# SI Appendix

## S1 Experimental details

### S1.1 Yeast strains and plasmids

For live transcript analysis, we engineered the haploid strains of *Saccharomyces cerevisiae* (BY4742 and BY4741), which are isogenic to S288C (Research Genetics/Invitrogen, Huntsville, AL). 14X PP7 binding sites (hairpins) were amplified from pTL031[36] using primers T1053, T1054 and integrated in one of the haploids by replacing one of the *CUP1* ORFs in the yeast genome by homologous recombination[37]. For expressing PP7-GFP coat protein, we constructed a MET3 integrative vector (pTSK630) to express PP7-NLS-GFP from *SEC61* promoter and integrated this vector by SacI XhoI digestion in both the haploids. This vector can be available upon request. For the photobleaching correction, we used a diploid strain YTK1231 in which both the *CUP1* arrays are replaced by 256 copies of *LacO* and lacI-GFP-NLS is expressed from *pHIS3*. For preparing the calibration curve for the number of GFP molecules verses brightness/intensity, we used three yeast strains (YTK541, YTK1231 and YTK1268) with known numbers of GFP molecules per locus. YTK541, contains a tandem array of 10 copies of *CUP1* locus with 40 binding sites for the transcription activator Ace1p-GFP. *CUP1* is activated by Cu, and at the peak of activity *CUP1* array binds 120 molecules of GFP. In YTK1231, each lacO binding site may be associated with a dimer of the lac Repressor (LacI-GFP)[38]. Therefore, the array of 256 tandem lacO binding sites is associated with 512 LacI-GFP molecules. In YTK1268, the spindle pole body of the diploid yeast strain contains approximately 1000 molecules of Spc42-GFP. Strain genotypes are provided in Table S8, plasmids in Table S9 and primer sequences in Table S10.

### S1.2 Media and growth conditions

For live transcript analysis, cells of YTK1799 were plated on CSM-URA plate (from −80 °C frozen glycerol stock) and grown for 48 h at 28 °C. 3 to 5 colonies were inoculated in 3 mL CSM-URA media (in 14 mL polypropylene tubes, Cat no. 352059, Falcon, Maxico) and grown for overnight at 28 °C, 230 RPM. 250 µL of this overnight grown culture was inoculated in 25 mL CSM-URA (in 250 mL flask) and grown at 28 °C, 230 RPM for 24 h. This flask was removed from the shaker and kept in refrigerator at 4 °C. We used this refrigerated culture for daily inoculations for a month to get consistent results (to avoid day to day variations in transcription induction kinetics due to difference in the age of the culture). From this refrigerated culture, we inoculated 60 µL in 3 mL of fresh CSM-URA media (in 14 mL polypropylene tubes, Cat no. 352059, Falcon, Maxico) in the morning and grew the cultures for 5 h at 28 °C, 230 RPM. Cells were harvested by centrifugation (2200 RPM for 1 min) and cells were placed under the CSM-URA agarose pad (100 µM $CuSO_4$) for imaging. For the photobleaching correction and GFP calibration curve, strains YTK541, YTK1231 and YTK1268 were grown under the same conditions, except YTK541 and YTK1268 were grown in CSM-HIS media.

### S1.3 Quantitative RT-PCR (RT-qPCR)

Samples were harvested at indicated time points after Cu induction. RNA was extracted (from yeast cells) using the ISOLATE II RNA Mini kit (Bioline, UK, Cat no. BIO-52072). cDNA was prepared using the iScript cDNA synthesis kit (BioRad, Cat no.: 1708891) starting with 1 mg of total RNA. Quantitative real-time PCR (qPCR) was performed as described[27]. For normalization, the expression of the housekeeping gene ACT1 was quantified. Primers used for this quantification are listed in Table S10 (T531, T532 for CUP1 and T1055, T1056 for ACT1). To confirm the absence of contaminating genomic DNA in cDNA preparations, reverse transcriptase negative (-RT) samples were used as a control, which produced the Ct value difference of >10 cycles between "-RT" and "+RT"samples, indicating a negligible amount of genomic DNA contamination in cDNA samples. mRNA extraction, cDNA synthesis, and qPCR were repeated at least twice, and qPCR was performed in duplicates for each experiment. Error bars indicate SEM.

### S1.4 Microscope settings and imaging conditions

For imaging live cells, 5 h grown cultures were harvested by centrifugation (2200 RPM for 1 min) and 3 µL of cell pellet were placed in Lab-Tek II chambered coverglass (1.5 Borosilicate Gass, Nunc, ThermoFisher Scientific, MA, US), mixed with equal volume of 200 µM $CuSO_4$ containing CSM-URA and covered by 1cm x 1cm CSM-URA agarose pad (100 µM $CuSO_4$). A timer was started immediately upon mixing the cells with 200 µM $CuSO_4$ containing CSM-URA. 3D time-lapse movies were acquired at the room temperature on the DeltaVIsion Elite Microscope, using 100x 1.4 NA oil immersion objective lens, sCMOS camera, FITC filter set (15 ms exposure, Ex 488/27; Em 505/45, Chroma Technology Corp, Bellows Falls, VT), 15 z-steps at every 400 nm, 1x1 binning and 1024x1024 pixels. Time-lapse movies with 3 s time interval were acquired for 90 s, followed by changing the field within next 90 s, imaging new field for another 90 s and so on. This imaging regime was repeated 9 times within 30 min to cover the entire slow cycle. E.g. to cover the entire slow cycle of 30 min with 3 s time interval, first set of movies were started for new field of cells after 3, 6, 9, 12, 15, 18, 21, 24, 27 min after copper induction and acquired for 90 s. Remaining 90 s (between 3 min and 6 min time points, between 6 min and 9 min time points, and so on) were used for moving to

the next field of cells. Second set of movies were acquired for new field of cells after 4.5, 7.5, 10.5, 13.5, 16.5, 19.5, 22.5, 25.5 and 28.5 min after copper induction and acquired for 90 s to compensate the missing time points from the first set. Similarly, time-lapse movies for 12 s time interval were recorded for 5 min, followed by changing the field of view within the next 1 min. A first set of movies was started after 3, 9, 15, 21, 27 min after induction. A second set of movies was stared at 6, 12, 18, 24 min after induction. For the photobleaching correction, strain YTK1231 was imaged under the same condition with 3 s time interval.

**Table S1.** Yeast strains used in this study

| Strain ID | Application | Source | Genotype |
|---|---|---|---|
| YTK541 | GFP calibration | This study | MATα, his3-D1 leu2-D0 ura3-D0 lys2-D0 ace1-D1::KAN TRP1::TRP1ORF-pCap2-ACE1-tripleGFP-HIS3 |
| YTK1231 | GFP and photobleaching calibration | This study | MATa/MATα, his3-D1/his3-D1 leu2-D0/leu2-D0 trp1/TRP1 ura3-D0/ura3-D0 lys2-d::pHIS3-lacI-GFP-NLS-Nat1/ lys2-d::pHIS3-lacI-GFP-NLS-Nat1 MET15/met15-D0 Cu1::KAN-(LacO)256/Cu3::(LacO)256 [pRS426 pHIS3-LacI-GFP-URA3] |
| YTK1268 | GFP calibration | This study | MATa/MATα, his3-d1/his3-d1 leu2-d0/leu2-d0 met15-d0/MET15 LYS2/lys2-d0 ura3-d0/ura3-d0 SPC42-GFP-HIS3/SPC42-GFP-HIS3 |
| YTK1799 | 3s movies, 12s movies | This study | MATa/MATα, his3-d1/his3-d1 leu2-d0/leu2-d0 lys2-d0/LYS2 MET15/met15-d0 ura3-d0/ura3-d0 pdr5-d::LoxP/pdr5-d::LoxP trp1d::pADH-AFB2::LEU2/trp1d::pADH-AFB2::LEU2 ace1-d::pCAP2-ACE1-mCherry-TRP1/ ace1-d::pCAP2-ACE1-mCherry-TRP1 RSC2-AID-9Myc::HIS5/RSC2-AID-9Myc::HIS5 CUP1-PP7hairpins(14)-KANMX/CUP1 MET3::pSEC61-PP7-GFP-CTCT-URA3/MET3::pSEC61-PP7-GFP-CTCT-URA3 |

**Table S2.** Plasmids used in this study

| Plasmid Name | Application | Primers | Source |
|---|---|---|---|
| pTL031 | For amplifying 14x PP7 binding sites (hairpins) | T1053-T1054 | Lenstra and Larson, 2016[36] |
| pTSK630 | For integrating pSEC61-PP7-GFP-CYCT at MET3 locus using URA3 (K. lactis) as a selection marker. Cut with SacI and XhoI to integrate into MET3 | | This study |

**Table S3.** Sequences of the primers used in this study

| Primer ID | Application | Sequence |
|---|---|---|
| T1053 | For replacing one copy of *CUP1* ORF with *14x PP7* binding sites (Forward) | gatattaagaaaaacaaactgtacaatcaatcaatcaat-catcacataaa gtaaaacgacggccagtgagcg* |
| T1054 | For replacing one copy of *CUP1* ORF with *14x PP7* binding sites (Reverse) | aaaattaaaaacagcaaatagttagatgaatatattaaagactattcgtt gttttcgacactggatggcggcg* |
| T531 | RT-qPCR | CATTTCCCAGAGCAGCATGA |
| T532 | RT-qPCR | GTCATGAGTGCCAATGCCAA |
| T1055 | RT-qPCR | ggttgctgctttggttattgataacgg |
| T1056 | RT-qPCR | gttcttctggggcaactctc |

*Homologous sequences for site-specific integration are shown in blue. Sequences homologous to plasmid DNA are shown in red.

**Table S4.** Structures used for GFP calibration

| Strain | Sub-cellular structure | Expected GFP |
|---|---|---|
| YTK541 | CUP1 array / Ace1p-3xGFP | 120 |
| YTK1231 | lacO (256) / LacI-GFP | 512 |
| YTK1231Enh | lacO (256) / LacI-GFP | 512 |
| YTK1268 | SPB/ Spc42-GFP | 1000 |

**Table S5.** Datasets collected in this study

| Dataset | Strain | No.Movies |
|---|---|---|
| 3s | YTK1799 | 82 |
| 12s | YTK1799 | 71 |
| Photobleaching 3s | YTK1231 | 25 |
| Photobleaching 12s | YTK1231 | 20 |
| GFP calibration | YTK541, YTK1231, YTK1799, YTK1268 | 20 |

## S2 Data acquisition and pre-processing

### S2.1 Cell selection

As a first step, we extracted movies of individual cells from the full movies. In order to extract cells with spots, we computed two projections: a maximum projection over all frames leading to a 2D movie and a maximum projection over all time frames and slices producing a single image. The latter single image allowed to identify all cells that develop an active transcription site during imaging. For the identified cells we drew ROIs in the initial time frame of the 2D video. These initial ROIs were tracked over time and used to extract 3D movies of individual cells. For each cell, the full image stack over time was stored for trace extraction. Fluorescence intensity traces were obtained as described in Sec. S6.2. A more detailed description of the custom tracker is given in Sec. S12.

### S2.2 Trace extraction

To track fluorescence spots and quantify fluorescence levels, we developed a custom 3D method based on sequential filtering. The location of the spot within an image stack at time $t_k$ is given by $r_k = (x_k, y_k, z_k)$. The spot is modeled as a diffraction limited point source such that the intensity $\hat{I}$ of the pixel at position $r$ can be described as

$$\hat{I}_k(r) = b_k + I_k \exp\left[-\frac{1}{2}(r - r_k)^T \Sigma_{\mathrm{PSF}}^{-1}(r - r_k)\right]$$

where the diagonal matrix $\Sigma_{\mathrm{PSF}}$ describes the shape of the point spread function of the optical system, $b_k$ is a local background and $I_k$ is the peak intensity of the spot that is related to the underlying intensity of the point source by a constant factor. In addition, we introduce a binary variable $s_k$ that describes whether the spot is visible in frame $k$. This leads to a total state

$x_k = (r_k, b_k, y_k, s_k)$. The observation model is given by the predicted intensity per pixel with additive Gaussian noise. State estimation is carried out using the standard recursive filtering approach[43]. Due to the non-linear observation model, a Laplace approximation is used to evaluate filter updates. The details are given in S12. The filter provides estimates of $I_k$ for $k = 1, \ldots, n$ which are treated as the noisy measurements $y_1, \ldots, y_n$ for the main analysis. Traces showing dividing spots due to DNA replication were excluded from further analysis. In this study we do not take into account the noise from the fraction of the cells that completed the replication of the reporter gene, but did not separate sister chromatids, which may lead to transcription of two alleles in the same spot.

### S2.3 Quality Control

In the images, some of the cells showed too high or too low GFP expression or were otherwise malfunctioning. In the 12s data set, we identified 3260 cells where an active TS is visible in at least one time frame during imaging. After tracking, all traces were compared manually to the corresponding cell video for quality checking. During this screening, we removed all potentially problematic traces, e.g. floating or dividing cells. The remaining 3036 high quality traces were used for further analysis. The same procedure was applied to the 3s data set leading to a selection of 3685 of an initial 4053 cells. A detailed summary of extracted cells per time interval is given in Table S13.

**Table S6.** Cells extracted from the 3 *s* and 12 *s* data sets

| Time since induction | 3*s* data set | | 12*s* data set | |
|---|---|---|---|---|
| | Extracted | Selected | Extracted | Selected |
| 3 min | 20 | 17 | 223 | 208 |
| 4.5 min | 78 | 68 | | |
| 6 min | 170 | 157 | 361 | 339 |
| 7.5 min | 227 | 202 | | |
| 9 min | 308 | 282 | 627 | 589 |
| 10.5 min | 322 | 285 | | |
| 12 min | 391 | 365 | 498 | 448 |
| 13.5 min | 315 | 300 | | |
| 15 min | 427 | 399 | 533 | 510 |
| 16.5 min | 284 | 270 | | |
| 18 min | 351 | 313 | 477 | 440 |
| 19.5 min | 248 | 228 | | |
| 21 min | 255 | 231 | 251 | 232 |
| 22.5 min | 143 | 120 | | |
| 24 min | 214 | 195 | 172 | 162 |
| 25.5 min | 108 | 93 | | |
| 27 min | 125 | 105 | 118 | 108 |
| 28.5 min | 50 | 40 | | |
| 30 min | 17 | 15 | | |

## S3 Exploratory analysis

Before turning to the model-based Bayesian inference approach, we performed descriptive statistical analysis to confirm the cyclic activity of the *CUP1* promoter known from earlier studies.

### S3.1 Total image brightness

The movies within different dataset were collected on different days. To ensure that these movies are comparable, we check the total intensity distributions of the first time frame of every movie. By focusing on the first time step, we can avoid confounding effects of bleaching. Boxplot representations of the intensity distributions (Fig. S11) reveal some variations in the intensity distribution. To understand these variations, it is important to note that the images consist mostly of extra-cellular background. Biological differences mainly affect the nuclear brightness and can thus not be responsible for large distribution shifts. Possible factors that could affect total brightness are the agarose medium in the background or variations in illumination or other factors of the optical system.

The most striking difference is that the movies collected on day 4 in the 3s dataset are significantly brighter than all other movies in this dataset. To make the dataset more homogenous, we exclude the day 4 movies from all further analysis. We also

observe that the 12s dataset is significantly brighter than the 3s dataset. The reason for this difference is unclear as several months are between the recording dates. However, as we analyze the dataset separately and optical factors such as the gfp scaling are inferred from the dataset, we do not view this as a problem. The third observation we discuss here is that the brightness tends to decrease for movies take from the same coverslip. This is most likely due to some cross-bleaching from imaging neighboring positions. Since the effect is rather small, we will ignore it in the following analysis.

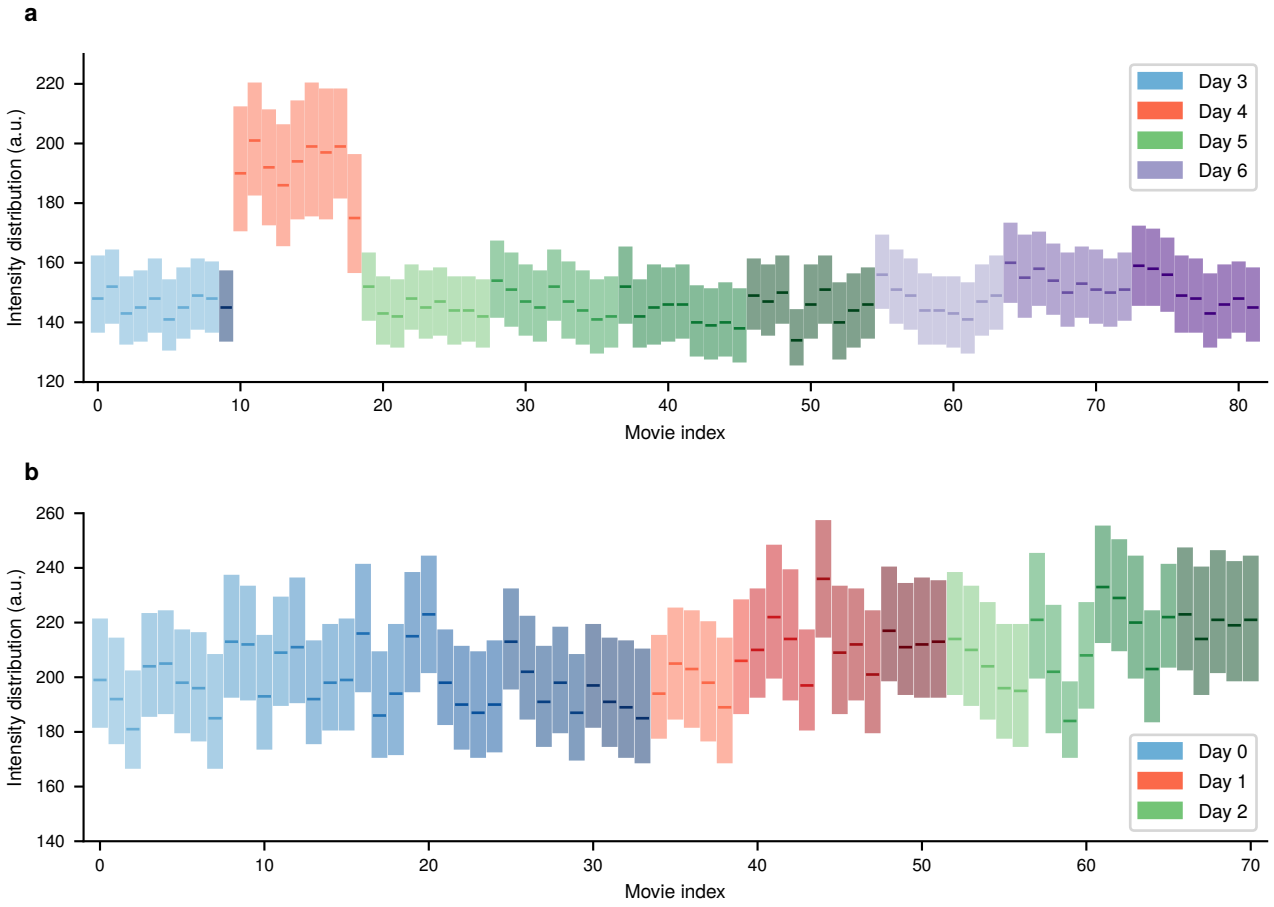


**Figure S6.** Pixel intensity distributions for the first time frame of every movie in form of a box plot shown for the 3 s dataset (**a**) and the 12 s dataset (**b**). Lower and upper border of the rectangles represent 25th and 75th percentile, respectively. The bold line within the box indicates the median. Every color family represents one experiment day, while the shading withing a color family indicates movies taken from the same coverslip

### S3.2 Spot Brightness

We now take a first look at the spot brightness. Similarly as an the previous section, we consider only the first frame of every movie to avoid effects of bleaching. For every movie, the spot intensities of all active cells in the first frame were pooled. The corresponding distributions are shown in the form of boxplots for both 3s and 12s dataset in Fig. S12.

