Supplementary Information
Automating Bayesian inference and design to quantify acoustic particle levitation

Kiran Dhatt-Gauthier, Dimitri Livitz, and Kyle J. M. Bishop[*]

Department of Chemical Engineering, Columbia University, New York, NY, USA

Contents

---

[*]Address correspondence to kyle.bishop@columbia.edu

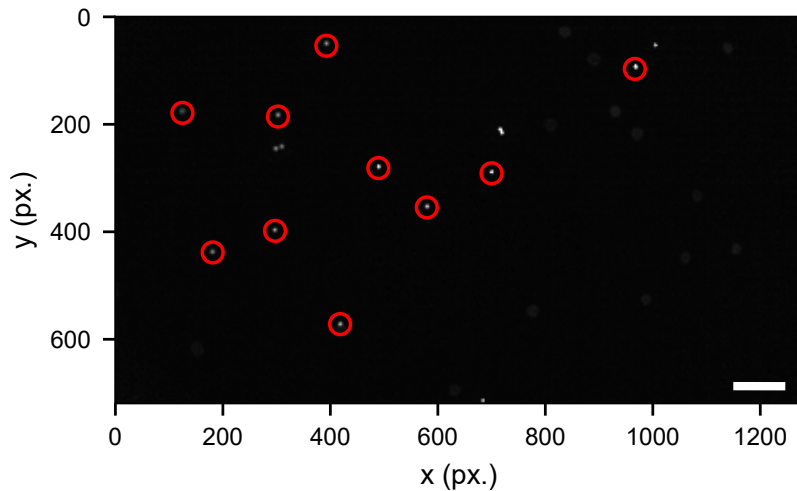# 1 Observation

1.1 Particle imaging and tracking

Figure S1: Optical micrograph of PS particles levitating at the nodal plane from the first optimal experiment in the sequence shown in Figures 5 and 6. Scale bar is 250 μm (1 pixel = 2.14 μm). The particles appear somewhat blurry because the focal plane of the microscope is positioned ca. 15 μm above the height of the particles. This choice ensures that the apparent particle size y varies linearly with their height h (see Figure S2 below). Some particles aggregate at the nodal plane due to lateral radiation forces caused by heterogeneity in the acoustic field. Only isolated particles (colored circles) are considered in our analysis.

1.2 Linear relation between particle height h and size y

To quantify the relationship between the particle size y returned by TrackPy and the particle height h within the acoustic cell, we measured the apparent size of a particle as a function of time during its steady sedimentation through the focal plane of the microscope. Figure S2 shows the result of this analysis for 4 different particles. During sedimentation, the height of the particle varies linearly with time as $h = h_0 - Ut$ where $U = 6.1$ μm/s is the estimated speed of particle sedimentation. Using this estimate, Figure S2 shows the particle size as a function of height h measured relative to the focal plane h_{fp} where the apparent size is minimal. During our experiments, we position the focal plane ca. 15 μm above the nodal plane of the acoustic field. Consequently, all measurements are conducted within the shaded region of Figure S2 where the relationship between size y and height h is approximately linear.
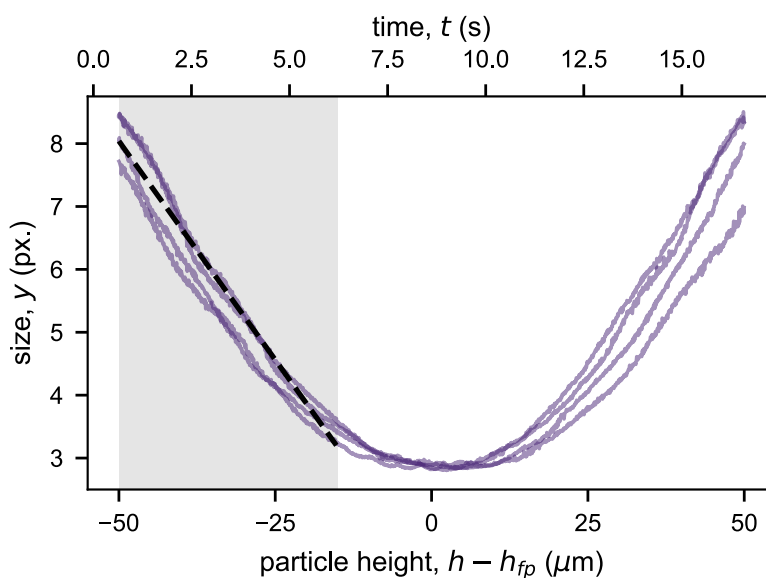
Figure S2: Particle size y as a function of time (top axis) for four PS particles sedimenting through the focal plane of the microscope. The bottom axis shows the particle height $h - h_{\mathrm{fp}}$ relative to that of the focal plane using an estimated sedimentation speed of $U = 6.6$ μm/s. All measurements reported the main text are made within the shaded region, where the relationship between y and h is approximately linear (dashed black line).

