, consistent with mixing between the reservoirs making the gradient shallower. The amount by which g_{est} was smaller than g scaled with g. This is consistent with the fact that diffusive and convective fluxes scale linearly with concentration, so in the same amount of time (the typical time it took for the gradient to form), steeper gradients should have proportionally more molecules of MeAsp transported from the high reservoir to the low reservoir. The average percent reduction in gradient steepness relative to the intended steepness was 4.4%.

The chemotactic coefficient was estimated by linearly fitting $v_d = \chi\, g$ to the scatter plot of $v_{d,i}$ versus g_{est} from all gradient experiments, with each data point given weight inversely proportional to the squared standard error of $v_{d,i}$, $1/\sigma_i^2$, using MATLAB's *fit* function. Since the error bars on g_{est} were small, performing the fit in a way that accounted for errors in both $v_{d,i}$ and g_{est} had little effect on the result. Allowing the y-intercept of the fit to be non-zero, i.e. fitting $v_d = a + \chi\, g$, only increases χ by 3%, but increases its uncertainty by 70%.

Throughout, we were careful to estimate parameters of a median cell. But here, we are computing the population-average drift speed. But drift speed depends nonlinearly on behavioral parameters, and the average of a nonlinear parameter combination does not in general equal the nonlinear function evaluated with median parameters. Therefore, the properties of the phenotype that achieves the population-average drift speed could be different from the median phenotype. To address this, we note that from theory (i.e. Eqn. (17)), most of the dependence of the drift speed on behavioral parameters is captured by $v_0^2 \frac{(1-\alpha)\lambda_{R0}}{(1-\alpha)\lambda_{R0}+2\,D_r} P_{run}$. From swimming trajectories, we computed the average of each parameter in bins of P_{run}, i.e. $v_0(P_{run})$, $\lambda_{R0}(P_{run})$, etc (see Supplementary Fig. S1). Then, we computed this expression for each bin (Supplementary Fig. S9), as well as the distribution of P_{run}. Finally, we compared the average of this expression with respect to the distribution of P_{run} to the value one gets from plugging in the parameters corresponding to the median P_{run}. We find that these two are similar: the population average gives $\left\langle v_0^2(P_{run}) \frac{(1-\alpha)\lambda_{R0}}{(1-\alpha)\lambda_{R0}+2\,D_r} P_{run} \right\rangle \sim 375 \pm 1 \left(\frac{\mu\text{m}}{\text{s}}\right)^2$, whereas plugging parameters in the median bin of P_{run} gives $v_0^2 \frac{(1-\alpha)\lambda_{R0}}{(1-\alpha)\lambda_{R0}+2\,D_r} P_{run} \sim 410 \pm 3 \left(\frac{\mu\text{m}}{\text{s}}\right)^2$. This justifies our comparison of population-average drift speeds to bounds quantified using a median cell's parameters. Uncertainties in the distribution of P_{run} and in these quantities was determined by bootstrapping, as described in the **Section 21: Estimating behavioral parameters, run speed, and rotational diffusion** section.

Supplementary Information References

1. Taute, K. M., Gude, S., Tans, S. J. & Shimizu, T. S. High-throughput 3D tracking of bacteria on a standard phase contrast microscope. *Nat Commun* **6**, 8776 (2015).
2. Waite, A. J. *et al.* Non-genetic diversity modulates population performance. *Molecular Systems Biology* **12**, 895 (2016).
3. Govern, C. C. & Wolde, P. R. ten. Optimal resource allocation in cellular sensing systems. *PNAS* **111**, 17486–17491 (2014).
4. Sourjik, V. & Berg, H. C. Binding of the Escherichia coli response regulator CheY to its target measured in vivo by fluorescence resonance energy transfer. *PNAS* **99**, 12669–12674 (2002).
5. Francis, N. R. *et al.* Subunit Organization in a Soluble Complex of Tar, CheW, and CheA by Electron Microscopy. *J. Biol. Chem.* **277**, 36755–36759 (2002).
6. Levit, M. N., Grebe, T. W. & Stock, J. B. Organization of the Receptor-Kinase Signaling Array That Regulates Escherichia coli Chemotaxis. *J. Biol. Chem.* **277**, 36748–36754 (2002).
7. Lansey, J. C. Beautiful and distinguishable line colors + colormap. *MATLAB Central File Exchange* https://www.mathworks.com/matlabcentral/fileexchange/42673-beautiful-and-distinguishable-line-colors-colormap (2015).
8. Schreiber, T. Measuring Information Transfer. *Phys. Rev. Lett.* **85**, 461–464 (2000).
9. Locsei, J. T. Persistence of direction increases the drift velocity of run and tumble chemotaxis. *J. Math. Biol.* **55**, 41–60 (2007).
10. Celani, A. & Vergassola, M. Bacterial strategies for chemotaxis response. *PNAS* **107**, 1391–1396 (2010).