Especially in the 3s dataset one can observe a clear dependence of spot brightness on the time since induction. Since the movies taken from the same coverslip are ordered sequentially, they correspond to spot measurements every 3 minutes or every 6 minutes for 3 s dataset and 12 s dataset, respectively. The observed sharp increase followed by a longer decrease is well in accordance with measurements from RT-qPCR. In addition, the results from different coverslips are quite similar and show less variation than the total intensity distributions in Fig. S11. The reason for this is that spot intensities are measured relative to local nuclear background such that additive noise affecting the total brightness of images is filtered out by the spot tracking. The one exception from this are the movies from day 4, which we have already decided to treat as outliers due to their much higher overall brightness.

The dependence on time since induction is also visible in the 12 s dataset albeit less clear than for the 12 s dataset. One reason

for this is that there are fewer measurements per coverslip due to the longer duration of the individual movies. In addition, some series of the movies are shifted by 3 minutes leading to a less regular appearance.

Finally, it seems that the spots in the 12 s dataset are overall brighter than in the 3 s dataset. Two possible explanations come to mind. The first possibility is that the cell cultures differ in activity which is possible due the time between the collection of the datasets. Alternatively, multiplicative noise affecting the total intensity distribution could explain the differences as it would not be filtered out by the spot tracking. In practice, it could also be a combination of both factors. We deal with this by estimating the observation model (cf. Sec. S9) parameters individually for every dataset. Any multiplicative noise will be captured by the estimated GFP scaling factor such that remaining differences can be attributed to Biological causes.

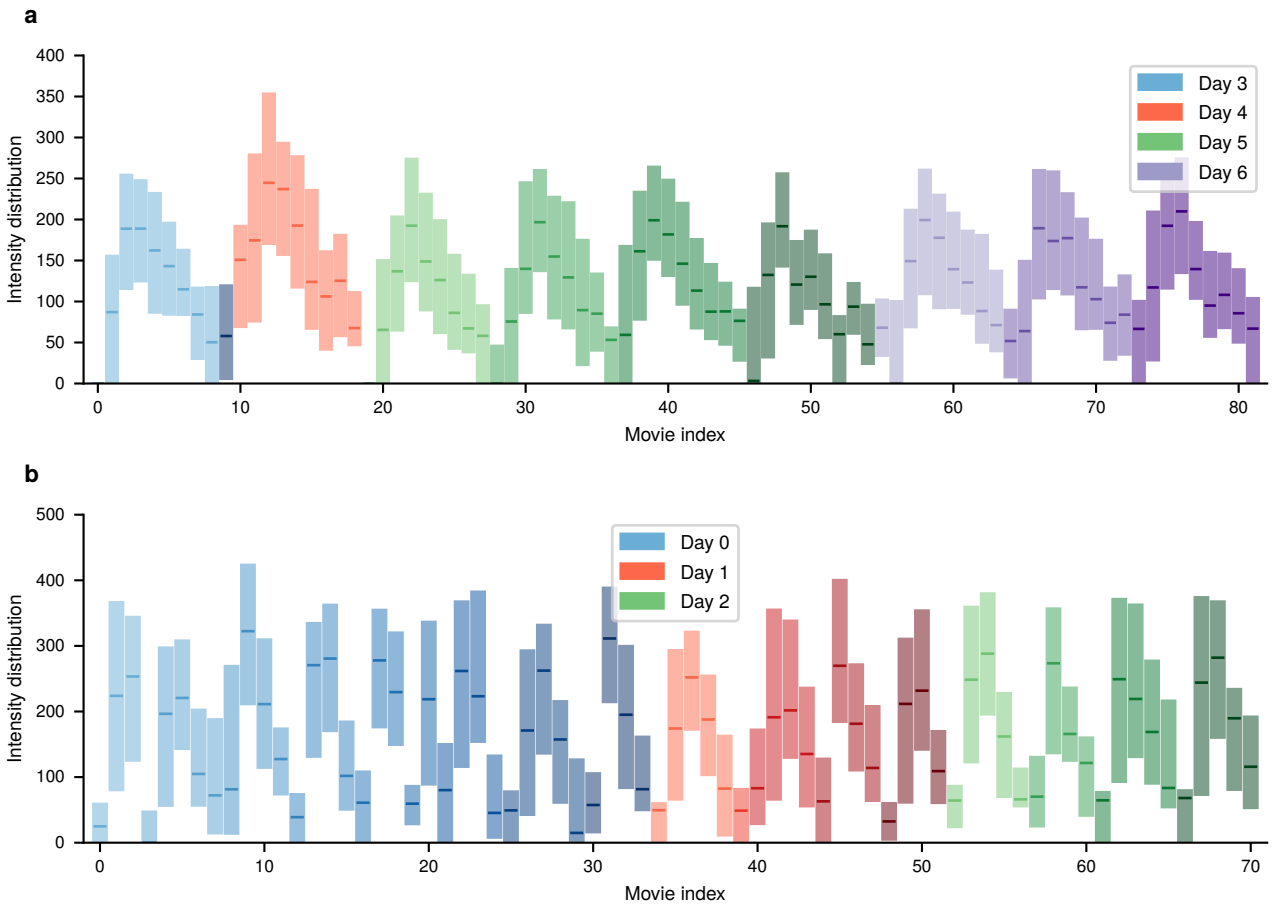


**Figure S7.** Pixel intensity distributions for the first time frame of every movie in form of a box plot shown for the 3 s dataset (**a**) and the 12 s dataset (**b**). Lower and upper border of the rectangles represent 25th and 75th percentile, respectively. The bold line within the box ndicates the median. Every color family represents one experiment day, while the shading withing a color family indicates movies taken from the same coverslip.

### S3.3 Investigating time-dependant activity

The presence of non-stationary transcription dynamics can be already seen from the movie histograms in Fig. S12. We will now take a closer look by inspecting the average spot intensities pooled over all movies that started with the same time delay with respect to induction. As before, we will only consider the first time frame of each movie to avoid the effects of bleaching. The first row in Fig. S13 shows the average spot intensity over the cycle. For both 3 s dataset (left) and 12 s dataset (right) wie see a sharp rise of the intensity in the early part of the observed interval followed by a slower decay. While the general shape of both curves is similar, the spots in the 12s dataset are generally brighter as we have seen from the per-movie distributions (cf. Fig. S12). The second row of Fig. S13 shows the number of responding cells divided by the total number of cells per time window. To assign cells to the class of responders or non-responders, we used a standard Gaussian mixture classifier on the spot intensity distributions of the datasets. In order to avoid bias by the overall different brightness, we applied this classifier individually on

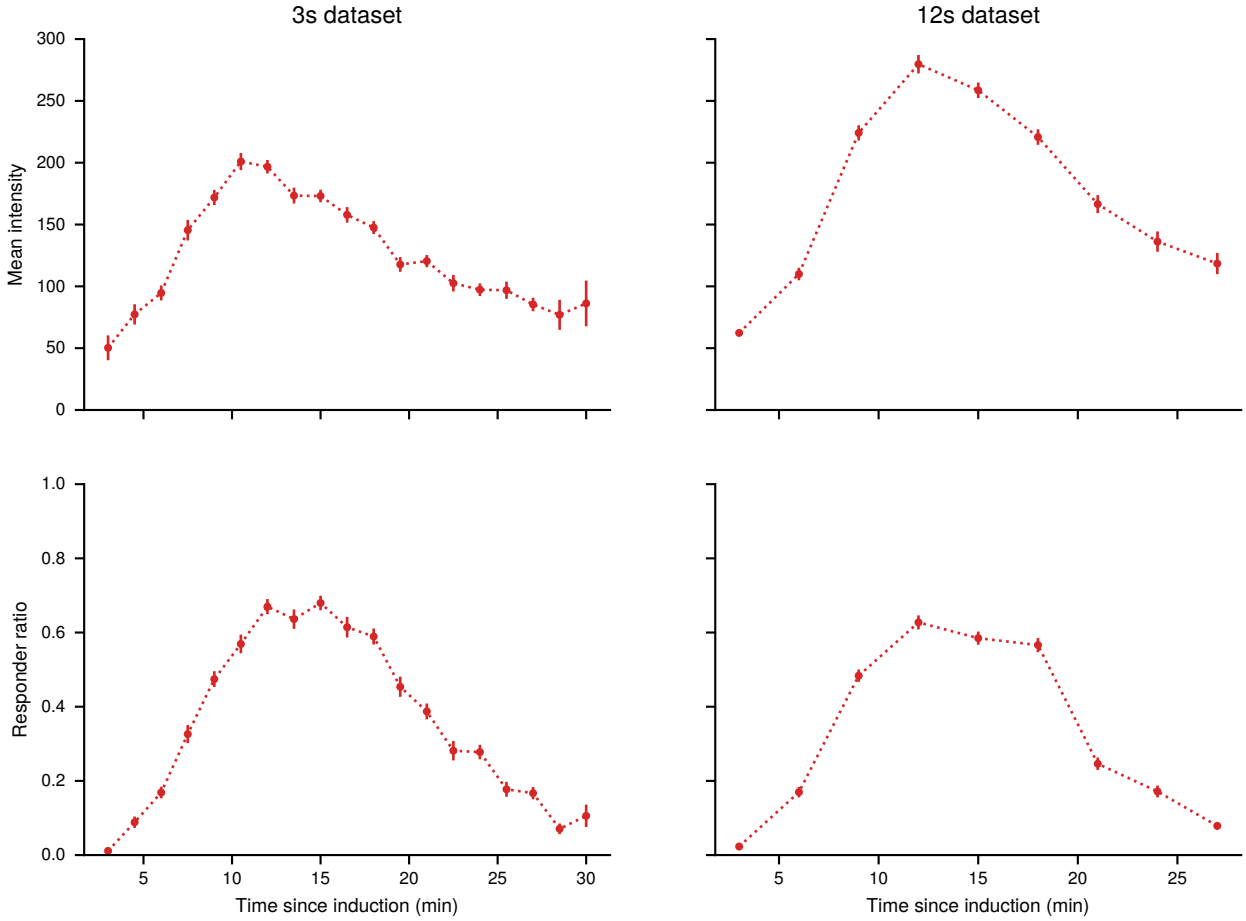


**Figure S8.** Dependence of summary statistics on the time since induction. The first row shows the mean of the spot intensities pooled over all videos with the same time delay since induction. The second row shows the number of responders divided by the total number of cells per time window. Error bars indicate standard error. Results for the 3 s dataset are also shown in the main text (Fig. 1f, g)

each dataset. The resulting responder ratios of both datasets agree well.

## S4 Model

### S4.1 Kinetic Transcription Model

As described in the main text, the model is an example of a Markov jump process that satisfies the master equation (1). The transition function $Q$ defines the probability of an event to happen in an infinitesimal interval $h$

$$P(X(t+h) = x' \mid X(t) = x) = Q(x, x' \mid \theta)h + o(h)\,.$$

The transition function is fully specified by the vector of parameters $\theta = (k_{\text{on}}, k_{\text{off}}, k_{\text{i}}, k_{\text{e}}, k_{\text{t}})^\top$ and the conditions under which transitions can occur, leading to

$$Q(x, x' \mid \theta) = \begin{cases} k_{\text{on}} & x_0 = 0, x_0' = 1\,, \\ k_{\text{off}} & x_0 = 1, x_0' = 1\,, \\ k_{\text{i}} & x_0 = 1, x_1 = 0, x_1' = 1\,, \\ k_{\text{e}} & x_i = 1, x_{i+1} = 0, x_i' = 0, x_{i+1}' = 1\,, \quad i = 1, \ldots, L-1\,, \\ k_{\text{t}} & x_l' = x_l - 1 \end{cases}.$$

For parameter inference, we need to evaluate the system for many different parameter configurations. It is therefore convenient to represent the transition functions as

$$Q(x, x' \mid \theta) = \sum_{i=1}^{5} \theta_i R_i(x, x') \,. \tag{S2}$$

The operators $R_i$ are independent of the parameters. By enumerating the states of the system, the probability $p(x, t)$ can be represented by a vector $\mathbf{p}(t)$ and the transition function $Q$ becomes a sparse matrix $\mathbf{Q}$. With this, the master equation becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{p}(t) = \mathbf{Q}\mathbf{p}(t) \,.$$

A formal solution of this system is given by the matrix exponential

$$\mathbf{p}(t) = \exp\left(\mathbf{Q}t\right)\mathbf{p}_0$$

with initial distribution $\mathbf{p}_0$. For sparse $\mathbf{Q}$, this can be efficiently solved for fairly large state spaces by the Krylov subspace approximation for matrix exponentials[41,42].

### S4.2 Observation Model

The kinetic model discussed above is a continuous time model. In practice, one cannot observe such a systems in continuous time but rather at discrete sample times $t_1, \ldots, t_n$. In addition, we do not observe the process $X(t_k)$ directly. Here, our measurement $Y(t_k)$ is provided by the total intensity of the fluorescence spots as measured by the tracking algorithm. In the following, we construct a model that relates $X$ and $Y$. First, note that as the polymerase traverses the gene, an additional stem loop is added for every site until at some point the maximum number of stem loops is acquired. For the remaining part of the transcription process, the number of stem loops stays constant. After termination, the mRNA is released and rapidly diffuses away from transcription site. The corresponding spot is thus no longer visible and we observe a sharp drop in intensity. Hence, if $a \in \mathbb{N}^{L+1}$ encodes the number of stem loops associated with the sites of the gene, the variable

$$N(t) = \sum_{i=0}^{L} \alpha_i X_i(t) \tag{S3}$$

corresponds to the number of visible stem loops at time $t$. Assuming that stem loops are occupied by GFP fast compared to the elongation speed, the total spot signal can be described as

$$I(t) = b_0 + e^{-\lambda t}\left(b_1 + \gamma N(t)\right) \,. \tag{S4}$$

Here, $\lambda$ is the bleaching factor and $b_0$, $b_1$ correspond to baseline background and a bleachable part of the background respectively. The factor $\gamma$ encodes the intensity contribution per GFP molecule. During image acquisition and intensity estimation, the signal is corrupted by various forms of noise such as z-diffusion of the transcription site, photon counting noise on the camera chip, variations in the media, mismatch of the point-spread function with the Gaussian approximation, irregular background illumination, etc. We subsume all these effects into a single multiplicative noise variable leading to the relation

$$Y_1(t_k) = I(t_k)\exp(\sigma\epsilon_k) \,, \tag{S5}$$

where $\epsilon_k$ are independent and standard normally distributed. While (S10) is a reasonable approximation for larger signals, it is not suitable for very small signals. The reason for this is that at low intensity, due to fundamental limitations of the spot detection, there is an increased probability that a spot is missed or that a local fluctuation is confused with a signal. To take account of this effect, we introduce the random background signal

$$Y_0(t_k) = I_0\exp(\sigma\epsilon_k)$$

and an additional unobserved random variable $Z(t_k) \in \{0, 1\}$ with

$$\Pr(Z(t_k) = 1 \mid I(t_k)) = \mathrm{sigmoid}\left(w \log\left(\frac{I(t_k)}{I_0}\right)\right)$$

where $\mathrm{sigmoid}(x) = (1 + \exp(-x))^{-1}$ and $w$ is a response parameter. The final observation model is then given by

$$Y(t_k) = Z(t_k)Y_1(t_k) + (1 - Z(t_k))Y_0(t_k) \,. \tag{S6}$$

This can be understood as a soft threshold for spot detection. When the true spot intensity $I$ is larger than $I_0$, $Z$ will likely be one and we measure the signal $Y_1$ with high probability. When $I$ is smaller than $I_0$, $Z$ is likely zero and we observe the spurious signal $Y_0$ with high probability. The likelihood corresponding to (S11) is that of a mixture of two lognormal distributions with parameters $\log(I)$ and $\log(I_0)$. An overview of all parameters involved in kinetic and observation model is given in Table S14.

### S4.3 Elongation times

Many models for transcription assume independent movement of individual polymerases. This leads to a simple relation between elongation rate and elongation time as $t_e = \frac{l}{k_e}$ where $l$ is the size of the translated region. Due to possible interactions between polymerases, this relation is not valid in the TASEP model. As shown in Fig. S14, the TASEP model requires a higher value compared to an independence model to produce the same expected elongation time. This difference becomes smaller for higher elongation rates, since a fast movement of individual polymerases decreases the probability of interaction. As an example, consider the dotted black line in Fig. S14 corresponding to an elongation time of 12 s. To produce such an elongation time, the independent model requires an elongation rate of 100 nt $s^{-1}$, while the TASEP model requires and elongation rate of roughly 120 nt $s^{-1}$.

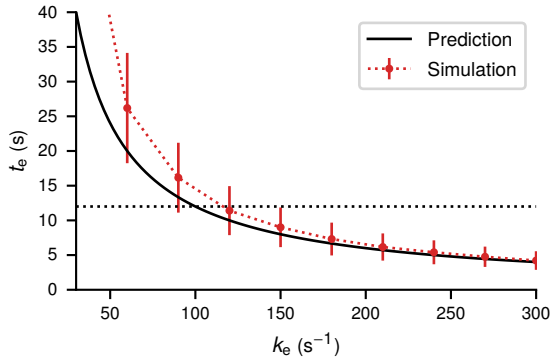


**Figure S9.** Expected elongation times of a gene template with size $l = 1200$ nt for different values of the elongation rate $k_e$. The thick black curve indicates the value obtained from independent polymerase movement by the relation $t_e = \frac{l}{k_e}$. Red dots show the expected elongation time from simulations of a TASEP model with a site size of 120 nt. Error bars indicate the standard deviation of the distribution. Initiation rate $k_i$ and termination rate $k_t$ are fixed and chosen such that typically multiple polymerases are on the template and no traffic jams are caused by the termination site.