2 Inference

2.1 Marginalizing over ϕ_{ij}

To facilitate analysis of the full probability model, it is convenient to marginalize over the particle parameters ϕ_{ij} to obtain

$$p(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{y} \mid \mathbf{d}) = p(\boldsymbol{\theta}) \prod_{i=1}^{N_{\mathrm{e}}} \prod_{j=1}^{N_{\mathrm{p}}} p(\lambda_{ij} \mid \boldsymbol{\theta}, \mathbf{d}_i) p(\mathbf{y}_{ij} \mid \lambda_{ij}, \mathbf{d}_i) \tag{S1}$$

where the marginal likelihood for the data $\mathbf{y}_{ij}$ for particle j in experiment i is given by

$$p(\mathbf{y}_{ij} \mid \lambda_{ij}) = \int p(\mathbf{y}_{ij} \mid \lambda_{ij}, \boldsymbol{\phi}_{ij}, \mathbf{d}_i) p(\boldsymbol{\phi}_{ij}) d\boldsymbol{\phi}_{ij} \tag{S2}$$

This integral can be performed analytically using the properties of the multivariate normal distribution. In the analysis below, we omit temporarily the subscripts ij to streamline the notation.

We first consider the posterior distribution for the parameters $p(\boldsymbol{\phi} \mid \mathbf{y}, \lambda)$, which can be obtained using standard methods of linear regression. Using Bayes theorem, the posterior is given by

$$p(\boldsymbol{\phi} \mid \mathbf{y}, \lambda) = \frac{p(\mathbf{y} \mid \boldsymbol{\phi}, \lambda) p(\boldsymbol{\phi})}{p(\mathbf{y} \mid \lambda)} \tag{S3}$$

Substituting the likelihood and the prior, the posterior can be expressed as

$$\begin{aligned}
p(\boldsymbol{\phi} \mid \mathbf{y}, \lambda) &= c \exp\left(-\frac{1}{2s^2}(\mathbf{y} - X^\top \boldsymbol{\phi})^\top (\mathbf{y} - X^\top \boldsymbol{\phi})\right) \exp\left(-\frac{1}{2}\boldsymbol{\phi}^\top \Sigma^{-1} \boldsymbol{\phi}\right) \\
&= c \exp\left(-\frac{1}{2}(\boldsymbol{\phi} - \boldsymbol{\phi}_o)^\top A^{-1}(\boldsymbol{\phi} - \boldsymbol{\phi}_o)\right) \exp\left(-\frac{\mathbf{y}^\top \mathbf{y}}{2s^2} + \frac{1}{2}\boldsymbol{\phi}_o^\top A^{-1} \boldsymbol{\phi}_o\right)
\end{aligned} \tag{S4}$$

where X is the $2 \times N_{\mathrm{t}}$ design matrix, Σ is the prior covariance matrix, $\boldsymbol{\phi}_o = AX\mathbf{y}/s^2$ is the posterior mean, $A = (\Sigma^{-1} + XX^\top/s^2)^{-1}$ is the posterior convariance matrix, and the normalization constant is $c = 1/p(\mathbf{y} \mid \lambda)(2\pi s^2)^{N_{\mathrm{t}}/2}|2\pi\Sigma|^{1/2}$. Integrating over the nuisance parameters $\boldsymbol{\phi}$, one obtains the the following expression for the marginal likelihood

$$p(\mathbf{y} \mid \lambda) = \frac{1}{(2\pi s^2)^{N_{\mathrm{t}}/2}} \frac{|A|^{1/2}}{|\Sigma|^{1/2}} \exp\left(-\frac{\mathbf{y}^\top \mathbf{y}}{2s^2} + \frac{1}{2}\boldsymbol{\phi}_o^\top A^{-1} \boldsymbol{\phi}_o\right) \tag{S5}$$