11. Masson, J.-B., Voisinne, G., Wong-Ng, J., Celani, A. & Vergassola, M. Noninvasive inference of the molecular chemotactic response using bacterial trajectories. *PNAS* **109**, 1802–1807 (2012).

12. Wong-Ng, J., Melbinger, A., Celani, A. & Vergassola, M. The Role of Adaptation in Bacterial Speed Races. *PLOS Computational Biology* **12**, e1004974 (2016).

13. Long, J., Zucker, S. W. & Emonet, T. Feedback between motion and sensation provides nonlinear boost in run-and-tumble navigation. *PLOS Computational Biology* **13**, e1005429 (2017).

14. Sourjik, V. & Berg, H. C. Functional interactions between receptors in bacterial chemotaxis. *Nature* **428**, 437–441 (2004).

15. Shimizu, T. S., Tu, Y. & Berg, H. C. A modular gradient-sensing network for chemotaxis in Escherichia coli revealed by responses to time-varying stimuli. *Molecular Systems Biology* **6,** 382 (2010).

16. Lazova, M. D., Ahmed, T., Bellomo, D., Stocker, R. & Shimizu, T. S. Response rescaling in bacterial chemotaxis. *PNAS* **108**, 13870–13875 (2011).

17. Tu, Y. Quantitative Modeling of Bacterial Chemotaxis: Signal Amplification and Accurate Adaptation. *Annual Review of Biophysics* **42**, 337–359 (2013).

18. Kamino, K., Keegstra, J. M., Long, J., Emonet, T. & Shimizu, T. S. Adaptive tuning of cell sensory diversity without changes in gene expression. *Science Advances* (2020).

19. Monod, J., Wyman, J. & Changeux, J.-P. On the nature of allosteric transitions: A plausible model. *Journal of Molecular Biology* **12**, 88–118 (1965).

20. Mello, B. A. & Tu, Y. An allosteric model for heterogeneous receptor complexes: Understanding bacterial chemotaxis responses to multiple stimuli. *PNAS* **102**, 17354–17359 (2005).

21. Keymer, J. E., Endres, R. G., Skoge, M., Meir, Y. & Wingreen, N. S. Chemosensing in Escherichia coli: Two regimes of two-state receptors. *PNAS* **103**, 1786–1791 (2006).

22. Kalinin, Y. V., Jiang, L., Tu, Y. & Wu, M. Logarithmic Sensing in Escherichia coli Bacterial Chemotaxis. *Biophysical Journal* **96**, 2439–2448 (2009).

23. Berg, H. C. & Brown, D. A. Chemotaxis in Escherichia coli analysed by Three-dimensional Tracking. *Nature* **239**, 500–504 (1972).

24. Korobkova, E., Emonet, T., Vilar, J. M. G., Shimizu, T. S. & Cluzel, P. From molecular noise to behavioural variability in a single bacterium. *Nature* **428**, 574–578 (2004).

25. Park, H., Oikonomou, P., Guet, C. C. & Cluzel, P. Noise Underlies Switching Behavior of the Bacterial Flagellum. *Biophysical Journal* **101**, 2336–2340 (2011).

26. Keegstra, J. M. *et al.* Phenotypic diversity and temporal variability in a bacterial signaling network revealed by single-cell FRET. *eLife* **6**, e27455 (2017).

27. Colin, R., Rosazza, C., Vaknin, A. & Sourjik, V. Multiple sources of slow activity fluctuations in a bacterial chemosensory network. *eLife* **6**, e26796 (2017).

28. Sneddon, M. W., Pontius, W. & Emonet, T. Stochastic coordination of multiple actuators reduces latency and improves chemotactic response in bacteria. *PNAS* **109**, 805–810 (2012).