### S4.4 Coarse-graining

The most natural quantization of the gene into sites would be to associate every site with a single nucleotide. This would, however, lead to more than 1000 sites and since the state space of the model scales as $2^L$ would make inference intractable. In addition, we only observe the system by one-dimensional summary statistic every few seconds such that most of the detailed dynamics are not captured. It is thus necessary to combine several nucleotides into a single site. A good candidate for such a coarse-graining is the DNA footprint of a stem loop, which is roughly 60 nt, as the appearance of a stem loop is the most fine-grained observable. For a DNA template consisting of 1200 nt, this would lead to $L = 22$ sites. While the master equation this model is still tractable, it requires significant computational effort such that only a small number of cells can be handled this way. As argued in the main text, it is important to pool many traces in order to overcome structural identifiability limitations of the system. As a compromise between tractability and resolution, we choose a partition size of 120 nt corresponding to two stem loops leading to a system of $L = 12$ sites.

The coarse-graining changes the waiting time distribution between appearance of two stem loops. If one starts from a fine-grained model where every site corresponds to a single nucleotide, the waiting time for jumps of size 120 bp is much more peaked compared to the exponential distribution. To investigate the robustness of the inference against this kind of mismatch, we simulated 100 trajectories from a fine-grained model described in[23] Supp. M. The model is similar to the TASEP model we use for inference, but uses one site for every nucleotide. In addition, RNAP molecules have a footprint of 40 sites and individual stems loops appear every 60 sites with 14 stem loops in total. The number of sites was set to 1200 roughly corresponding to the gene investigates experimentally. The elongation rate was set to $k_{elong} = 100$ nt $s^{-1}$, for all other parameters we used the standard values of the coarse-grained model. We generated 100 trajectories with 3 s time-lapse and performed pooled inference. For simplicity, we assume a constitutive promoter an drop the switching site. The results are presented in Fig. S15a. They indicate that while initiation and termination rate are inferred quite accurately, the estimated elongation rate is, however, significantly larger than the ground truth used to generate data in the fine-grained model. This can be explained by the exclusion property of the TASEP process. In the coarse-grained model, collisions will happen more frequently on average since the space is limited. In order to achieve a similar total elongation time, the rate has to be increased to account for the possible blocking. To verify this, we generated trajectories from fine-grained model and coarse-grained model and extracted the distribution of total transcription times (Fig. S15b). The mean elongation time predicted by the learned coarse-grained model is quite close to the mean elongation time of the fine-grained model. This demonstrates that even though there is a mismatch, the model learns physically meaningful quantities. However, we stress that this is merely an example to illustrate model mismatch. In practice, the distribution of elongation times predicted by the fine-grained model is likely too narrow as it ignores effects such as pausing, reverse steps, chromatin remodelling, etc. Thus, a coarse-grained model with wider waiting time distribution may be a better approximation to the real data, as long as these effects are not modeled explicitly.

### S4.5 Overview of the model parameters

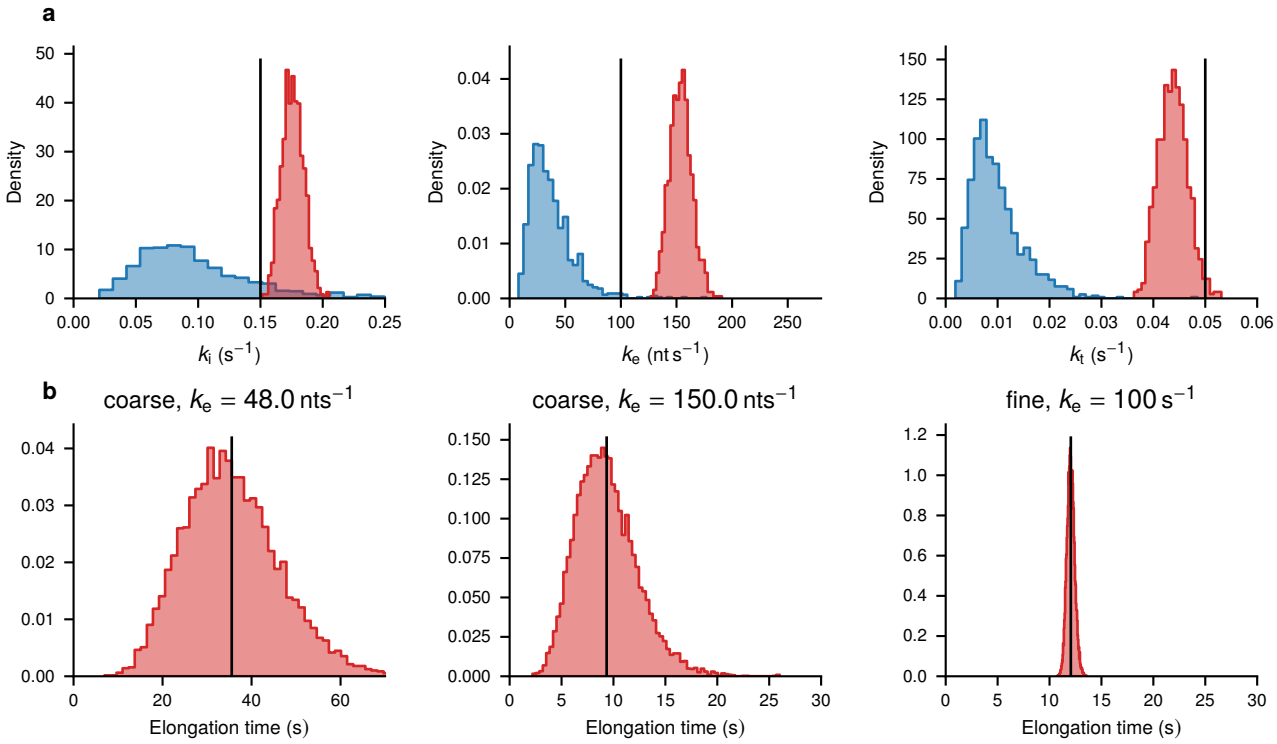


**Figure S10. a** Pooled posterior inference of a non-switching model on traces generated by the fine-grained TASEP model with a time laps of 3 s. Observation parameters are not shown but were also estimated during inference. The rows show histogram approximations of the prior distribution (blue) and the posterior distribution (red) for the model parameters. Black lines indicate the parameter value used to generate the data. **b** Distribution of total elongation times obtained from forward simulation of the coarse-grained model (left and middle) and the fine-grained model (right). The elongation right of the left plot is a typical value of the prior distribution, the elongation rate of the middle plot is a typical value of the posterior distribution. Black lines indicate the empirical mean.

**Table S7.** Overview of the model parameters

| Parameter | Default value | Explanation |
|---|---|---|
| $k_{on}$ | $0.025\,s^{-1}$ | On-switching rate of the promoter state |
| $k_{off}$ | $0.025\,s^{-1}$ | Off-switching rate of the promoter state |
| $k_i$ | $0.3\,s^{-1}$ | Initiation rate of polymerases, given the promoter is in the active state |
| $k_e$ | $100\,nt\,s^{-1}$ | Elongation rate per polymerase. Must be divided by the site size to convert to a jump rate between sites |
| $k_t$ | $0.3\,s^{-1}$ | Termination rate from the pooled termination site. $\frac{1}{k_t}$ is the average time a transcript is visible at the transcription site after elongation. |
| $L$ | 11 | Number of TASEP lattice sites. Due to the additional promoter site $X_0$, the full model has $L+1$ sites. |
| $b_0$ | 5 a.u. | Non-bleachable baseline fluorescence intensity |
| $b_1$ | 5 a.u. | Bleachable background fluorescence intensity |
| $\gamma$ | 1.1 a.u. | Fluorescence intensity per unit of GFP |
| $\lambda$ | $1 \times 10^{-3}\,s^{-1}$ | Exponential bleaching rate |
| $I_0$ | 25 a.u. | Soft detection threshold of a true spot signal |
| $w$ | 4 | Tuning parameter to determine the sharpness of the soft detection |
| $\sigma$ | 0.1 | Noise level standard deviation in the log-domain of the signal |

## S5 Experimental details

### S5.1 Yeast strains and plasmids

For live transcript analysis, we engineered the haploid strains of *Saccharomyces cerevisiae* (BY4742 and BY4741), which are isogenic to S288C (Research Genetics/Invitrogen, Huntsville, AL). 14X PP7 binding sites (hairpins) were amplified from pTL031[36] using primers T1053, T1054 and integrated in one of the haploids by replacing one of the *CUP1* ORFs in the yeast genome by homologous recombination[37]. For expressing PP7-GFP coat protein, we constructed a MET3 integrative vector (pTSK630) to express PP7-NLS-GFP from *SEC61* promoter and integrated this vector by SacI XhoI digestion in both the haploids. This vector can be available upon request. For the photobleaching correction, we used a diploid strain YTK1231 in which both the *CUP1* arrays are replaced by 256 copies of *LacO* and lacI-GFP-NLS is expressed from *pHIS3*. For preparing the calibration curve for the number of GFP molecules verses brightness/intensity, we used three yeast strains (YTK541, YTK1231 and YTK1268) with known numbers of GFP molecules per locus. YTK541, contains a tandem array of 10 copies of *CUP1* locus with 40 binding sites for the transcription activator Ace1p-GFP. *CUP1* is activated by Cu, and at the peak of activity *CUP1* array binds 120 molecules of GFP. In YTK1231, each lacO binding site may be associated with a dimer of the lac Repressor (LacI-GFP)[38]. Therefore, the array of 256 tandem lacO binding sites is associated with 512 LacI-GFP molecules. In YTK1268, the spindle pole body of the diploid yeast strain contains approximately 1000 molecules of Spc42-GFP. Strain genotypes are provided in Table S8, plasmids in Table S9 and primer sequences in Table S10.

### S5.2 Media and growth conditions

For live transcript analysis, cells of YTK1799 were plated on CSM-URA plate (from −80 °C frozen glycerol stock) and grown for 48 h at 28 °C. 3 to 5 colonies were inoculated in 3 mL CSM-URA media (in 14 mL polypropylene tubes, Cat no. 352059, Falcon, Maxico) and grown for overnight at 28 °C, 230 RPM. 250 µL of this overnight grown culture was inoculated in 25 mL CSM-URA (in 250 mL flask) and grown at 28 °C, 230 RPM for 24 h. This flask was removed from the shaker and kept in refrigerator at 4 °C. We used this refrigerated culture for daily inoculations for a month to get consistent results (to avoid day to day variations in transcription induction kinetics due to difference in the age of the culture). From this refrigerated culture, we inoculated 60 µL in 3 mL of fresh CSM-URA media (in 14 mL polypropylene tubes, Cat no. 352059, Falcon, Maxico) in the morning and grew the cultures for 5 h at 28 °C, 230 RPM. Cells were harvested by centrifugation (2200 RPM for 1 min) and cells were placed under the CSM-URA agarose pad (100 µM $CuSO_4$) for imaging. For the photobleaching correction and GFP calibration curve, strains YTK541, YTK1231 and YTK1268 were grown under the same conditions, except YTK541 and YTK1268 were grown in CSM-HIS media.

### S5.3 Quantitative RT-PCR (RT-qPCR)

Samples were harvested at indicated time points after Cu induction. RNA was extracted (from yeast cells) using the ISOLATE II RNA Mini kit (Bioline, UK, Cat no. BIO-52072). cDNA was prepared using the iScript cDNA synthesis kit (BioRad, Cat no.: 1708891) starting with 1 mg of total RNA. Quantitative real-time PCR (qPCR) was performed as described[27]. For normalization, the expression of the housekeeping gene ACT1 was quantified. Primers used for this quantification are listed in Table S10 (T531, T532 for CUP1 and T1055, T1056 for ACT1). To confirm the absence of contaminating genomic DNA in cDNA preparations, reverse transcriptase negative (-RT) samples were used as a control, which produced the Ct value difference of >10 cycles between "-RT" and "+RT"samples, indicating a negligible amount of genomic DNA contamination in cDNA samples. mRNA extraction, cDNA synthesis, and qPCR were repeated at least twice, and qPCR was performed in duplicates for each experiment. Error bars indicate SEM.

### S5.4 Microscope settings and imaging conditions

For imaging live cells, 5 h grown cultures were harvested by centrifugation (2200 RPM for 1 min) and 3 µL of cell pellet were placed in Lab-Tek II chambered coverglass (1.5 Borosilicate Gass, Nunc, ThermoFisher Scientific, MA, US), mixed with equal volume of 200 µM $CuSO_4$ containing CSM-URA and covered by 1cm x 1cm CSM-URA agarose pad (100 µM $CuSO_4$). A timer was started immediately upon mixing the cells with 200 µM $CuSO_4$ containing CSM-URA. 3D time-lapse movies were acquired at the room temperature on the DeltaVIsion Elite Microscope, using 100x 1.4 NA oil immersion objective lens, sCMOS camera, FITC filter set (15 ms exposure, Ex 488/27; Em 505/45, Chroma Technology Corp, Bellows Falls, VT), 15 z-steps at every 400 nm, 1x1 binning and 1024x1024 pixels. Time-lapse movies with 3 s time interval were acquired for 90 s, followed by changing the field within next 90 s, imaging new field for another 90 s and so on. This imaging regime was repeated 9 times within 30 min to cover the entire slow cycle. E.g. to cover the entire slow cycle of 30 min with 3 s time interval, first set of movies were started for new field of cells after 3, 6, 9, 12, 15, 18, 21, 24, 27 min after copper induction and acquired for 90 s. Remaining 90 s (between 3 min and 6 min time points, between 6 min and 9 min time points, and so on) were used for moving to

the next field of cells. Second set of movies were acquired for new field of cells after 4.5, 7.5, 10.5, 13.5, 16.5, 19.5, 22.5, 25.5 and 28.5 min after copper induction and acquired for 90 s to compensate the missing time points from the first set. Similarly, time-lapse movies for 12 s time interval were recorded for 5 min, followed by changing the field of view within the next 1 min. A first set of movies was started after 3, 9, 15, 21, 27 min after induction. A second set of movies was stared at 6, 12, 18, 24 min after induction. For the photobleaching correction, strain YTK1231 was imaged under the same condition with 3 s time interval.

**Table S8.** Yeast strains used in this study

| Strain ID | Application | Source | Genotype |
|---|---|---|---|
| YTK541 | GFP calibration | This study | MATα, his3-D1 leu2-D0 ura3-D0 lys2-D0 ace1-D1::KAN TRP1::TRP1ORF-pCap2-ACE1-tripleGFP-HIS3 |
| YTK1231 | GFP and photobleaching calibration | This study | MATa/MATα, his3-D1/his3-D1 leu2-D0/leu2-D0 trp1/TRP1 ura3-D0/ura3-D0 lys2-d::pHIS3-lacI-GFP-NLS-Nat1/ lys2-d::pHIS3-lacI-GFP-NLS-Nat1 MET15/met15-D0 Cu1::KAN-(LacO)256/Cu3::(LacO)256 [pRS426 pHIS3-LacI-GFP-URA3] |
| YTK1268 | GFP calibration | This study | MATa/MATα, his3-d1/his3-d1 leu2-d0/leu2-d0 met15-d0/MET15 LYS2/lys2-d0 ura3-d0/ura3-d0 SPC42-GFP-HIS3/SPC42-GFP-HIS3 |
| YTK1799 | 3s movies, 12s movies | This study | MATa/MATα, his3-d1/his3-d1 leu2-d0/leu2-d0 lys2-d0/LYS2 MET15/met15-d0 ura3-d0/ura3-d0 pdr5-d::LoxP/pdr5-d::LoxP trp1d::pADH-AFB2::LEU2/trp1d::pADH-AFB2::LEU2 ace1-d::pCAP2-ACE1-mCherry-TRP1/ ace1-d::pCAP2-ACE1-mCherry-TRP1 RSC2-AID-9Myc::HIS5/RSC2-AID-9Myc::HIS5 CUP1-PP7hairpins(14)-KANMX/CUP1 MET3::pSEC61-PP7-GFP-CTCT-URA3/MET3::pSEC61-PP7-GFP-CTCT-URA3 |

**Table S9.** Plasmids used in this study

| Plasmid Name | Application | Primers | Source |
|---|---|---|---|
| pTL031 | For amplifying 14x PP7 binding sites (hairpins) | T1053-T1054 | Lenstra and Larson, 2016[36] |
| pTSK630 | For integrating pSEC61-PP7-GFP-CYCT at MET3 locus using URA3 (K. lactis) as a selection marker. Cut with SacI and XhoI to integrate into MET3 | | This study |

**Table S10.** Sequences of the primers used in this study

| Primer ID | Application | Sequence |
|---|---|---|
| T1053 | For replacing one copy of *CUP1* ORF with *14x PP7* binding sites (Forward) | gatattaagaaaaacaaactgtacaatcaatcaatcaat-catcacataaa gtaaaacgacggccagtgagcg* |
| T1054 | For replacing one copy of *CUP1* ORF with *14x PP7* binding sites (Reverse) | aaaattaaaaacagcaaatagttagatgaatatattaaagactattcgtt gttttcgacactggatggcggcg* |
| T531 | RT-qPCR | CATTTCCCAGAGCAGCATGA |
| T532 | RT-qPCR | GTCATGAGTGCCAATGCCAA |
| T1055 | RT-qPCR | ggttgctgctttggttattgataacgg |
| T1056 | RT-qPCR | gttcttctggggcaactctc |

*Homologous sequences for site-specific integration are shown in blue. Sequences homologous to plasmid DNA are shown in red.

**Table S11.** Structures used for GFP calibration

| Strain | Sub-cellular structure | Expected GFP |
|---|---|---|
| YTK541 | CUP1 array / Ace1p-3xGFP | 120 |
| YTK1231 | lacO (256) / LacI-GFP | 512 |
| YTK1231Enh | lacO (256) / LacI-GFP | 512 |
| YTK1268 | SPB/ Spc42-GFP | 1000 |

**Table S12.** Datasets collected in this study

| Dataset | Strain | No.Movies |
|---|---|---|
| 3s | YTK1799 | 82 |
| 12s | YTK1799 | 71 |
| Photobleaching 3s | YTK1231 | 25 |
| Photobleaching 12s | YTK1231 | 20 |
| GFP calibration | YTK541, YTK1231, YTK1799, YTK1268 | 20 |

## S6 Data acquisition and pre-processing

### S6.1 Cell selection

As a first step, we extracted movies of individual cells from the full movies. In order to extract cells with spots, we computed two projections: a maximum projection over all frames leading to a 2D movie and a maximum projection over all time frames and slices producing a single image. The latter single image allowed to identify all cells that develop an active transcription site during imaging. For the identified cells we drew ROIs in the initial time frame of the 2D video. These initial ROIs were tracked over time and used to extract 3D movies of individual cells. For each cell, the full image stack over time was stored for trace extraction. Fluorescence intensity traces were obtained as described in Sec. S6.2. A more detailed description of the custom tracker is given in Sec. S12.