This result applies independently to each particle j and experiment i.

2.2 Marginalizing over $\boldsymbol{\lambda}$

Marginalizing over λ is more challenging as the likelihood $p(\mathbf{y} \mid \lambda)$ of equation (S5) is, in general, a complex function of λ.[1] However, for effective designs, this function is well approximated by a lognormal distribution

$$p(\mathbf{y} \mid \lambda) \approx c_\lambda \, \mathrm{Lognormal}(\lambda \mid \mu_\lambda, \sigma_\lambda^2) \tag{S6}$$

---

[1]As in the previous section, we omit the subscripts i and j throughout this section; these results apply to each experiment i and particle j.
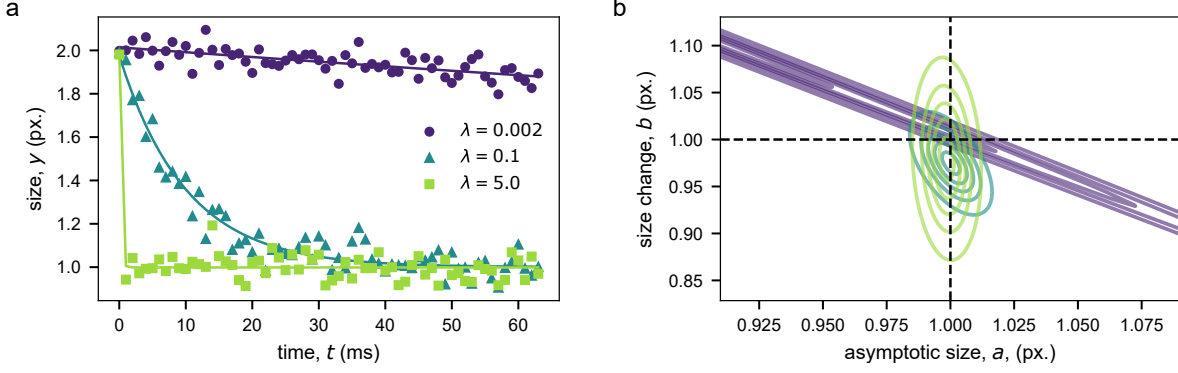
Figure S3: Analysis of simulated data illustrating the marginalization over $\phi = \{a, b\}$. (a) Simulated size data y as a function of time for three different relaxation rates λ. Markers are the $N_t = 64$ simulated data points with noise level $s = 0.1$ pixels; the solid curves show the predicted behavior for the MAP parameters. The constant time step is $\Delta t = 1$ ms. The prior standard deviations for a and b are $\sigma_a = \sigma_b = 2$ pixels. (b) Posterior distributions for the asymptotic size a and the size change b for the three data sets in (a) fitted using the true values for λ. The dashed lines show the true values used to generate the data: $a = 1$ pixel and $b = 1$ pixel.

where the data-dependent quantities c_λ, μ_λ, and σ_λ^2 serve as sufficient statistics for each particle j in experiment i. With this approximation, one can marginalize over λ analytically as

$$p(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{d}) = \int p(\mathbf{y} \mid \lambda)p(\lambda \mid \boldsymbol{\theta}, \mathbf{d})d\lambda$$
$$= \frac{c_\lambda}{(2\pi(\sigma_\lambda^2 + \sigma^2))^{1/2}} \exp\left(-\frac{(\mu_\lambda - \mu)^2 + 2\mu\sigma_\lambda^2 + 2\mu_\lambda\sigma^2 - \sigma_\lambda^2\sigma^2}{2(\sigma_\lambda^2 + \sigma^2)}\right) \tag{S7}$$

where the parameters c_λ, μ_λ, and σ_λ depend on the data $\mathbf{y}$ while the quantities μ and σ depend on the global parameters $\boldsymbol{\theta}$.