29. Flores, M., Shimizu, T. S., ten Wolde, P. R. & Tostevin, F. Signaling Noise Enhances Chemotactic Drift of E. coli. *Phys. Rev. Lett.* **109**, 148101 (2012).

30. Lovely, P. S. & Dahlquist, F. W. Statistical measures of bacterial motility and chemotaxis. *Journal of Theoretical Biology* **50**, 477–496 (1975).

31. Barkai, N. & Leibler, S. Robustness in simple biochemical networks. *Nature* **387**, 913–917 (1997).

32. Alon, U., Surette, M. G., Barkai, N. & Leibler, S. Robustness in bacterial chemotaxis. **397**, 4 (1999).

33. Berg, H. C. & Purcell, E. M. Physics of chemoreception. *Biophysical Journal* **20**, 193–219 (1977).

34. Segall, J. E., Block, S. M. & Berg, H. C. Temporal comparisons in bacterial chemotaxis. *PNAS* **83**, 8987–8991 (1986).

35. Bialek, W., Nemenman, I. & Tishby, N. Predictability, complexity, and learning. *Neural Comput* **13**, 2409–2463 (2001).

36. Bialek, W., De Ruyter Van Steveninck, R. R. & Tishby, N. Efficient representation as a design principle for neural coding and computation. in *2006 IEEE International Symposium on Information Theory* 659–663 (2006). doi:10.1109/ISIT.2006.261867.

37. Palmer, S. E., Marre, O., Berry, M. J. & Bialek, W. Predictive information in a sensory population. *PNAS* **112**, 6908–6913 (2015).

38. Becker, N. B., Mugler, A. & ten Wolde, P. R. Optimal Prediction by Cellular Signaling Networks. *Phys. Rev. Lett.* **115**, 258103 (2015).

39. Shannon, C. E. Communication in the Presence of Noise. *Proceedings of the IRE* **37**, 10–21 (1949).

40. Tostevin, F. & ten Wolde, P. R. Mutual Information between Input and Output Trajectories of Biochemical Networks. *Phys. Rev. Lett.* **102**, 218101 (2009).

41. Sourjik, V. & Berg, H. C. Receptor sensitivity in bacterial chemotaxis. *Proceedings of the National Academy of Sciences* **99**, 123–127 (2002).

42. Sourjik, V., Vaknin, A., Shimizu, T. S. & Berg, H. C. [17] - In Vivo Measurement by FRET of Pathway Activity in Bacterial Chemotaxis. in *Methods in Enzymology* (eds. Simon, M. I., Crane, B. R. & Crane, A.) vol. 423 365–391 (Academic Press, 2007).

43. Zal, T. & Gascoigne, N. R. J. Photobleaching-Corrected FRET Efficiency Imaging of Live Cells. *Biophysical Journal* **86**, 3923–3939 (2004).

44. Miyawaki, A. Development of Probes for Cellular Functions Using Fluorescent Proteins and Fluorescence Resonance Energy Transfer. *Annual Review of Biochemistry* **80**, 357–373 (2011).

45. Balleza, E., Kim, J. M. & Cluzel, P. Systematic characterization of maturation time of fluorescent proteins in living cells. *Nature Methods* **15**, 47–51 (2018).

46. Elowitz, M. B., Surette, M. G., Wolf, P.-E., Stock, J. B. & Leibler, S. Protein Mobility in the Cytoplasm ofEscherichia coli. *Journal of Bacteriology* **181**, 197–203 (1999).

47. MacKay, D. J. C. Information Theory, Inference, and Learning Algorithms. 640.

48. Sarkka, S. *Bayesian Filtering and Smoothing*. (Cambridge University Press, 2013). doi:10.1017/CBO9781139344203.

49. Parthasarathy, R. Rapid, accurate particle tracking by calculation of radial symmetry centers. *Nature Methods* **9**, 724–726 (2012).

50. Park, H. *et al.* Interdependence of behavioural variability and response to small stimuli in bacteria. *Nature* **468**, 819–823 (2010).

51. Cluzel, P., Surette, M. & Leibler, S. An Ultrasensitive Bacterial Motor Revealed by Monitoring Signaling Proteins in Single Cells. *Science* **287**, 1652–1655 (2000).