### S6.2 Trace extraction

To track fluorescence spots and quantify fluorescence levels, we developed a custom 3D method based on sequential filtering. The location of the spot within an image stack at time $t_k$ is given by $r_k = (x_k, y_k, z_k)$. The spot is modeled as a diffraction limited point source such that the intensity $\hat{I}$ of the pixel at position $r$ can be described as

$$\hat{I}_k(r) = b_k + I_k \exp\left[-\frac{1}{2}(r - r_k)^T \Sigma_{\mathrm{PSF}}^{-1}(r - r_k)\right]$$

where the diagonal matrix $\Sigma_{\mathrm{PSF}}$ describes the shape of the point spread function of the optical system, $b_k$ is a local background and $I_k$ is the peak intensity of the spot that is related to the underlying intensity of the point source by a constant factor. In addition, we introduce a binary variable $s_k$ that describes whether the spot is visible in frame $k$. This leads to a total state

$x_k = (r_k, b_k, y_k, s_k)$. The observation model is given by the predicted intensity per pixel with additive Gaussian noise. State estimation is carried out using the standard recursive filtering approach[43]. Due to the non-linear observation model, a Laplace approximation is used to evaluate filter updates. The details are given in S12. The filter provides estimates of $I_k$ for $k = 1, \ldots, n$ which are treated as the noisy measurements $y_1, \ldots, y_n$ for the main analysis. Traces showing dividing spots due to DNA replication were excluded from further analysis. In this study we do not take into account the noise from the fraction of the cells that completed the replication of the reporter gene, but did not separate sister chromatids, which may lead to transcription of two alleles in the same spot.

### S6.3 Quality Control

In the images, some of the cells showed too high or too low GFP expression or were otherwise malfunctioning. In the 12s data set, we identified 3260 cells where an active TS is visible in at least one time frame during imaging. After tracking, all traces were compared manually to the corresponding cell video for quality checking. During this screening, we removed all potentially problematic traces, e.g. floating or dividing cells. The remaining 3036 high quality traces were used for further analysis. The same procedure was applied to the 3s data set leading to a selection of 3685 of an initial 4053 cells. A detailed summary of extracted cells per time interval is given in Table S13.

**Table S13.** Cells extracted from the 3 *s* and 12 *s* data sets

| Time since induction | 3*s* data set | | 12*s* data set | |
|---|---|---|---|---|
| | Extracted | Selected | Extracted | Selected |
| 3 min | 20 | 17 | 223 | 208 |
| 4.5 min | 78 | 68 | | |
| 6 min | 170 | 157 | 361 | 339 |
| 7.5 min | 227 | 202 | | |
| 9 min | 308 | 282 | 627 | 589 |
| 10.5 min | 322 | 285 | | |
| 12 min | 391 | 365 | 498 | 448 |
| 13.5 min | 315 | 300 | | |
| 15 min | 427 | 399 | 533 | 510 |
| 16.5 min | 284 | 270 | | |
| 18 min | 351 | 313 | 477 | 440 |
| 19.5 min | 248 | 228 | | |
| 21 min | 255 | 231 | 251 | 232 |
| 22.5 min | 143 | 120 | | |
| 24 min | 214 | 195 | 172 | 162 |
| 25.5 min | 108 | 93 | | |
| 27 min | 125 | 105 | 118 | 108 |
| 28.5 min | 50 | 40 | | |
| 30 min | 17 | 15 | | |

## S7 Exploratory analysis

Before turning to the model-based Bayesian inference approach, we performed descriptive statistical analysis to confirm the cyclic activity of the *CUP1* promoter known from earlier studies.

### S7.1 Total image brightness

The movies within different dataset were collected on different days. To ensure that these movies are comparable, we check the total intensity distributions of the first time frame of every movie. By focusing on the first time step, we can avoid confounding effects of bleaching. Boxplot representations of the intensity distributions (Fig. S11) reveal some variations in the intensity distribution. To understand these variations, it is important to note that the images consist mostly of extra-cellular background. Biological differences mainly affect the nuclear brightness and can thus not be responsible for large distribution shifts. Possible factors that could affect total brightness are the agarose medium in the background or variations in illumination or other factors of the optical system.

The most striking difference is that the movies collected on day 4 in the 3s dataset are significantly brighter than all other movies in this dataset. To make the dataset more homogenous, we exclude the day 4 movies from all further analysis. We also

observe that the 12s dataset is significantly brighter than the 3s dataset. The reason for this difference is unclear as several months are between the recording dates. However, as we analyze the dataset separately and optical factors such as the gfp scaling are inferred from the dataset, we do not view this as a problem. The third observation we discuss here is that the brightness tends to decrease for movies take from the same coverslip. This is most likely due to some cross-bleaching from imaging neighboring positions. Since the effect is rather small, we will ignore it in the following analysis.

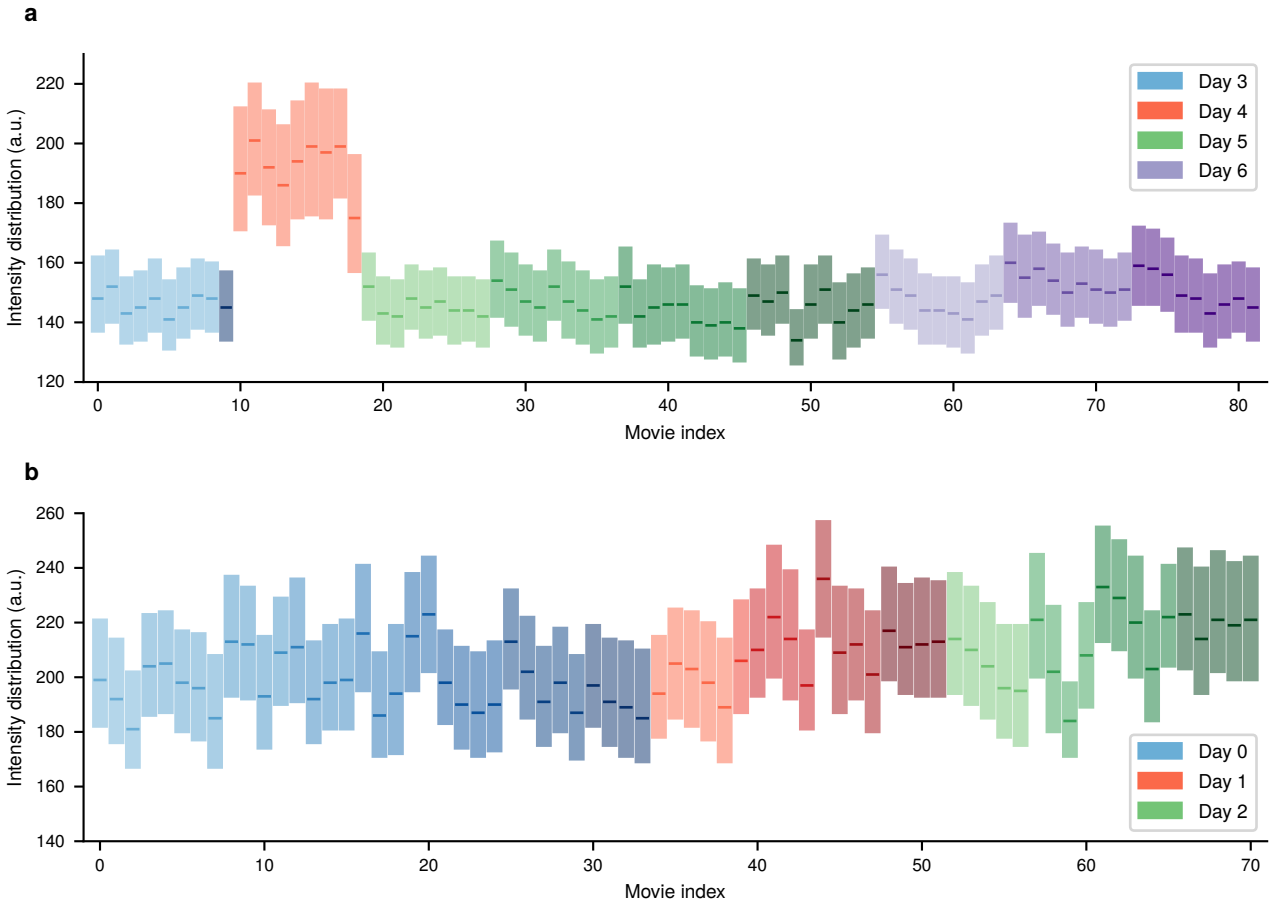


**Figure S11.** Pixel intensity distributions for the first time frame of every movie in form of a box plot shown for the 3 s dataset (**a**) and the 12 s dataset (**b**). Lower and upper border of the rectangles represent 25th and 75th percentile, respectively. The bold line within the box indicates the median. Every color family represents one experiment day, while the shading withing a color family indicates movies taken from the same coverslip

### S7.2 Spot Brightness

We now take a first look at the spot brightness. Similarly as an the previous section, we consider only the first frame of every movie to avoid effects of bleaching. For every movie, the spot intensities of all active cells in the first frame were pooled. The corresponding distributions are shown in the form of boxplots for both 3s and 12s dataset in Fig. S12.

Especially in the 3s dataset one can observe a clear dependence of spot brightness on the time since induction. Since the movies taken from the same coverslip are ordered sequentially, they correspond to spot measurements every 3 minutes or every 6 minutes for 3 s dataset and 12 s dataset, respectively. The observed sharp increase followed by a longer decrease is well in accordance with measurements from RT-qPCR. In addition, the results from different coverslips are quite similar and show less variation than the total intensity distributions in Fig. S11. The reason for this is that spot intensities are measured relative to local nuclear background such that additive noise affecting the total brightness of images is filtered out by the spot tracking. The one exception from this are the movies from day 4, which we have already decided to treat as outliers due to their much higher overall brightness.

The dependence on time since induction is also visible in the 12 s dataset albeit less clear than for the 12 s dataset. One reason

for this is that there are fewer measurements per coverslip due to the longer duration of the individual movies. In addition, some series of the movies are shifted by 3 minutes leading to a less regular appearance.

Finally, it seems that the spots in the 12 s dataset are overall brighter than in the 3 s dataset. Two possible explanations come to mind. The first possibility is that the cell cultures differ in activity which is possible due the time between the collection of the datasets. Alternatively, multiplicative noise affecting the total intensity distribution could explain the differences as it would not be filtered out by the spot tracking. In practice, it could also be a combination of both factors. We deal with this by estimating the observation model (cf. Sec. S9) parameters individually for every dataset. Any multiplicative noise will be captured by the estimated GFP scaling factor such that remaining differences can be attributed to Biological causes.

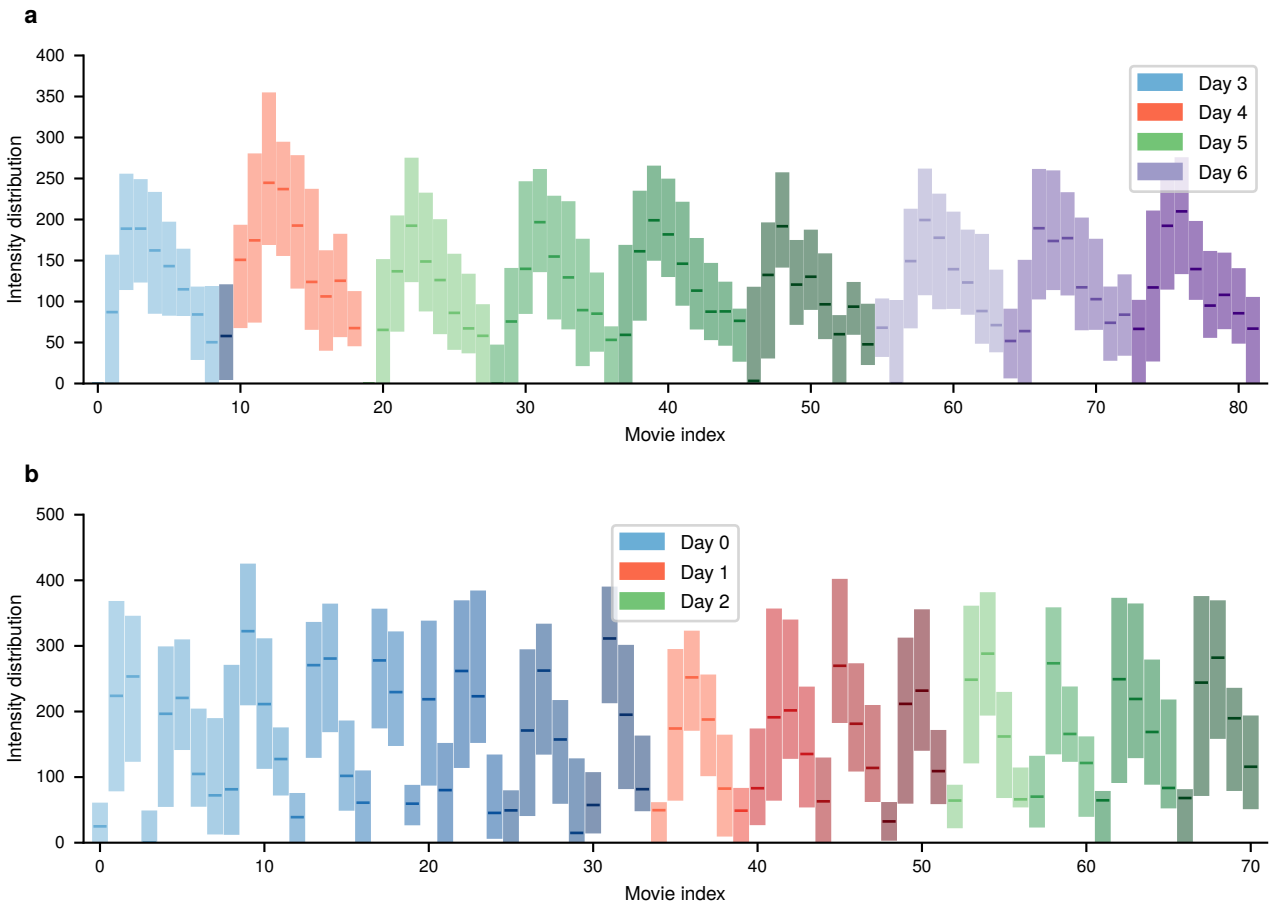


**Figure S12.** Pixel intensity distributions for the first time frame of every movie in form of a box plot shown for the 3 s dataset (**a**) and the 12 s dataset (**b**). Lower and upper border of the rectangles represent 25th and 75th percentile, respectively. The bold line within the box ndicates the median. Every color family represents one experiment day, while the shading withing a color family indicates movies taken from the same coverslip.

### S7.3 Investigating time-dependant activity

The presence of non-stationary transcription dynamics can be already seen from the movie histograms in Fig. S12. We will now take a closer look by inspecting the average spot intensities pooled over all movies that started with the same time delay with respect to induction. As before, we will only consider the first time frame of each movie to avoid the effects of bleaching. The first row in Fig. S13 shows the average spot intensity over the cycle. For both 3 s dataset (left) and 12 s dataset (right) wie see a sharp rise of the intensity in the early part of the observed interval followed by a slower decay. While the general shape of both curves is similar, the spots in the 12s dataset are generally brighter as we have seen from the per-movie distributions (cf. Fig. S12). The second row of Fig. S13 shows the number of responding cells divided by the total number of cells per time window. To assign cells to the class of responders or non-responders, we used a standard Gaussian mixture classifier on the spot intensity distributions of the datasets. In order to avoid bias by the overall different brightness, we applied this classifier individually on

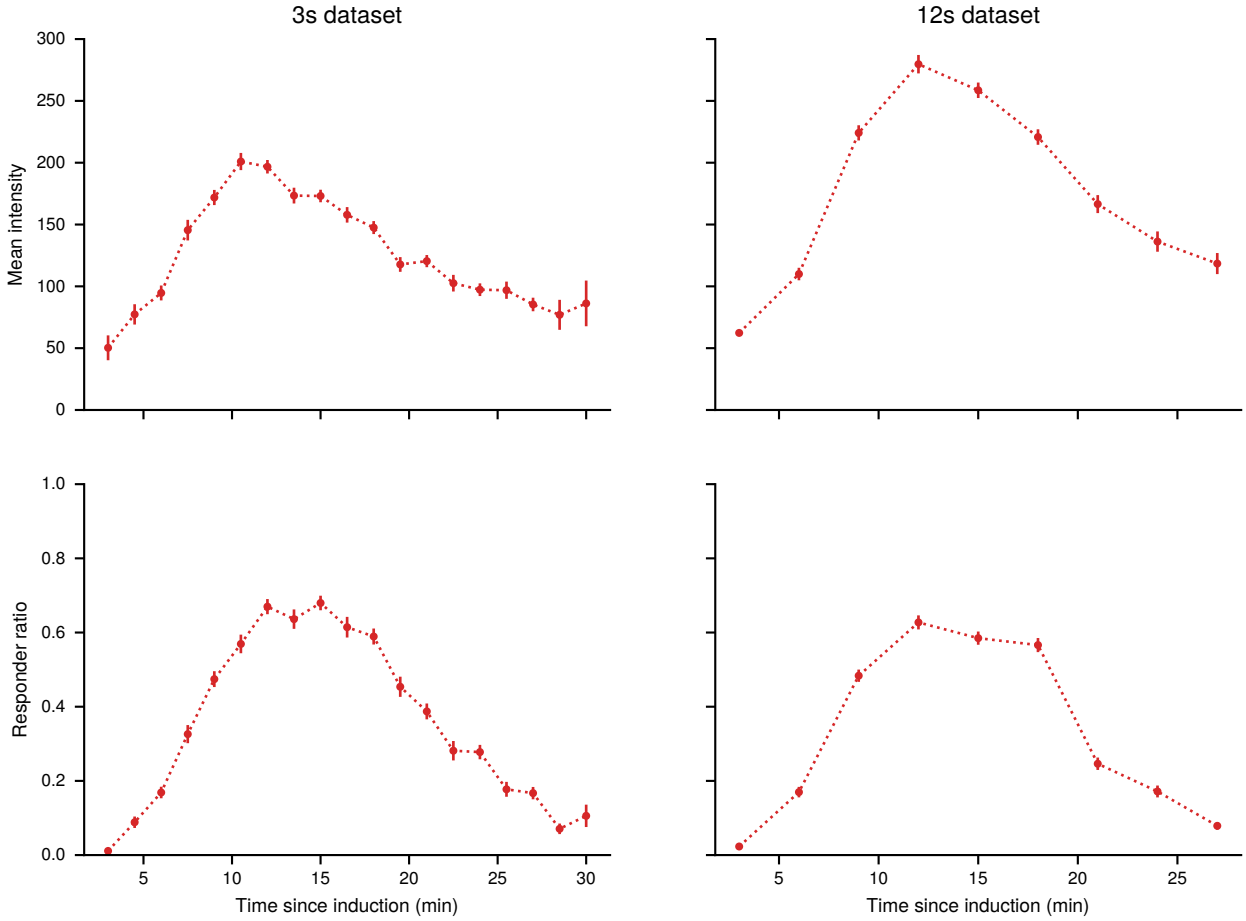


**Figure S13.** Dependence of summary statistics on the time since induction. The first row shows the mean of the spot intensities pooled over all videos with the same time delay since induction. The second row shows the number of responders divided by the total number of cells per time window. Error bars indicate standard error. Results for the 3 s dataset are also shown in the main text (Fig. 1f, g)

each dataset. The resulting responder ratios of both datasets agree well.