The approximation of equation (S6) is performed by numerical computation of the zeroth, first, and second moments of $p(\mathbf{y} \mid \lambda)$ within a specified range $\lambda_{\min} < \lambda < \lambda_{\max}$. Specifically, the parameters c_λ, μ_λ, and σ_λ^2 are defined as

$$c_\lambda = \int_{\lambda_{\min}}^{\lambda_{\max}} p(\mathbf{y} \mid \lambda)d\lambda \tag{S8}$$

$$\mu_\lambda = \frac{1}{c_\lambda} \int_{\lambda_{\min}}^{\lambda_{\max}} \ln\lambda \ p(\mathbf{y} \mid \lambda)d\lambda \tag{S9}$$

$$\sigma_\lambda^2 = \frac{1}{c_\lambda} \int_{\lambda_{\min}}^{\lambda_{\max}} (\ln\lambda - \mu_\lambda)^2 \ p(\mathbf{y} \mid \lambda)d\lambda \tag{S10}$$

Figure S4 illustrates the performance of this approximation for three different values of the decay rate λ.

We set the upper limit of the integral to $\lambda_{\max} = 100/\Delta t$ where Δt is the interval between successive time points. The value of 100 is selected to ensure an uninformative measurement (large σ_λ) when $\lambda > \Delta t^{-1}$. In the limit of large λ, the marginal log-likelihood has the asymptotic form: $\ln p(y \mid \lambda) = b_0 + b_1 e^{-\lambda t} + \ldots$ for $\lambda > 1$. For small λ, the marginal log-likelihood approaches a constant value as $\ln p(y \mid \lambda) = c_0 + c_1\lambda + c_2\lambda^2 + \ldots$, which is valid when $\lambda \ll c_1/c_2$. We use this

approximation to select $\lambda_{\min}$ such that the deviation of $\ln p(\mathbf{y} \mid \lambda)$ from its asymptotic value of c_0 is equal to a specified tolerance of 10^{-3}. The moment integrals are computed numerically using the cubature package for adaptive integration in Python.[1]
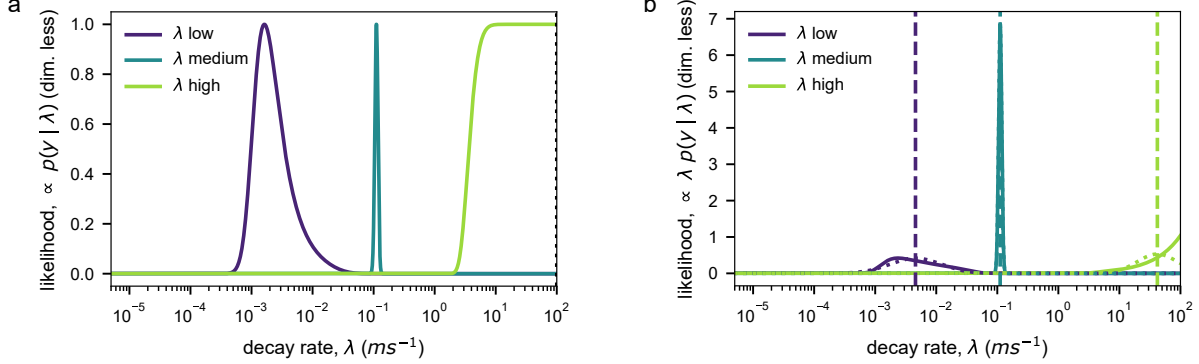

Figure S4: (a) Marginal likelihood $p(\mathbf{y} \mid \lambda)$ for the simulated data in Figure S3a. The height of each curve is scaled to one. (b) Marginal likelihood $p(\mathbf{y} \mid \ln\lambda) = \lambda p(\mathbf{y} \mid \lambda)$ parameterized by the logarithm of λ along with the log-normal fit of equation (S6) (dotted curve). With this parameterization the fits are proportional to the familiar bell-shaped curve of the normal distribution. The vertical dashed lines show the parameter μ_λ. The distributions are normalized to unity with respect to integration over $\ln\lambda$.