## S8 Model

### S8.1 Kinetic Transcription Model

As described in the main text, the model is an example of a Markov jump process that satisfies the master equation (1). The transition function $Q$ defines the probability of an event to happen in an infinitesimal interval $h$

$$P(X(t+h) = x' \mid X(t) = x) = Q(x, x' \mid \theta)h + o(h)\,.$$

The transition function is fully specified by the vector of parameters $\theta = (k_{\text{on}}, k_{\text{off}}, k_{\text{i}}, k_{\text{e}}, k_{\text{t}})^\top$ and the conditions under which transitions can occur, leading to

$$Q(x, x' \mid \theta) = \begin{cases} k_{\text{on}} & x_0 = 0, x_0' = 1\,, \\ k_{\text{off}} & x_0 = 1, x_0' = 1\,, \\ k_{\text{i}} & x_0 = 1, x_1 = 0, x_1' = 1\,, \\ k_{\text{e}} & x_i = 1, x_{i+1} = 0, x_i' = 0, x_{i+1}' = 1\,, \quad i = 1, \ldots, L-1\,, \\ k_{\text{t}} & x_l' = x_l - 1 \end{cases}.$$

For parameter inference, we need to evaluate the system for many different parameter configurations. It is therefore convenient to represent the transition functions as

$$Q(x, x' \mid \theta) = \sum_{i=1}^{5} \theta_i R_i(x, x') \,. \tag{S7}$$

The operators $R_i$ are independent of the parameters. By enumerating the states of the system, the probability $p(x, t)$ can be represented by a vector $\mathbf{p}(t)$ and the transition function $Q$ becomes a sparse matrix $\mathbf{Q}$. With this, the master equation becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{p}(t) = \mathbf{Q}\mathbf{p}(t) \,.$$

A formal solution of this system is given by the matrix exponential

$$\mathbf{p}(t) = \exp\left(\mathbf{Q}t\right)\mathbf{p}_0$$

with initial distribution $\mathbf{p}_0$. For sparse $\mathbf{Q}$, this can be efficiently solved for fairly large state spaces by the Krylov subspace approximation for matrix exponentials[41,42].

### S8.2 Observation Model

The kinetic model discussed above is a continuous time model. In practice, one cannot observe such a systems in continuous time but rather at discrete sample times $t_1, \ldots, t_n$. In addition, we do not observe the process $X(t_k)$ directly. Here, our measurement $Y(t_k)$ is provided by the total intensity of the fluorescence spots as measured by the tracking algorithm. In the following, we construct a model that relates $X$ and $Y$. First, note that as the polymerase traverses the gene, an additional stem loop is added for every site until at some point the maximum number of stem loops is acquired. For the remaining part of the transcription process, the number of stem loops stays constant. After termination, the mRNA is released and rapidly diffuses away from transcription site. The corresponding spot is thus no longer visible and we observe a sharp drop in intensity. Hence, if $a \in \mathbb{N}^{L+1}$ encodes the number of stem loops associated with the sites of the gene, the variable

$$N(t) = \sum_{i=0}^{L} \alpha_i X_i(t) \tag{S8}$$

corresponds to the number of visible stem loops at time $t$. Assuming that stem loops are occupied by GFP fast compared to the elongation speed, the total spot signal can be described as

$$I(t) = b_0 + e^{-\lambda t}\left(b_1 + \gamma N(t)\right) \,. \tag{S9}$$

Here, $\lambda$ is the bleaching factor and $b_0$, $b_1$ correspond to baseline background and a bleachable part of the background respectively. The factor $\gamma$ encodes the intensity contribution per GFP molecule. During image acquisition and intensity estimation, the signal is corrupted by various forms of noise such as z-diffusion of the transcription site, photon counting noise on the camera chip, variations in the media, mismatch of the point-spread function with the Gaussian approximation, irregular background illumination, etc. We subsume all these effects into a single multiplicative noise variable leading to the relation

$$Y_1(t_k) = I(t_k)\exp(\sigma\epsilon_k) \,, \tag{S10}$$

where $\epsilon_k$ are independent and standard normally distributed. While (S10) is a reasonable approximation for larger signals, it is not suitable for very small signals. The reason for this is that at low intensity, due to fundamental limitations of the spot detection, there is an increased probability that a spot is missed or that a local fluctuation is confused with a signal. To take account of this effect, we introduce the random background signal

$$Y_0(t_k) = I_0\exp(\sigma\epsilon_k)$$

and an additional unobserved random variable $Z(t_k) \in \{0, 1\}$ with

$$\Pr(Z(t_k) = 1 \mid I(t_k)) = \mathrm{sigmoid}\left(w \log\left(\frac{I(t_k)}{I_0}\right)\right)$$

where $\mathrm{sigmoid}(x) = (1 + \exp(-x))^{-1}$ and $w$ is a response parameter. The final observation model is then given by

$$Y(t_k) = Z(t_k)Y_1(t_k) + (1 - Z(t_k))Y_0(t_k) \,. \tag{S11}$$

This can be understood as a soft threshold for spot detection. When the true spot intensity $I$ is larger than $I_0$, $Z$ will likely be one and we measure the signal $Y_1$ with high probability. When $I$ is smaller than $I_0$, $Z$ is likely zero and we observe the spurious signal $Y_0$ with high probability. The likelihood corresponding to (S11) is that of a mixture of two lognormal distributions with parameters $\log(I)$ and $\log(I_0)$. An overview of all parameters involved in kinetic and observation model is given in Table S14.

### S8.3 Elongation times

Many models for transcription assume independent movement of individual polymerases. This leads to a simple relation between elongation rate and elongation time as $t_e = \frac{l}{k_e}$ where $l$ is the size of the translated region. Due to possible interactions between polymerases, this relation is not valid in the TASEP model. As shown in Fig. S14, the TASEP model requires a higher value compared to an independence model to produce the same expected elongation time. This difference becomes smaller for higher elongation rates, since a fast movement of individual polymerases decreases the probability of interaction. As an example, consider the dotted black line in Fig. S14 corresponding to an elongation time of 12 s. To produce such an elongation time, the independent model requires an elongation rate of 100 nt $s^{-1}$, while the TASEP model requires and elongation rate of roughly 120 nt $s^{-1}$.

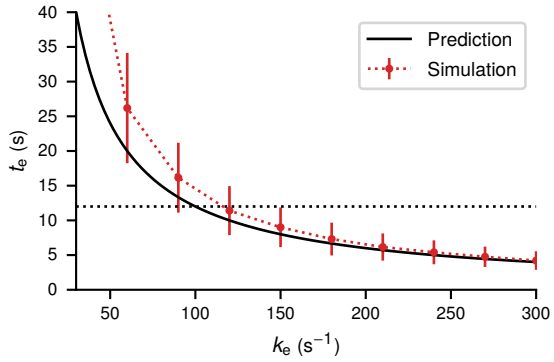


**Figure S14.** Expected elongation times of a gene template with size $l = 1200$ nt for different values of the elongation rate $k_e$. The thick black curve indicates the value obtained from independent polymerase movement by the relation $t_e = \frac{l}{k_e}$. Red dots show the expected elongation time from simulations of a TASEP model with a site size of 120 nt. Error bars indicate the standard deviation of the distribution. Initiation rate $k_i$ and termination rate $k_t$ are fixed and chosen such that typically multiple polymerases are on the template and no traffic jams are caused by the termination site.

### S8.4 Coarse-graining

The most natural quantization of the gene into sites would be to associate every site with a single nucleotide. This would, however, lead to more than 1000 sites and since the state space of the model scales as $2^L$ would make inference intractable. In addition, we only observe the system by one-dimensional summary statistic every few seconds such that most of the detailed dynamics are not captured. It is thus necessary to combine several nucleotides into a single site. A good candidate for such a coarse-graining is the DNA footprint of a stem loop, which is roughly 60 nt, as the appearance of a stem loop is the most fine-grained observable. For a DNA template consisting of 1200 nt, this would lead to $L = 22$ sites. While the master equation this model is still tractable, it requires significant computational effort such that only a small number of cells can be handled this way. As argued in the main text, it is important to pool many traces in order to overcome structural identifiability limitations of the system. As a compromise between tractability and resolution, we choose a partition size of 120 nt corresponding to two stem loops leading to a system of $L = 12$ sites.

The coarse-graining changes the waiting time distribution between appearance of two stem loops. If one starts from a fine-grained model where every site corresponds to a single nucleotide, the waiting time for jumps of size 120 bp is much more peaked compared to the exponential distribution. To investigate the robustness of the inference against this kind of mismatch, we simulated 100 trajectories from a fine-grained model described in[23] Supp. M. The model is similar to the TASEP model we use for inference, but uses one site for every nucleotide. In addition, RNAP molecules have a footprint of 40 sites and individual stems loops appear every 60 sites with 14 stem loops in total. The number of sites was set to 1200 roughly corresponding to the gene investigates experimentally. The elongation rate was set to $k_{elong} = 100$ nt $s^{-1}$, for all other parameters we used the standard values of the coarse-grained model. We generated 100 trajectories with 3 s time-lapse and performed pooled inference. For simplicity, we assume a constitutive promoter an drop the switching site. The results are presented in Fig. S15a. They indicate that while initiation and termination rate are inferred quite accurately, the estimated elongation rate is, however, significantly larger than the ground truth used to generate data in the fine-grained model. This can be explained by the exclusion property of the TASEP process. In the coarse-grained model, collisions will happen more frequently on average since the space is limited. In order to achieve a similar total elongation time, the rate has to be increased to account for the possible blocking. To verify this, we generated trajectories from fine-grained model and coarse-grained model and extracted the distribution of total transcription times (Fig. S15b). The mean elongation time predicted by the learned coarse-grained model is quite close to the mean elongation time of the fine-grained model. This demonstrates that even though there is a mismatch, the model learns physically meaningful quantities. However, we stress that this is merely an example to illustrate model mismatch. In practice, the distribution of elongation times predicted by the fine-grained model is likely too narrow as it ignores effects such as pausing, reverse steps, chromatin remodelling, etc. Thus, a coarse-grained model with wider waiting time distribution may be a better approximation to the real data, as long as these effects are not modeled explicitly.

### S8.5 Overview of the model parameters

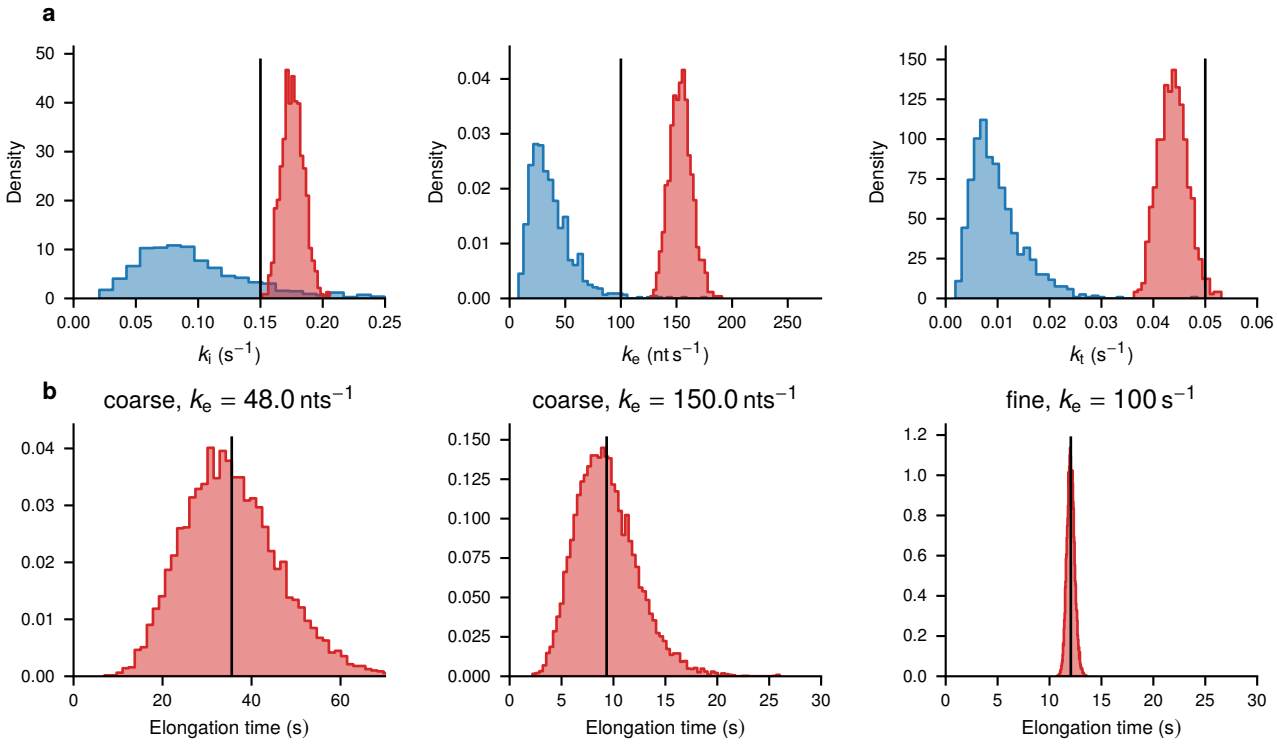


**Figure S15. a** Pooled posterior inference of a non-switching model on traces generated by the fine-grained TASEP model with a time laps of 3 s. Observation parameters are not shown but were also estimated during inference. The rows show histogram approximations of the prior distribution (blue) and the posterior distribution (red) for the model parameters. Black lines indicate the parameter value used to generate the data. **b** Distribution of total elongation times obtained from forward simulation of the coarse-grained model (left and middle) and the fine-grained model (right). The elongation right of the left plot is a typical value of the prior distribution, the elongation rate of the middle plot is a typical value of the posterior distribution. Black lines indicate the empirical mean.

**Table S14.** Overview of the model parameters

| Parameter | Default value | Explanation |
|---|---|---|
| $k_{on}$ | $0.025\,s^{-1}$ | On-switching rate of the promoter state |
| $k_{off}$ | $0.025\,s^{-1}$ | Off-switching rate of the promoter state |
| $k_i$ | $0.3\,s^{-1}$ | Initiation rate of polymerases, given the promoter is in the active state |
| $k_e$ | $100\,nt\,s^{-1}$ | Elongation rate per polymerase. Must be divided by the site size to convert to a jump rate between sites |
| $k_t$ | $0.3\,s^{-1}$ | Termination rate from the pooled termination site. $\frac{1}{k_t}$ is the average time a transcript is visible at the transcription site after elongation. |
| $L$ | 11 | Number of TASEP lattice sites. Due to the additional promoter site $X_0$, the full model has $L+1$ sites. |
| $b_0$ | 5 a.u. | Non-bleachable baseline fluorescence intensity |
| $b_1$ | 5 a.u. | Bleachable background fluorescence intensity |
| $\gamma$ | 1.1 a.u. | Fluorescence intensity per unit of GFP |
| $\lambda$ | $1 \times 10^{-3}\,s^{-1}$ | Exponential bleaching rate |
| $I_0$ | 25 a.u. | Soft detection threshold of a true spot signal |
| $w$ | 4 | Tuning parameter to determine the sharpness of the soft detection |
| $\sigma$ | 0.1 | Noise level standard deviation in the log-domain of the signal |

## S9 Calibration

The observation model as described in Sec. S8.2 contains a number of unknown parameters. In particular, the bleaching rate $\lambda$ and the GFP scaling factor $\gamma$ can have a significant effect on the inference results. It is therefore helpful to obtain independent estimates of these quantities from dedicated control data sets which can be used as prior distributions in the live cell inference procedure. The last subsection deals with the point spread function (PSF) of the optical system. While the PSF is not directly contained in the generative model, it is used to extract the intensity measurements from the images.

### S9.1 GFP-intensity scaling

For calibration measurements, we activated *CUP1* array in YTK541 with Cu and imaged the cells between 5 and 15 min of the Cu treatment, at the peak of their transcriptional activity to ensure that all the 40 binding sites are occupied with Ace1p-3xGFP fusion, and thus the total number of GFP per CUP1 array is 120. In YTK1231, we measured the brightness of the lacO/LacI-GFP array in telophase or G1 cells to ensure that we observe single non-duplicated array. LacI-GFP was present on multicopy plasmid, and some of the cells in the population were overexpressing LacI-GFP and thus displayed a high nuclear background. Therefore, we excluded the cells with abnormally bright nucleoplasm or arrays. In diploid cells of YTK1268, two peaks are observed in the SPB size distribution and about 5 % of the SPB in the population are significantly larger than average[39,40]. This happens because the SPB grows over time through the cell cycle. At G2 stage the SPB is split in two. Therefore, we avoided newly split SPB and selected only single SPB in the cells in telophase or in G1, where maximal size of the SPB is expected. As a control we also measured the cells in G2 with two SPB well separated by spindle and observed diminished brightness of those structures, as expected. Also, we selected the sub-population of SPB based on the range of brightness excluding from the measurements the SPB that were abnormally bright. Spot intensities of all structures were measured as for the live cell experiments (see Sec. S6.2) but only for a single time point.

A field view for each of the different strains used for this is shown in Fig. S16a. A first overview of the calibration data as shown in Fig. S16b reveals that the linear approximation assumed in (S9) is reasonable. In addition, the noise increases with expected number of GFP confirming the multiplicative noise model (S10). For a more detailed analysis, note that we have single images rather than videos as calibration data. Thus, with $t = 0$ and (S9) reduces to

$$I = b_0^* + \gamma N_{\mathrm{GFP}}\,, \tag{S12}$$

where $b_0^* = b_0 + b_1$ and $N_{\mathrm{GFP}}$ corresponds to the number of GFP associated with a particular structure. Together with the noise model (S10) and priors for $b_0^*$ and $\gamma$, we can perform Bayesian inference by Hamiltonian Monte Carlo. The obtained posterior is shown in Fig. S16c. It is well approximated by a Gamma distribution $\Gamma(\alpha, \beta)$ with $\alpha \approx 4226.4$, $\beta = 3822.8$. This distribution is used as a prior for the main inference part.

### S9.2 Bleaching Rate

To obtain an independent estimate of the bleaching rate, we recorded videos with 12 s intervals between observations with strain YTK1231 with *lac*O/LacI-GFP. Since GFP load is fixed for this structure, we expect any systematic changes in brightness over time to be caused by bleaching. This allows for an independent estimate of the bleaching factor. The mean intensity of this data over time is shown in Fig. S17a along with an exponential fit. We observe that the effect of bleaching is not very strong in the time window considered. To calibrate the observation model for Bayesian inference, we use Markov chain Monte Carlo sampling so get a posterior distribution over the bleaching factor that can be used as a prior for the main experiments. We again start from (S9). In contrast to the GFP calibration, time course data is available. However, $b_1$ and $I$ are not distinguishable due to the lack of spot dynamics leading to te reduced model

$$I(t) = b_0 + I^* \exp(-\lambda t)\,.$$

The corresponding Bayesian posterior for the parameter $\lambda$ is shown in Fig. S17b leading to an estimate of $\lambda = 0.007\,\mathrm{s}^{-1}$. For the 3 s interval movies, we do not use separate calibration data. Instead, we observe that light exposure is the main factor determining the bleaching rate. A four-fold increase in imaging frequency should therefore correspond to a four-fold increase in the bleaching rate. As this calibration only serves to construct a prior distribution, this estimate is deemed sufficient.

### S9.3 Point Spread Function

We use a model-based approach to evaluate the spot intensity where the spot is described as a local background $b$ plus a point source with intensity at $(x_0, y_0)$. In the image space, this corresponds to an intensity profile of the form

$$I(x,y) = b + I_0 \exp\left(-\frac{1}{2\sigma_{\mathrm{PSF}}^2}(x - x_0^2) - \frac{1}{2\sigma_{\mathrm{PSF}}^2}(y - y_0^2)\right)\,. \tag{S13}$$

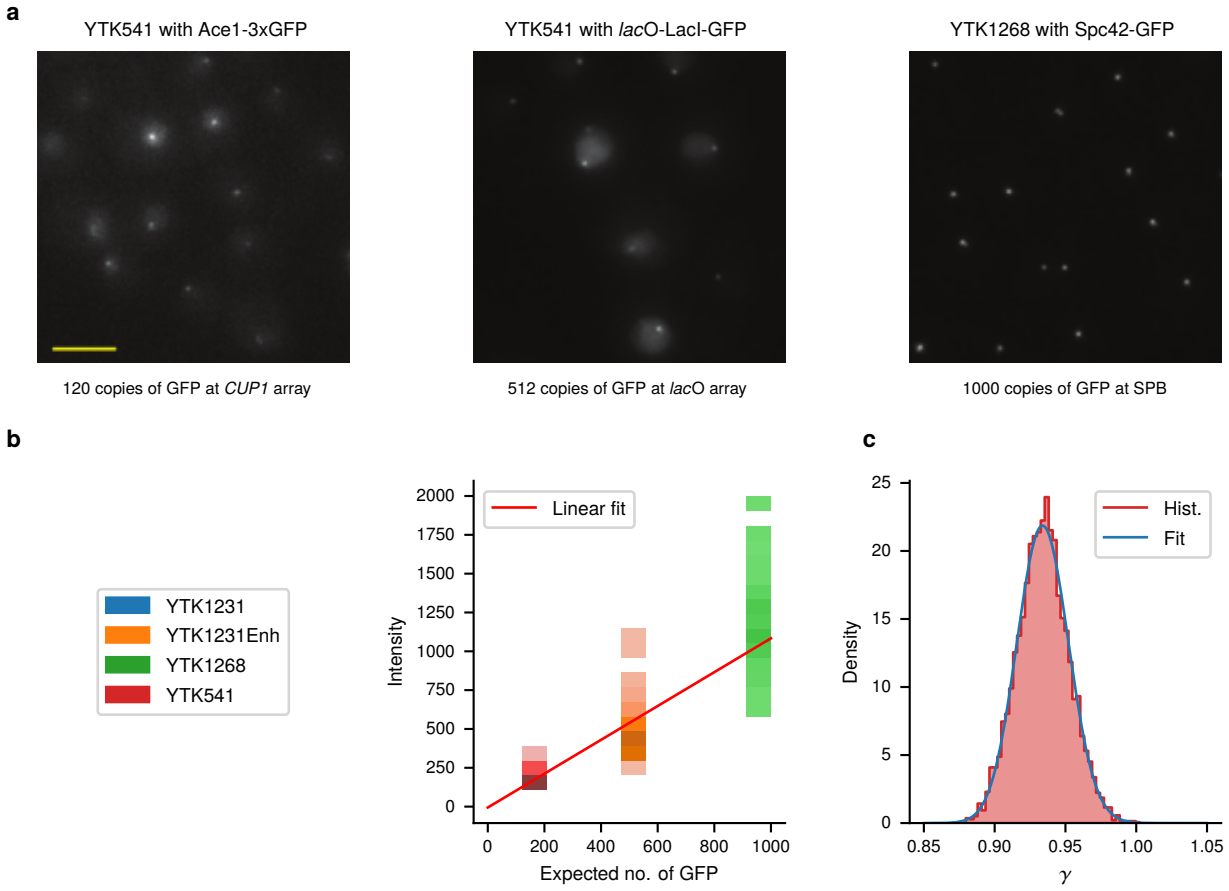


**Figure S16.** **a** Representative field views of cells with known numbers of GFP molecules per locus. **b** Color-coded histogram representations of the intensity distributions for the different strains. Darker colors indicate areas of higher density. The red line shows a least-squares fit of (S12) to the distribution. **c** Histogram of the posterior samples of $p(\gamma \mid I_1, \ldots, I_n)$ based on intensity estimates from $n = 509$ images (red). The blue curve corresponds to a gamma distribution fitted to the posterior samples.

The exponential term in (S13) corresponds to the point spread function with $\sigma_{PSF}$ and peak intensity $I_0$ describing the shape of the point in the image space. The parameter $\sigma_{PSF}$ depends on the properties of the optical system. While there are approximate formulas, these estimates are often not very accurate in systems with high aperture. We therefore infer $\sigma_{PSF}$ directly from a calibration dataset.

In fluorescence microscopy, the image is usually acquired by a CCD camera that discretizes the image space into a pixel grid. An example of a spot image from the strain *YTK1268* with spindle pole body is shown in Fig. S16a. The predicted intensity $p(i, j)$ of the pixel at position $i, j$ is given by

$$p(i,j) = \int_{x_i}^{x_{i+1}} \int_{y_i}^{y_{i+1}} I(x,y)\mathrm{d}x\mathrm{d}y$$

with pixel boundaries $x_i, x_{i+1}, y_i, y_{i+1}$. We assume now that the pixels have square shape and choose a coordinate system such that the pixel area is one. The intensity can then be expressed by

$$p(i,j) = b + \frac{\pi}{2}\sigma_{\mathrm{PSF}}^2 I_0 \left(\mathrm{erf}\left(\frac{x_{i+1}-x_0}{\sqrt{2}\sigma_{\mathrm{PSF}}}\right) - \mathrm{erf}\left(\frac{x_i - x_0}{\sqrt{2}\sigma_{\mathrm{PSF}}}\right)\right)\left(\mathrm{erf}\left(\frac{y_{i+1}-y_0}{\sqrt{2}\sigma_{\mathrm{PSF}}}\right) - \mathrm{erf}\left(\frac{y_i - y_0}{\sqrt{2}\sigma_{\mathrm{PSF}}}\right)\right).$$

Finally, in the measurement process, the intensity is corrupted by multiplicative noise. The measured intensity $\hat{I}(i, j)$ is thus given by

$$\hat{I}(i,j) = p(i,j)\exp(\sigma_{\mathrm{obs}}\epsilon_{ij}) \tag{S14}$$

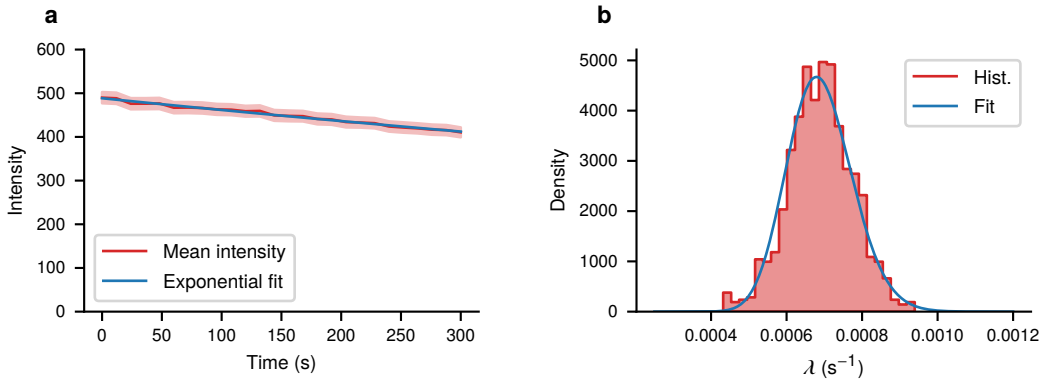


**Figure S17. a** Mean spot intensity of $n = 252$ cells of YTK1231 shown over time (red). The shaded region indicates the standard error. The blue line is a least squares fit of an exponential function. **b** Histogram of the posterior samples of $p(\lambda \mid I_1, \ldots, I_n)$ based on intensity estimates from $n = 252$ videos (red). The blue curve corresponds to a gamma distribution fitted to the posterior samples.

where $\sigma_{\text{obs}}$ is the noise level and $\epsilon_{ij}$ are i.i.d. standard normal random variables. By choosing priors for $b$, $I_0$, $x_0$, $y_0$ and $\sigma_{\text{PSF}}$, we have constructed a generative probabilistic model for spot images. We can then compute the posterior $p(\sigma_{\text{PSF}} \mid \hat{I})$ given an image $\hat{I}$. Generalization to multiple image samples is straightforward. A graphical model representation is shown in Fig. S18b. Using a non-informative prior, we run Hamiltonian Monte Carlo for inference. As shown in Fig. S18c, the resulting posterior is quite concentrated which justifies using a point estimate of $\sigma_{\text{PSF}} \approx 1.926$ during spot intensity estimation.

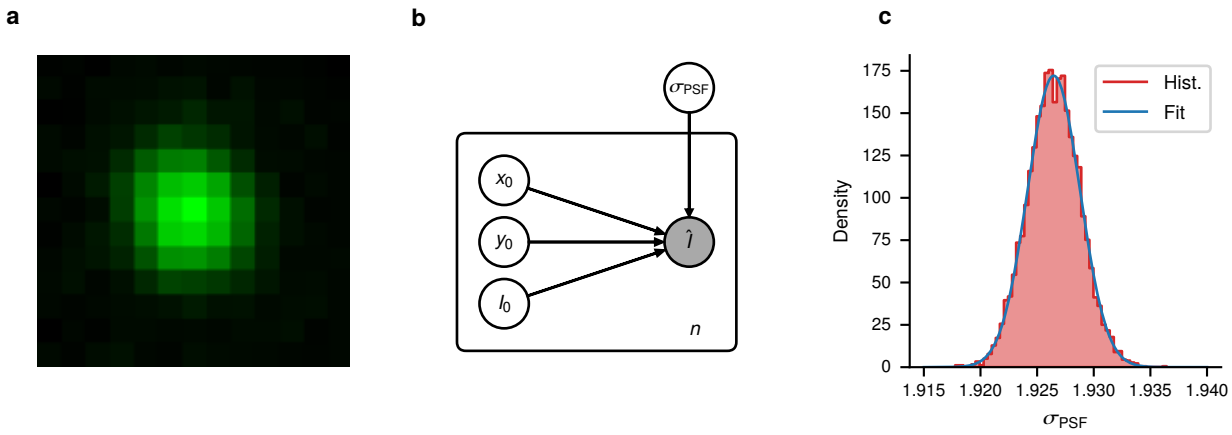


**Figure S18. a** Example of a spot image obtained from the strain YTK1268 with a spindle pole body. **b** Probabilistic graphical model for inferring the point spread function parameter $\sigma_{\text{PSF}}$ jointly from $n$ images. **c** Histogram of the posterior samples of $p(\gamma \mid \hat{I}_1, \ldots, \hat{I}_n)$ based on $n = 509$ images. The red line corresponds to a log-normal distribution fitted to the posterior samples.

## S10 Inference

### S10.1 Fully Bayesian inference

Given an observed trace $y = (y_1, \ldots, y_n)$ we are interested in estimating the latent state $x_{[0,T]}$, the model parameters $\theta$ and the observation parameters $\omega$. In a Bayesian approach, this corresponds to computing the joint posterior $p(\theta, \omega, x_{[0,T]} \mid y_1, \ldots, y_n)$. Targeting this posterior directly with Monte Carlo methods is difficult. Exploiting the relation

$$p(\theta, \omega, x_{[0,T]} \mid y_1, \ldots, y_n) = p(\theta, \omega \mid y_1, \ldots, y_n) p(x_{[0,T]} \mid \theta, \omega, y_1, \ldots, y_n)$$

full posterior can be split joint into two parts: The marginal parameter posterior

$$p(\theta, \omega \mid y) \propto p(\theta) p(\omega) p(y \mid \theta, \omega) \tag{S15}$$

and the conditional state posterior $p(x_{[0,T]} \mid \theta, \omega, y_1, \ldots, y_n)$. Intuitively, the conditional state posterior characterizes the distribution of paths of the underlying continuous-time process that have most likely created the observed trace $y$. One can generate samples from such a conditional Markov process using the backward filtering forward sampling approach. This is based on the observation that for a Markov jump process with generator $Q(x, x')$ the smoothing process is equivalent to a modified process with time-dependent generator[46]

$$\tilde{Q}(x, x', t) = \frac{\beta(x', t)}{\beta(x, t)} Q(x, x') \,. \tag{S16}$$

Here, $\beta(x, t) = p(y_k, \ldots, k_n \mid X(t) = x)$ with $k = \min\{i : t_i > t\}$ is the probability density of future observations given the current state. Now $\beta(x, t)$ satisfies a backward equation

$$\frac{\mathrm{d}}{\mathrm{d}_t}\beta(x, t) = -\sum_{x'} Q(x, x')\beta(x', t) \tag{S17}$$

with reset conditions

$$\beta(x, t_k^-) = \beta(x, t_k) p(y_k \mid X(t_k) = x), \quad k = 1, \ldots, n.$$

Sampling proceeds by solving (S17) backward in time and then using (S16) within a time-dependent version of the stochastic simulation algorithm[44]. The marginal parameter posterior (S15) is a continuous, finite-dimensional distribution and can, in principle, be tackled by a number of MCMC algorithms. A computational challenge, however, is that sampling from (S15) using MCMC requires evaluation of the marginal data likelihood $p(y \mid \theta, \omega)$. The marginal likelihood can be evaluated using filtering theory for Markov processes. Consider the filter distribution $\alpha(x, t) = \Pr(x(t) = x \mid y_1, \ldots, y_k)$, $k = \max\{i : t_i \leq t\}$ describing the best estimate of the current state of the system given all past observations. From general recursive filtering theory, we obtain that $\alpha(x, t)$ obeys the master equation in between the observation times, and at the observation times satisfies the update conditions

$$\begin{aligned} \alpha(x, t_k) &= C_k^{-1} p(x, t_k^-) p(y_k \mid x, \omega) \,, \\ C_k &= \sum_x \alpha(x, t_k^-) p(y_k \mid x, \omega) \,. \end{aligned}$$

Furthermore

$$\log p(y \mid \theta, \omega) = \sum_{k=1}^{n} \log C_k \,.$$

A derivation of the filter solution of the marginal likelihood is discussed in Sec. S10.2. To evaluate the gradient of the marginal likelihood we use the adjoin method and obtain

$$\frac{\mathrm{d}}{\mathrm{d}\theta_i} \log p(y \mid \theta, \omega) = \int_0^T \sum_{x, x'} \beta(x, t) R_i(x, x') p(x', t) \mathrm{d}t \tag{S18}$$

where $\beta(x, t)$ is a backward filter as in (S17) but with modified reset conditions

$$\begin{aligned} \beta(x, t_k^-) = \frac{p(y_k \mid x, \omega)}{C_k} \Big( & \beta(x, t_k) \\ & - \sum_{x'} \beta(x', t_k) p(x', t_k^+) p(y_k \mid x', \omega) + 1 \Big) \end{aligned}$$

and $R_i$ is the parameter independent part of the generator (cf. (S7)). A derivation of the modified backward filter is provided in Sec. S10.3. Having access to $\log p(y \mid \theta, \omega)$ and $\nabla_\theta \log p(y \mid \theta, \omega)$ allows to use efficient Monte Carlo methods such as Hamiltonian Monte Carlo[47]. In addition, the model can be integrated with probabilistic programming[48] facilitating easy reuse and modification of our approach.

In practice, we use the marginal approach for parameter inference. If latent state inference is required, the full posterior can be reconstructed by resampling $\theta, \omega$ from the parameter posterior and generate a trajectory from the smoothing process. These samples can be used to investigate arbitrary summary statistics of the latent process given the data, for example initiation times and local polymerase speed.

The discussion above refers to inference from a single trace. In case of a dataset $\mathcal{D} = \{y^{(1)}, \ldots, y^{(m)}\}$ of $m$ traces, a joint analysis with a hierarchical model is possible. To generalize the above to such a scenario, we view each trace $y^{(i)} = (y_1^{(i)}, \ldots, y_n^{(i)})$ as one single random variable (cf. main text, Fig 2e,f). As $p(y^{(i)} \mid \theta, \omega)$ is a finite dimensional distribution, constructing hierarchical models for a collection of traces follows the usual rules of probabilistic modeling and Bayesian inference[35]. Hierarchical modeling is discussed further in Sec. S10.4. The joint analysis requires evaluating forward and backward filters for every trace in the dataset in every step of MCMC. Therefore, full MCMC becomes infeasible for more then a few hundred traces. A viable alternative in this case is stochastic variational inference with mini-batching, that allows to efficiently compute the best approximation of the posterior within a parametric family of distributions[28].