2.3 Estimating the evidence $p(\mathbf{y} \mid \mathbf{d})$

In evaluating the information gain of an experiment—actual or simulated—we require the marginal likelihood (also called the evidence)

$$p(\mathbf{y} \mid \mathbf{d}) = \int p(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{d}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \qquad \text{(S11)}$$

We approximate the evidence from the posterior parameter samples using the Monte Carlo algorithm of Heavens et al.[2] We first apply a linear whitening transformation to the sampled parameters $\{\boldsymbol{\theta}_m\}$ such that the resulting covariance matrix of the transformed parameters $\{\tilde{\boldsymbol{\theta}}_m\}$ is equal to the identity matrix. Briefly, we compute the eigendecomposition of the covariance matrix

$$\mathbf{C} = \langle (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^T \rangle = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T \qquad \text{(S12)}$$

where the angled brackets denote sample averages, $\bar{\boldsymbol{\theta}} = \langle \boldsymbol{\theta} \rangle$ is the sample mean, $\mathbf{Q}$ is an orthogonal matrix of eigenvectors, and $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues. The whitened parameters are then evaluated as

$$\tilde{\boldsymbol{\theta}} = \boldsymbol{\Lambda}^{-1/2} \mathbf{Q}^T (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) \qquad \text{(S13)}$$

Second, we compute the 1st nearest neighbor distances $\{\tilde{D}_n\}$ from each of the N_θ transformed parameters $\{\tilde{\boldsymbol{\theta}}_m\}$ (using the scikit-learn library). The maximum posterior value of the evidence $E = p(\mathbf{y})$ is then given by

$$E_o = \frac{N_\theta \sqrt{\det(\boldsymbol{\Lambda})}}{N_\theta + 1} \sum_{n=1}^{N_\theta} V_m(\tilde{D}_n) p(\mathbf{y} \mid \boldsymbol{\theta}_n) p(\boldsymbol{\theta}_n) \qquad \text{(S14)}$$

6

where $V_m(D) = \pi^{m/2} D^m / \Gamma(1 + m/2)$ is the volume of the m-ball of radius D (here, $m = 4$ corresponding to the four parameters of interest). Note that values of the un-normalized posterior $p(\mathbf{y}|\boldsymbol{\theta}_n)p(\boldsymbol{\theta}_n)$ are saved during MCMC sampling and need not be evaluated twice. The variance of the posterior is approximately, $\sigma_E^2 \approx E_o^2 / N_\theta$, assuming the sampled parameters are independent of one another.

3 Design

3.1 Two-stage design process

As described in the Methods, we select the design $\tilde{\mathbf{d}}$ for the next experiment that maximizes the mutual information between the future outcome $\tilde{\mathbf{y}}$ of that experiment and the cell level parameters $\boldsymbol{\theta}$

$$\mathbf{d}^\star = \arg\max U(\tilde{\mathbf{d}}) \quad \text{with} \quad U(\tilde{\mathbf{d}}) = I(\tilde{\mathbf{y}}; \boldsymbol{\theta}) \tag{S15}$$

The Venn diagram of Figure S5 illustrates the relationships of various information measures associated with the three model variables $\boldsymbol{\theta}$, $\tilde{\boldsymbol{\lambda}}$, and $\tilde{\mathbf{y}}$. Using these relationships, the objective function $U(\tilde{\mathbf{d}})$ can be expressed by differences between a mutual information and a conditional mutual information as

$$U(\tilde{\mathbf{d}}) = I(\boldsymbol{\theta}; \tilde{\boldsymbol{\lambda}}) - I(\boldsymbol{\theta}; \tilde{\boldsymbol{\lambda}} \mid \tilde{\mathbf{y}}) = I(\tilde{\boldsymbol{\lambda}}; \tilde{\mathbf{y}}) - I(\tilde{\boldsymbol{\lambda}}; \tilde{\mathbf{y}} \mid \boldsymbol{\theta}) \tag{S16}$$

where $I(\boldsymbol{\theta}; \tilde{\boldsymbol{\lambda}})$ depends only on the frequency $\tilde{\omega}$, and the other terms depend on both the frequency $\tilde{\omega}$ and the frame rate $\tilde{f}$.