### S10.2 Marginal likelihood by stochastic filtering

Formally, the marginal data likelihood is given by

$$p(y_{1,\ldots,n} \mid \theta, \omega) = \int \mathrm{d}x_{[0,T]} p(x_{[0,T]}, \theta) \prod_{k=1}^{n} p(y_k \mid X(t_k), \omega) \tag{S19}$$

where the integral is over all sample paths of stochastic process $X(t)$. Using the Markov property, the path integral (S19) reduces to the finite dimensional sum

$$\begin{aligned} p(y_{1,\ldots,n} \mid \theta, \omega) &= \sum_{x_1,\ldots,x_n} \Pr(X(t_1) = x_1, \ldots, X(t_n) = x_n) \prod_{k=1}^{n} p(y_k \mid X(t_k)) \\ &= \sum_{x_1} \Pr(X(t_1) = x_1) p(y_k \mid X(t_k)) \prod_{k=2}^{n} \sum_{x_k} \Pr(X(t_k) = x_k \mid X(t_{k-1}) = x_{k-1}) p(y_k \mid X(t_k)) \end{aligned}$$

which is analogous to the likelihood of a hidden Markov model[43]. The above expression can be computed recursively by stochastic filtering. To see this, we introduce the forward filter

$$\frac{\mathrm{d}}{\mathrm{d}t}\alpha(x,t) = \sum_{x'} Q(x',x) p(x',t) \tag{S20}$$

$$\alpha(x, t_k) = C_k^{-1} \alpha(x, t_k^-) p(y_k \mid x, \omega) \tag{S21}$$

$$C_k = \sum_{x} \alpha(x, t_k^-) p(y_k \mid x, \omega) \tag{S22}$$

For the normalization constant $C_k$ we have

$$C_k = \sum_{x_k} p(y_k \mid x_k) \alpha(x_k \mid y_1, \ldots, y_{k-1}, \theta) = p(y_k \mid y_1, \ldots, y_{n-1})$$

from which follows

$$p(y \mid \theta) = \prod_{k=1}^{n} C_k \,.$$

### S10.3 Gradient of the marginal likelihood

To calculate the gradient $\nabla_\theta p(y \mid \theta, \omega)$, we observe that the likelihood depends on the parameters only implicitly through the prediction step in (S20). More specifically, we require the derivative of

$$\begin{aligned} \log p(y_{1,\ldots,n} \mid \theta, \omega) &= \sum_{k=1}^{n} \log C_k(\theta) \\ \text{subject to} \qquad \dot{\alpha}(x,t) &= \sum_i \theta_i \sum_{x'} R_i(x',x) \alpha(x',t) \,, \\ p(x, t_k) &= \frac{1}{C_k} p(x, t_k^-) p(y_k \mid x) \,, \\ C_k &= \sum_x \alpha(x, t_k^-) p(y_k \mid x) \,, \end{aligned}$$

where we have used the parameter form of the transition function (S7) in the master equation within the constraint. To compute the gradient of such a constrained function, one can use variational calculus. First, the constrained problem is transformed to an augmented unconstrained functional known as the Lagrangian

$$\begin{aligned} J[\alpha,\theta,\beta,\eta] = &\sum_{k=1}^{n} \log(C_k) + \sum_x \int_0^T \beta(t,x) \left( \sum_i \theta_i \sum_{x'} R_i(x',x)\alpha(x',t) - \dot{\alpha}(x,t) \right) \mathrm{d}t \\ &+ \sum_{k=1}^{n} \sum_x \eta_k(x) \left( \frac{1}{C_k} \alpha(x,t_k^-) p(y_k \mid x) - \alpha(x,t_k) \right) . \end{aligned} \tag{S23}$$

The gradient of the original function can be computed from the stationary conditions of the Lagrangian. In particular

$$\frac{\mathrm{d}}{\mathrm{d}\theta_i} \log p(y \mid \theta) = \frac{\partial}{\partial \theta_i} J[\alpha,\theta,\beta,\eta] = \sum_{x,x'} \int_0^T \beta^*(t,x) R_i(x',x) p^*(x',t) \mathrm{d}t$$

where $\alpha$ is the solution of (S20) for given $\theta$ and $\beta$, $\eta$ satisfy the stationarity condition

$$\frac{\delta J}{\delta \alpha} = 0 \,. \tag{S24}$$

To calculate the functional derivative, we choose a suitable perturbation $\delta\alpha$ and linearize (S23) around $\alpha$. After reorganizing terms, we get

$$\begin{aligned} \delta J = &\sum_x \delta\alpha(x,t) \left( \dot{\beta}(x,t) + \sum_{x'} \sum_i \theta_i R_i(x,x') \right) \\ &+ \sum_k \sum_x \delta\alpha(x,t_k^-) \left( \frac{p(y_k \mid x)}{C_k} + \frac{\eta_k(x) p(y_k \mid x)}{C_k} - \frac{p(y_k \mid x)}{C_k^2} \sum_{x'} \eta_k(x') \alpha(x',t_k^-) p(y_k \mid x') - \beta(x,t_k^-) \right) \\ &+ \sum_x \delta\alpha(x,t_k^-) \left( \beta(x,t_k) - \eta_k(x) \right) . \end{aligned}$$

Since $\delta\alpha$ is arbitrary, the terms in brackets must vanish to satisfy (S24). This leads to the adjoint state equation

$$\dot{\beta}(x,t) = -\sum_{x'} \sum_i \theta_i R_i(x,x') \beta(x,x')$$

along with the jump conditions

$$\beta(x,t_{k1}^-) = \frac{p(y_k \mid x)}{C_k} \left( \beta(x,t_k) - \sum_{x'} \beta(x',t_k) p(x',t_k^-) p(y_k \mid x') + 1 \right) .$$

For applying HMC, we also require the gradient $\nabla_\omega p(y \mid \theta, \omega)$. This gradient is straightforward to compute since $\omega$ only affects the observation likelihood

$$\begin{aligned} \nabla_\omega p(y \mid \theta, \omega) &= \sum_k \nabla_\omega \log C_k(\omega) \\ &= \sum_k C_k^{-1}(\omega) \nabla_\omega C_k(\omega) \\ &= \sum_k C_k^{-1}(\omega) \sum_x \alpha(x,t_k^-) \nabla_\omega p(y_k \mid x, \omega) \end{aligned}$$

which corresponds to a weighted average of the gradients of the observation likelihood.

### S10.4 Hierarchical modelling

Consider a collection of traces $\mathcal{D} = \{y^{(1)}_{1,\dots,n}, \dots, y^{(m)}_{1,\dots,n}\}$. As discussed in *Methods — Bayesian inference*, we can consider the discretely observed trace $y_{1,\dots,n}$ as a single random variable generated from a distribution $p(y_{1,\dots,n} \mid \theta, \omega)$. This allows to apply the usual rules for probabilistic graphical models and Bayesian inference to construct hierarchical models. For convenience, we provide a few more explicit examples in this section.

**Pooling of multiple traces** The simples model for a collection of traces $\mathcal{D}$ is that of identical and independently distributed samples. This means that two parameter vectors $\theta$, $\omega$ are shared by all the traces. We are interested in the posterior

$$p(\theta, \omega \mid \mathcal{D}) \propto p(\theta)p(\omega) \prod_{k=1}^{m} p(y^{(k)}_{1,\dots,n} \mid \theta, \omega)$$

where $p(\theta)$ and $p(\omega)$ are appropriate priors for transcription and observation parameters, respectively, and each term in the product is a marginal likelihood for a single trajectory as in (S19). This model uses identical parameters for all time windows meaning that it allows no cycle dependence and is considered as a baseline. More complex models will be scored against the fully pooled model.

**Per-window pooling** Consider now traces $\mathcal{D} = \left\{ y^{(i,j)}_{1,\dots,n} : i \in \{1, \dots, n_w\}, j \in \{1, \dots, n_{y,i}\} \right\}$ where $n_w$ is the number of windows and $m_i$ is the number of traces in window $i$. For every window, we assume independent parameters $\theta^{(i)}$, $\omega^{(i)}$. In this case, the inference problem decomposes into individual posteriors for every window

$$p(\theta^{(i)}, \omega^{(i)} \mid \mathcal{D}) \propto p(\theta^{(i)})p(\omega^{(i)}) \prod_{k=1}^{m_i} p(y^{(i,k)}_{1,\dots,n} \mid \theta^{(i)}, \omega^{(i)}) \,.$$

This corresponds to repeating the simple pooling independently for each window. While straightforward to apply, this approach does not exploit shared information between traces of different windows, which can lead to poor accuracy if the number of traces per window is small.

**Mixed approach** In practice, some of the parameters may depend on the cycle while others are shared between all traces. Throughout, we assume a global observation parameter $\omega$. The kinetic transcription parameters $\theta$ are split into global parameters $\theta_g$ and local parameters $\theta_l^{(i)}$ that are allowed to vary for different windows. The corresponding posterior is of the form

$$p(\theta_g, \omega, \theta_l^{(i)}, \dots, \theta_l^{(n_w)} \mid \mathcal{D}) \propto p(\theta_g)p(\omega) \prod_{i=1}^{n_w} p(\theta_l^{(i)}) \prod_{k=1}^{m_i} p(y^{(i,k)}_{1,\dots,n} \mid \theta_g, \theta_l^{(i)}, \omega) \,.$$

By performing inference with different partitions into local and global parameters and comparing the different models, this provides some insight into which parameters are affected by the cycle.

### S10.5 Parameter identifiability

The presented model contains 5 kinetic parameters and additional 7 observation parameters. In addition, the latent stochastic process $X(t)$ is a fairly high-dimensional lattice model ($\approx 2 \times 10^5$ states in the configuration used for most experiments) that is observed by fluorescence intensity, which is a one-dimensional quantity. This setting raises the question of parameter identifiability, which we investigated based on simulated data. Preliminary numerical experiments indicated that in particular the GFP intensity calibration factor $\gamma$ has a major impact on inference quality. If an uninformative prior is used for $\gamma$, the system is not practically identifiable with a realistic number of traces. We therefore used calibration measurements to construct a tight prior for $\gamma$ (see Sec. S9.1). However, even with good prior knowledge of $\gamma$, a single trace is not very informative. Pooling of multiple traces can increase accuracy significantly. This was demonstrated in the main paper using the example of the initiation rate (cf. Fig. 1f). The full results with posteriors for all parameters are given in Fig. S19.

### S10.6 Computing infrastructure

Numerical experiments using the real data were run on the Hessian High Performance Computer (HHLR) located at TU Darmstadt. A typical run of the variational inference was performed on a single compute node consisting of 96 Intel Xeon Platinum 9242 processors for 24 hours which allowed for roughly 2000 gradient steps. The experiments based on simulated used fewer traces and were thus run on the local cluster of Self-Organizing Systems Lab on a 22 core machine with Intel Core Haswell architecture.

### S10.7 Custom Software

The core of the code consists of the following parts: A model builder that allows to straightforwardly implement arbitrary CTMC models, a simulation module for data generation, an integrator for solving the master equation based on the Krylov subspace approximation of the matrix exponential[41,42], programs for evaluating the data likelihood and the corresponding gradient with respect to the parameters and a simulator for the state posterior. The backend of the code is implemented in C++ and was compiled into a native Python extension using the Pybind11. Matrix operations are implemented using the Eigen library. The code has been parallelized using OpenMP such that it can process multiple traces simultaneously. For inference, our python extension has also been interfaced with the PyTorch package and the probabilistic programming language Pyro[45]. This gives access to state of the art inference algorithms and simplifies implementing hierarchical models. The code is available at XXX.

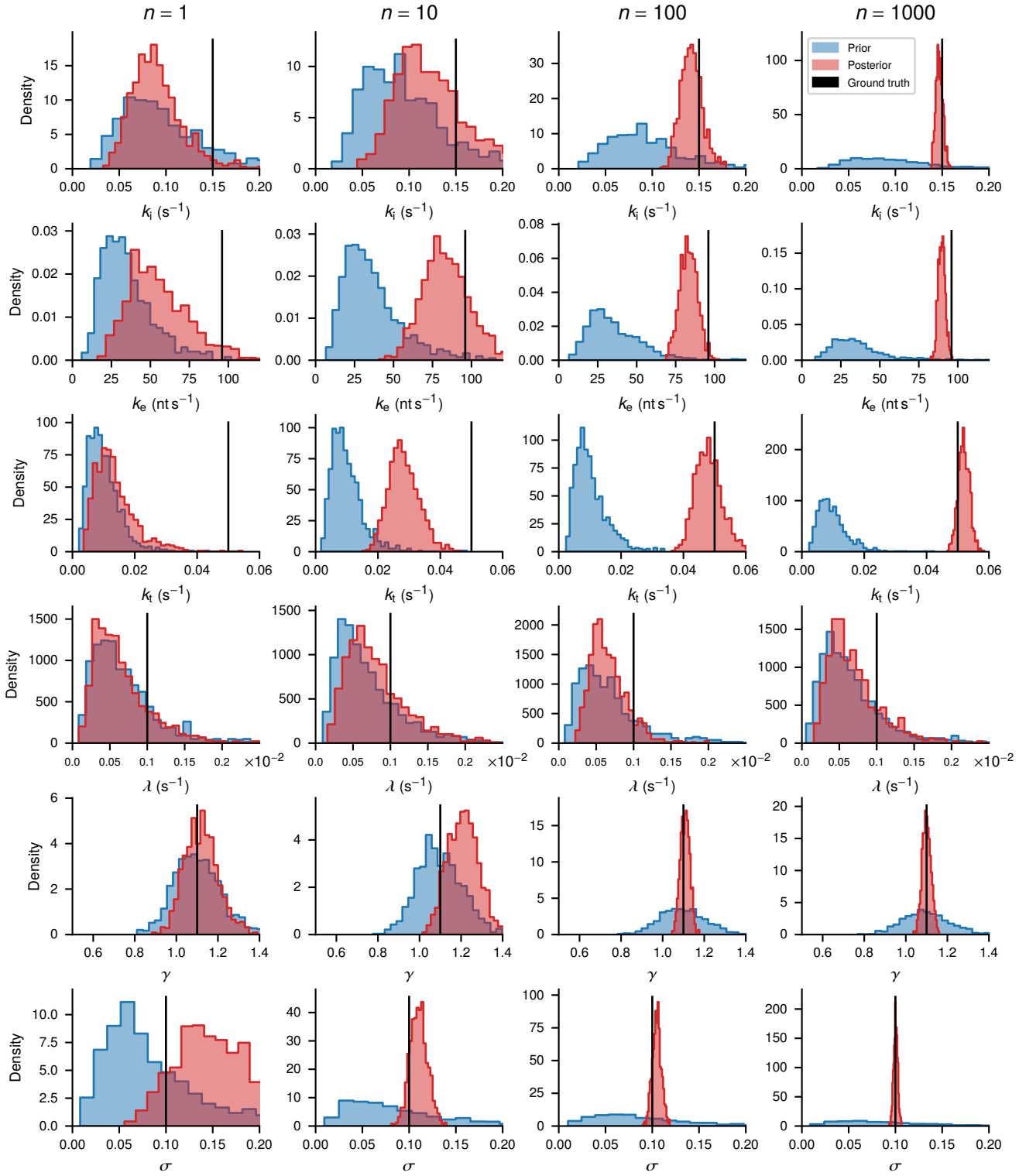


**Figure S19.** Joint posterior inference with different numbers of pooled traces (cf. Sec. S10.4) based on simulated data. The rows show histogram approximations of the prior distribution (blue) and the posterior distribution (red) for the model parameters. Black lines indicate the parameter value used to generate the data. The columns show realizations of the experiment with different numbers of trajectories. Some of the observation parameters are not shown, as results look similar.

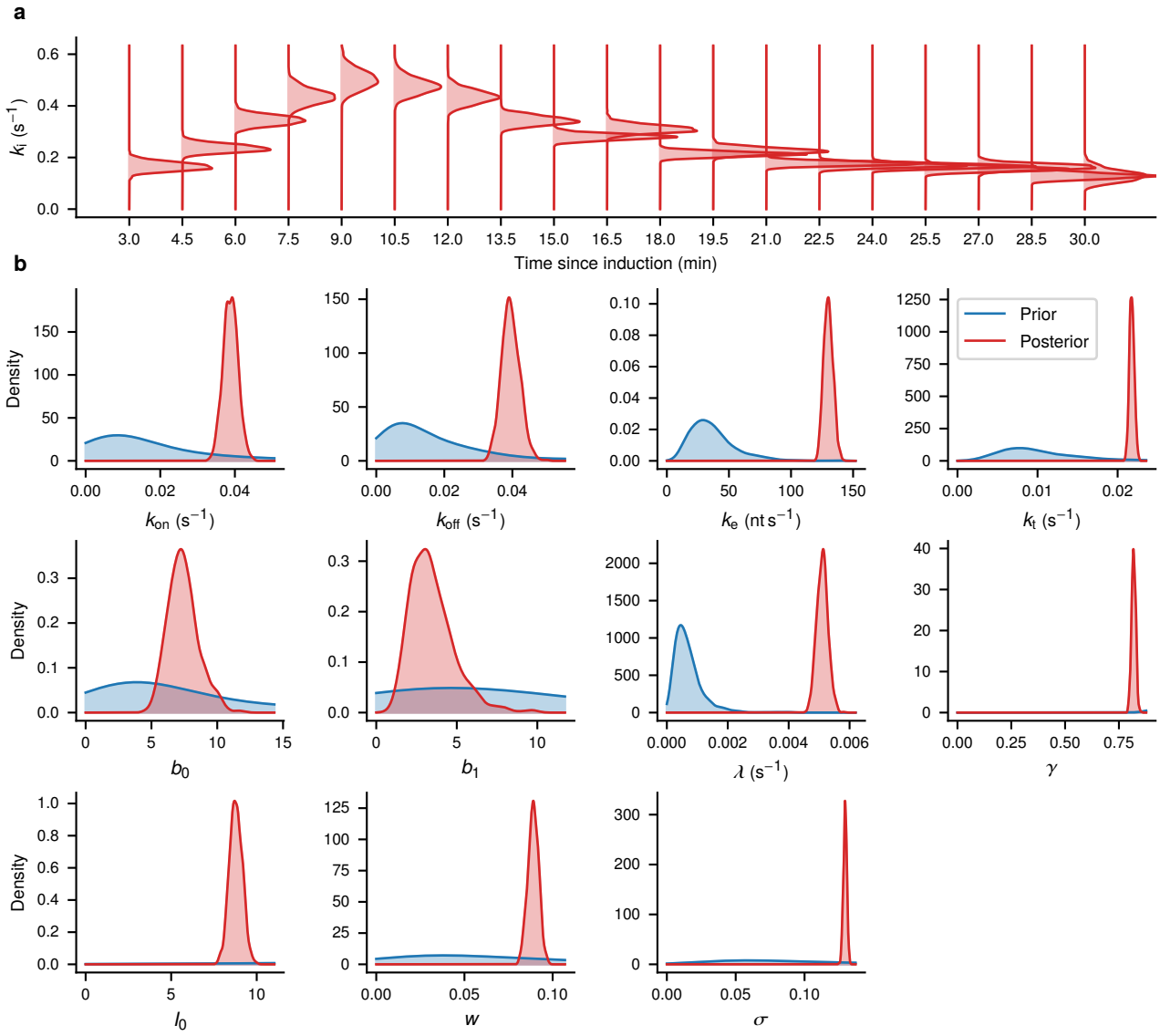


**Figure S20.** Posterior of the model with local initiation rate and remaining parameters global. This is an extended version of Fig. 4e,f of the main paper. **a** Local initiation rate per time window since induction. **b** Global kinetic and observation model parameters.