The information shared between the data $\tilde{\mathbf{y}}$ and the parameters $\boldsymbol{\theta}$ is transmitted by way of the latent variable $\tilde{\boldsymbol{\lambda}}$ (Fig. S5). The transmission of information along such a linear chain is often determined by an information "bottleneck", which limits the amount of information shared. Here, the bottleneck is the relationship between the cell-level parameters $\boldsymbol{\theta}$ and the latent rate parameters $\tilde{\boldsymbol{\lambda}}$, such that $U(\tilde{\mathbf{d}}) \approx I(\tilde{\boldsymbol{\lambda}}; \boldsymbol{\theta})$ for near optimal designs. As the mutual information $I(\tilde{\boldsymbol{\lambda}}; \boldsymbol{\theta})$ depends on the frequency $\tilde{\omega}$ but not the frame rate $\tilde{f}$, we first identify the optimal frequency using the approximate procedure

$$\omega^\star \approx \arg\max_{\tilde{\omega}} I(\tilde{\boldsymbol{\lambda}}; \boldsymbol{\theta}) \tag{S17}$$

where $I(\tilde{\boldsymbol{\lambda}}; \boldsymbol{\theta})$ is approximated using nested Monte Carlo integration (see Methods).
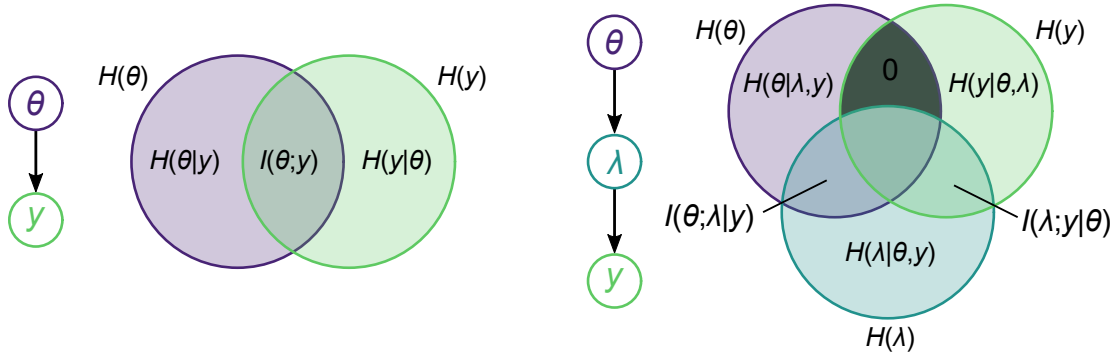
Figure S5: (left) Venn diagram showing information measures associated with the model variables $\boldsymbol{\theta}$ and $\mathbf{y}$. The circle on the left is the entropy $H(\boldsymbol{\theta})$, with the conditional entropy $H(\boldsymbol{\theta} \mid \mathbf{y})$ in purple. The circle on the entropy is $H(\mathbf{y})$, with the the conditional entropy $H(\mathbf{y} \mid \boldsymbol{\theta})$ in green. The overlap region between the two circles is the mutual information $I(\boldsymbol{\theta}; \mathbf{y})$. (right) Venn diagram showing information measures associated with three model variables $\boldsymbol{\theta}$, $\boldsymbol{\lambda}$, and $\mathbf{y}$. The conditional independence between the data $\mathbf{y}$ and the cell parameters $\boldsymbol{\theta}$ given the rate parameter $\boldsymbol{\lambda}$ implies that the conditional information $I(\boldsymbol{\theta}; \mathbf{y} \mid \boldsymbol{\lambda})$ is zero (grey region). The expected utility is the central overlapping region, $U = I(\boldsymbol{\theta}; \mathbf{y})$.