## S11 Model selection

In Bayesian inference there are two major approaches to model evaluation and comparison[35]. The first approach is based on the idea that we can perform posterior inference not only over parameters but also over models. Assuming there are a number of hypothesis $H_1, \ldots, H_m$ corresponding to $m$ models, some prior probabilities over the models $p(H)$ and data $\mathcal{D}$, we can compute $p(H \mid \mathcal{D})$ to find the most probable model given the data. Typically, a uniform prior over the models is chosen which leads to $p(H \mid \mathcal{D}) \propto p(\mathcal{D}|H)$. In particular, if we are interested in comparing the odds of two models, we get

$$\frac{p(H_1 \mid \mathcal{D})}{p(H_2 \mid \mathcal{D})} = \frac{p(\mathcal{D}|H_1)}{p(\mathcal{D}|H_2)} .$$

The fraction on the r.h.s. is called the Bayes factor, the term $p(\mathcal{D}|H)$ is called the marginal likelihood or evidence. In a parametric model the marginal likelihood is given by

$$p(\mathcal{D}|H) = \int p(\mathcal{D} \mid \theta, H) p(\theta \mid H) \mathrm{d}\theta .$$

Therefore, the marginal likelihood can be used to score the performance of different models for a given dataset. The main advantage of the marginal likelihood as a model evaluation criterion is that it automatically penalizes model complexity. If two models explain data equally well, the one with fewer parameters will show a higher score. Unfortunately, the marginal likelihood is difficult and costly to evaluate, so typically approximations have to be used. A comprehensive review of such approximation techniques can be found in[?]. If inference is performed using a variational approach, an approximation to the marginal likelihood is obtained for free in form of the evidence lower bounds (ELBO). Classical variational inference is based on the idea to approximate the posterior by a distribution $q$ in a tractable family $\boldsymbol{Q}$. The optimal approximation $q^*$ is obtained by minimizing the Kullback-Leibler divergence

$$q^* = \arg\min_{q \in \boldsymbol{Q}} D_{\mathrm{KL}}[q \,||\, p(\cdot \mid \mathcal{D}, H)] , \quad D_{\mathrm{KL}}[q \,||\, p] = \int q(\theta) \log\left(\frac{q(\theta)}{p(\theta)}\right) \mathrm{d}\theta$$

It can be shown that the above objective function decomposes as

$$D_{\mathrm{KL}}[q \,||\, p(\cdot \mid \mathcal{D})] = \log p(\mathcal{D} \mid H) - F(\mathcal{D}, H) , \quad F(\mathcal{D}, H) = \int q(\theta) \left(\log p(\theta \mid H) + \log p(D \mid \theta, H) - \log q(\theta)\right) \mathrm{d}\theta$$

The quantity $F(\mathcal{D}, H)$ is the ELBO and the maximization of the ELBO is equivalent to minimizing the KL divergence to the posterior. Furthermore, if the variational family $Q$ is sufficiently expressive, $D_{\mathrm{KL}}[q^* \,||\, p(\cdot \mid \mathcal{D})]$ is close to zero and we get

$$\log p(\theta \mid H) \approx F(\mathcal{D}, H) .$$

This immediately suggest an approximation of the Bayes factor as

$$\log\left(\frac{p(\mathcal{D}|H_1)}{p(\mathcal{D}|H_2)}\right) \approx F(\mathcal{D}, H_1) - F(\mathcal{D}, H_2) \equiv \ \Delta\mathrm{ELBO}$$

A conceptual drawback of evidence-based model selection is that it is only a relative measure of performance. If all models are bad and one model is slightly less bad, the latter one can still appear as a clear winner by evidence score. It is therefore advisable to combine evidence-based model selection with goodness of fit checks. In a Bayesian framework, the natural way to do this is by the posterior predictive distribution

$$p(\tilde{\mathcal{D}} \mid \mathcal{D}, H) = \int p(\tilde{\mathcal{D}} \mid \theta, H) p(\theta \mid \mathcal{D}, H) \mathrm{d}\theta . \tag{S25}$$

Intuitively, $p(\tilde{\mathcal{D}} \mid \mathcal{D}, H)$ is the distribution of data simulated from the fitted model. Goodness of fit in this framework is assessed by comparing the similarity of certain summary statistics $T(\mathcal{D})$ of the true data with the same summary statistics of the posterior predictive distribution. More explicitly, for a discrepancy measure $D(T(\mathcal{D}), T(\tilde{\mathcal{D}}))$ for two realization of the summary statistic, the predictive score is given by

$$S = \int D(T(\mathcal{D}), T(\tilde{\mathcal{D}})) p(\tilde{\mathcal{D}} \mid \mathcal{D}, H) \mathrm{d}\tilde{\mathcal{D}}$$

In practice, the integral with respect to the predictive distribution is performed by sampling from the parameter posterior $p(\theta \mid \mathcal{D}, H)$ and simulating a dataset $\tilde{\mathcal{D}}$ using the model $p(\tilde{\mathcal{D}} \mid \theta, H)$. The choice of these summary statistics depends on the problem and should reflect features that are deemed important. In many applications, low order moments such as the mean or quantiles of the empirical distribution are chosen and the discrepancy measure $D$ is, e.g. the $L_2$-norm. Here, we are interested in the evolution of the intensity distribution over time. Consider a collection of traces $\mathcal{D} = \{y^{(1)}_{1,\dots,n}, \dots, y^{(m)}_{1,\dots,n}\}$. Let $p_k(I, \mathcal{D}) = \frac{1}{m}\sum_{i=1}^{m} \delta(y_k^{(i)})$ be the empirical intensity distribution at time $t_k$. In this case, we are interested in the distribution of fluorescence intensity over time. As summary statistic we choose the collection of empirical distributions at the different time points, i.e. $T = (p_k)_{k=1^n}$. The same collection of distributions is computed for a simulated dataset $\tilde{\mathcal{D}}$. To compare individual distributions, we use the Wasserstein metric $W_1(p_i, p_i')$. The Wasserstein metric is rooted in optimal transport theory. Intuitively, it describes the minimal amount of work required to transform one distribution into another one by moving infinitesimal pieces of mass at a time[50]. Averaging over individual time points leads to

$$S = \int \left( \frac{1}{m} \sum_{i=1}^{m} W_1(p_i(\cdot, \mathcal{D}), p_i(\cdot, \tilde{\mathcal{D}})) \right) p(\tilde{\mathcal{D}} \mid \mathcal{D}, H) \mathrm{d}\tilde{\mathcal{D}}$$

This statistic is computed separately for all windows and then averages over the windows. We do not provide the explicit formula is the notation becomes quite cumbersome. Finally, after computing $S$ for all models of interest we can rank the models by how well the average predicted intensity distribution agrees with the empirical distribution.

Table S15 shows ΔELBO and the discussed posterior predictive score for various model configurations evaluated on both datasets. The relative ELBO is always provided for the simplest model, which can be considered as a baseline, for every dataset. A look at ΔELBO reveals that in both datasets there is a significant preference of the bursting models over the constitutive models. In addition, the time-dependent models score higher than the fully global model. Most of this gain is explained by a time-varying initiation rate. If other dynamic parameters are allows to vary, there is an additional but smaller gain in evidence. Interestingly, the fully local model shows a predictive performance close to other well fitting models but with lower evidence. This indicates that the additional degrees of freedom are not very useful for explaining the experimental data. In conclusion, S15 supports a bursting model with time-dependent parameters with most of the time dependence explained by the initiation rate.

| $k_{on}$ | $k_{off}$ | $k_i$ | $k_e$ | $k_t$ | 3s dataset ΔELBO | 3s dataset $S$ | 12s dataset ΔELBO | 12s dataset $S$ |
|---|---|---|---|---|---|---|---|---|
| - | - | g | g | g | 0 | - | 0 | - |
| - | - | l | g | g | $9.78 \times 10^2$ | 21.968281 | $1.55 \times 10^3$ | 22.716701 |
| - | - | l | l | l | $1.29 \times 10^3$ | 20.916601 | $1.73 \times 10^3$ | 21.943819 |
| g | g | g | g | g | $2.03 \times 10^3$ | - | $5.25 \times 10^3$ | - |
| l | g | g | g | g | $1.95 \times 10^3$ | 19.038088 | $6.06 \times 10^3$ | 22.791160 |
| g | g | g | l | g | $1.64 \times 10^3$ | 26.032090 | $5.43 \times 10^3$ | 45.164126 |
| g | g | l | g | g | $2.93 \times 10^3$ | 14.482935 | $6.07 \times 10^3$ | 22.861368 |
| l | l | g | g | g | $2.39 \times 10^3$ | 16.207827 | $4.98 \times 10^3$ | 19.721966 |
| l | l | l | g | g | $2.37 \times 10^3$ | 16.148590 | $6.13 \times 10^3$ | 20.363244 |
| l | l | l | l | l | $2.62 \times 10^3$ | 15.607396 | $6.17 \times 10^3$ | 20.185700 |

**Table S15.** Bayesian model selection by approximate evidence and predictive scores based on Wasserstein metric. The index column on the right indicate which parameters are shared for all cells (g) and which are local for every time window (l). The dash indicated missing values, meaning that the corresponding model is not bursting. The quantity ΔELBO is shown with respect to the simplest model (shared parameter, no bursting) for every dataset. Therefore, higher values are better. $S$ quantifies the average deviation of the intensity distribution between simulated and real dataset. Thus, lower values are better.

## S12 Tracking algorithm

Transcription site tracking in live cell fluorescence microscope imagery has to overcome three problems. First, fluorescent spots vary in intensity and may disappear altogether due to the underlying dynamics of the transcription process. Second, the background is non-homogenous and varies over time due to cellular clutter. Third, accumulations of fluorescent protein may cause spurious spots. Standard spot extraction algorithms typically run a detection step on each time frame and than combine possible spot candidates to a trajectory using a scoring scheme. We follow a joint approach combining detection and tracking within a stochastic filtering framework.

### S12.1 Sequential Filtering Framework

The central idea is that the current position of the spot and the current intensity will provide information on the likely places to find the spot in the next frame. We denote the position of the spot in frame $k$ as $r_k = (x_k, y_k, z_k)^\top$. In addition, we define $I_k$ and $b_k$ as the spot and local background intensity, respectively. This leads to a full state $\mathbf{x}_k = (r_k, b_k, I_k)$. We assume the evolution of these quantities between two consecutive images as given by

$$p(\mathbf{x}_k \mid \mathbf{x}_{k-1}) = \mathcal{N}(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \mathbf{Q}) \,. \tag{S26}$$

Eq. (S26) corresponds to a diffusive motion of the TS and and encourages the intensity variables at different times to be close. To deal with sudden vanishing and re-appearance of spots, we introduce a binary switching variable $s_k$ representing the visibility. The observation $\hat{I}_k$ corresponds to the 3D image stack at step $k$. We connect the latent state $\mathbf{x}_k$ to the observed image $\hat{I}_k$ by a point spread function model as described in Sec. S9.3

$$h^i(\mathbf{x}_k) = b_k + I_k \exp\left(-\frac{1}{2}(r^i - r_k)^\top \Sigma_{\mathrm{PSF}}^{-1}(r^i - r_k)\right) , \tag{S27}$$

where $r^i$ corresponds to the center of pixel $i$. In contrast to Sec. S9.3, we use a three-dimensional PSF model with $\Sigma_{\mathrm{PSF}} = \mathrm{diag}(\sigma_{\mathrm{PSF}}^2, \sigma_{\mathrm{PSF}}^2, \sigma_z^2)$. We also refrain from integrating over the pixel area and use the center of the pixel directly with the Gaussian profile. Together with a multiplicative noise as in (S14) we obtain

$$p(\log \hat{I}_k \mid s_k = 1, \mathbf{x}_k) = \mathcal{N}(\log h(\mathbf{x}_k), \sigma_{\mathrm{img}}^2)$$
$$p(\log \hat{I}_k \mid s_k = 1, \mathbf{x}_k) = \mathcal{N}(\log b_k, \sigma_{\mathrm{img}}^2)$$

where the second equation indicates that only noisy background is observed when visibility is zero. This leads to a hidden Markov model with spot position and intensity as latent state and image stacks as observations (cf. Fig. S21). Computing

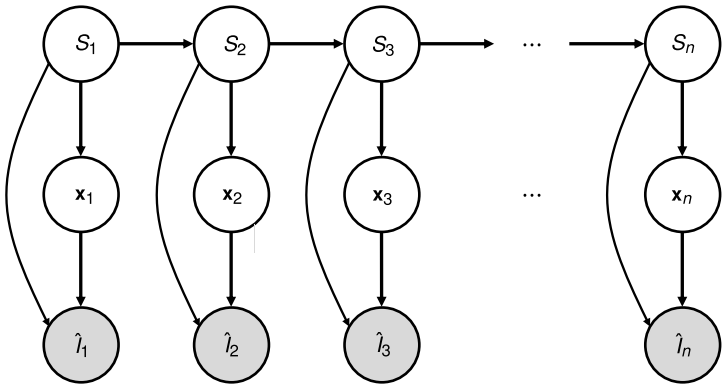


**Figure S21.** Probabilistic graphical model representation of the spot tracking problem with visibility variables $s_k$, spot state $\mathbf{x}_k$ and noise observed image stack $I_k$.

$p(s_k, \mathbf{x}_k \mid \hat{I}_1, \ldots, \hat{I}_k)$, the distribution of the latent state given all observations up to step $k$, is known as the filtering problem and can be reduced to sequential prediction and update steps[43].

### S12.2 Prediction Step

Assume that at step $k-1$ we have access to the filtering distribution

$$p(s_{k-1}, \mathbf{x}_{k-1} \mid \hat{I}_1, \ldots, \hat{I}_{k-1}) = p(s_{k-1} \mid \hat{I}_1, \ldots, \hat{I}_{k-1}) p(\mathbf{x}_{k-1} \mid s_{k-1}, \hat{I}_1, \ldots, \hat{I}_{k-1})$$

where $p(s_{k-1} \mid \hat{I}_1, \ldots, \hat{I}_{k-1})$ is a Bernoulli distribution and we assume that $p(\mathbf{x}_{k-1} \mid s_{k-1}, \hat{I}_1, \ldots, \hat{I}_{k-1})$ follows a Gaussian distribution. In this case, the prediction step can be computed explicitly as

$$p(s_k, \mathbf{x}_k \mid \hat{I}_1, \ldots, \hat{I}_{k-1}) = \sum_{s_{k-1}} \underbrace{p(s_k \mid s_{k-1}) p(s_{k-1} \mid \hat{I}_1, \ldots, \hat{I}_{k-1})}_{\text{weight}} \underbrace{p(\mathbf{x}_k \mid s_{k-1}, \hat{I}_1, \ldots, \hat{I}_{k-1})}_{\text{Gaussian mode}} \,. \tag{S28}$$

The last term on the r.h.s. can be computed analytically under the Gaussian assumption. Thus, (S28) becomes a Gaussian mixture with modes at the previous filter estimates for visible and invisible state and weights determined by the activity estimate.

### S12.3 Approximate Update Step

Given the prediction step, the update step is given by

$$p(s_k, \mathbf{x}_k \mid \hat{I}_1, \ldots, \hat{I}_k) = \frac{1}{p(\mathbf{y}_k \mid \hat{I}_1, \ldots, \hat{I}_{k-1})} p(\hat{I}_k \mid s_k, x_k) p(s_k, \mathbf{x}_k \mid \hat{I}_1, \ldots, \hat{I}_{k-1}) \,.$$

The normalizer is given by

$$\begin{aligned} p(\mathbf{y}_k \mid \hat{I}_1, \ldots, \hat{I}_{k-1}) &= \sum_{s_k} \int \mathrm{d}\mathbf{x}_k \, p(\hat{I}_k \mid s_k, x_k) p(s_k, \mathbf{x}_k \mid \hat{I}_1, \ldots, \hat{I}_{k-1}) \\ &= \sum_{s_k, s_{k-1}} p(s_k \mid s_{k-1}) p(s_{k-1} \mid \hat{I}_1, \ldots, \hat{I}_{k-1}) \int \mathrm{d}\mathbf{x}_k \, p(\hat{I}_k \mid s_k, x_k) p(\mathbf{x}_k \mid s_{k-1}, \hat{I}_1, \ldots, \hat{I}_{k-1}) \,. \end{aligned}$$

For the model discussed here, the update step cannot be solved in closed form due to the non-linear observation model (S27). In addition, the binary state $s_k$ causes the filtering distribution to become a mixture with $2^k$ components. To obtain a tractable approximation, we combine two approximations. First, we observe that the non-linear observation likelihood is strongly peaked. Thus it is reasonable to use a Laplace approximation

$$\begin{aligned} p(\hat{I}_k \mid s_k, x_k) p(\mathbf{x}_k \mid s_{k-1}, \hat{I}_1, \ldots, \hat{I}_{k-1}) &= \exp(-f(x)) \\ &\approx \exp(-f(x_0)) \exp\left( \frac{1}{2} (x - x^*)^\top H_f(x^*) (x - x^*) \right) \end{aligned}$$

where

$$f(x) = -\log p(\hat{I}_k \mid s_k, x_k) - \log p(\mathbf{x}_k \mid s_{k-1}, \hat{I}_1, \ldots, \hat{I}_{k-1})$$

and $x^* = \arg\min_x f(x)$. This leads to a representation of the filter in form of a Gaussian mixture distribution at every time. In order to keep the mixture from growing, we perform a mixture reduction via moment matching after every step.