In the second stage of the design process, the optimal frame rate $f^\star$ can be determined by minimizing the second term of equation (S16) evaluated at the optimal frequency $\omega^\star$

$$f^\star = \arg\min_{\tilde{f}} I(\tilde{\boldsymbol{\lambda}}; \boldsymbol{\theta} \mid \tilde{\mathbf{y}}) \tag{S18}$$

Importantly, it is not essential to identify the "optimal" frame rate but rather a "good" frame rate for which $I(\tilde{\boldsymbol{\lambda}}; \boldsymbol{\theta} \mid \tilde{\mathbf{y}}) \ll I(\tilde{\boldsymbol{\lambda}}; \boldsymbol{\theta})$. Under these conditions, the expected utility of the experiment is determined by the information bottleneck—namely, the relationship between the cell parameters $\tilde{\boldsymbol{\theta}}$ and the rate parameters $\tilde{\boldsymbol{\lambda}}$ independent of $\tilde{f}$. For this reason, we approximate the optimal frame rate using a more convenient objective function

$$f^\star \approx \arg\max_{\tilde{f}} I(\tilde{\boldsymbol{\lambda}}; \tilde{\mathbf{y}}) \tag{S19}$$

where $I(\tilde{\boldsymbol{\lambda}}; \tilde{\mathbf{y}})$ is estimated using nested Monte Carlo integration (see Methods). By choosing the frame rate that maximizes the information shared between the rate parameter $\tilde{\boldsymbol{\lambda}}$ and the data $\tilde{\mathbf{y}}$, we ensure that this relationship is not the bottleneck between the cell parameters $\boldsymbol{\theta}$ and the data $\tilde{\mathbf{y}}$.

The intuition underlying this heuristic is illustrated by a simple example containing three normal variables x, y, and z within a linear Bayesian network: $x \to y \to z$. The joint distribution of these variables is

$$p(x, y, z \mid \sigma_x, \sigma_y, \sigma_z) = \mathcal{N}(x \mid 0, \sigma_x^2)\mathcal{N}(y \mid x, \sigma_y^2)\mathcal{N}(z \mid y, \sigma_z^2) \tag{S20}$$

which is a multivariate normal distribution with zero mean. Without loss of generality, we can set $\sigma_x = 1$ such that each variable is measured in units of σ_x. The amount of information shared between the variables x and z is determined by the two parameters σ_y and σ_z, which characterize the respective relationships between x and y and between y and z

$$I(x; z) = -\frac{1}{2} \ln\left(1 - \frac{\sigma_x^2}{\sigma_x^2 + \sigma_y^2 + \sigma_z^2}\right) \tag{S21}$$

Figure S6 shows how this and other information measures depend on the parameter σ_z when σ_y is held constant. When σ_z falls below some critical value (namely, when $\sigma_z^2 \ll \sigma_y\sqrt{\sigma_x^2 + \sigma_y^2}$), the mutual information $I(x; z)$ is no longer influenced by this parameter. Under these conditions the relationship between x and y becomes the information bottleneck, and the overall objective function is well approximated as $I(x; z) \approx I(x; y)$. Moreover, we see that this desired outcome can be achieved by maximizing $I(y; z)$ with respect to σ_z.
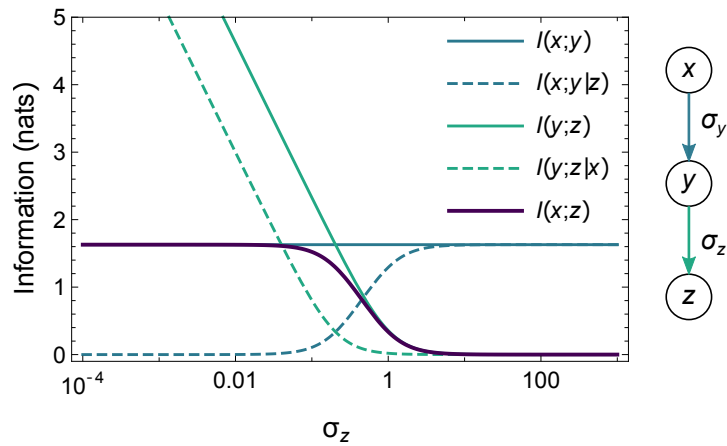
Figure S6: Various information measures among the variables x, y, z as a function of the parameter σ_z for the Gaussian Bayesian network show at the right with joint distribution given by equation (S20) with $\sigma_x = 1$ and $\sigma_y = 0.2$. These measures are related as $I(x;z) = I(x;y) - I(x;y \mid z) = I(y;z) - I(y;z \mid x)$.

4 Automated optimal experiments
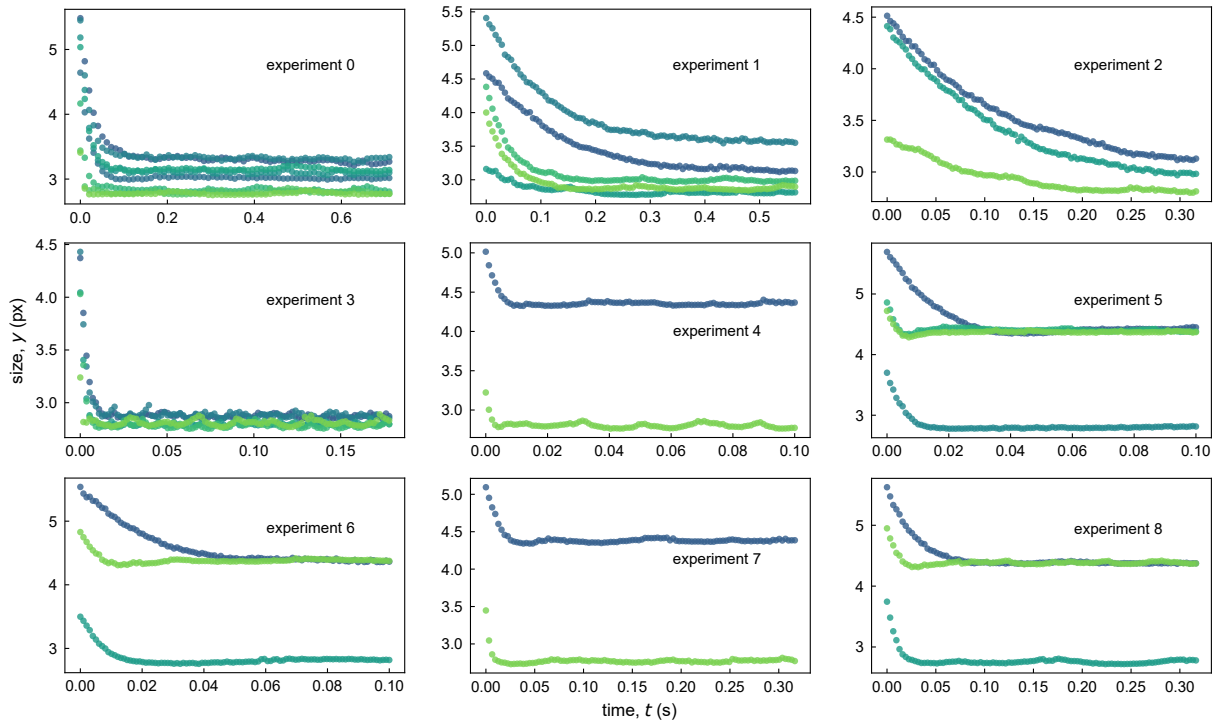
4.1 Size data for nine optimal experiments

Figure S7: Raw data for particle sizes vs. time for an automated series of nine optimal experiments (corresponding to those in Figs. 5 & 6). For each experiment, the different colored markers correspond to the different tracked particles. The number of tracked particles N_{p} typically decreases from one experiment to the next as particles begin to aggregate at the nodal plane (see Fig. S1). The number of time points for each experiment is fixed at $N_{\mathrm{t}} = 100$ although the frame rate, and thereby the total observation time, may vary.

4.2 Inference and design for nine optimal experiments
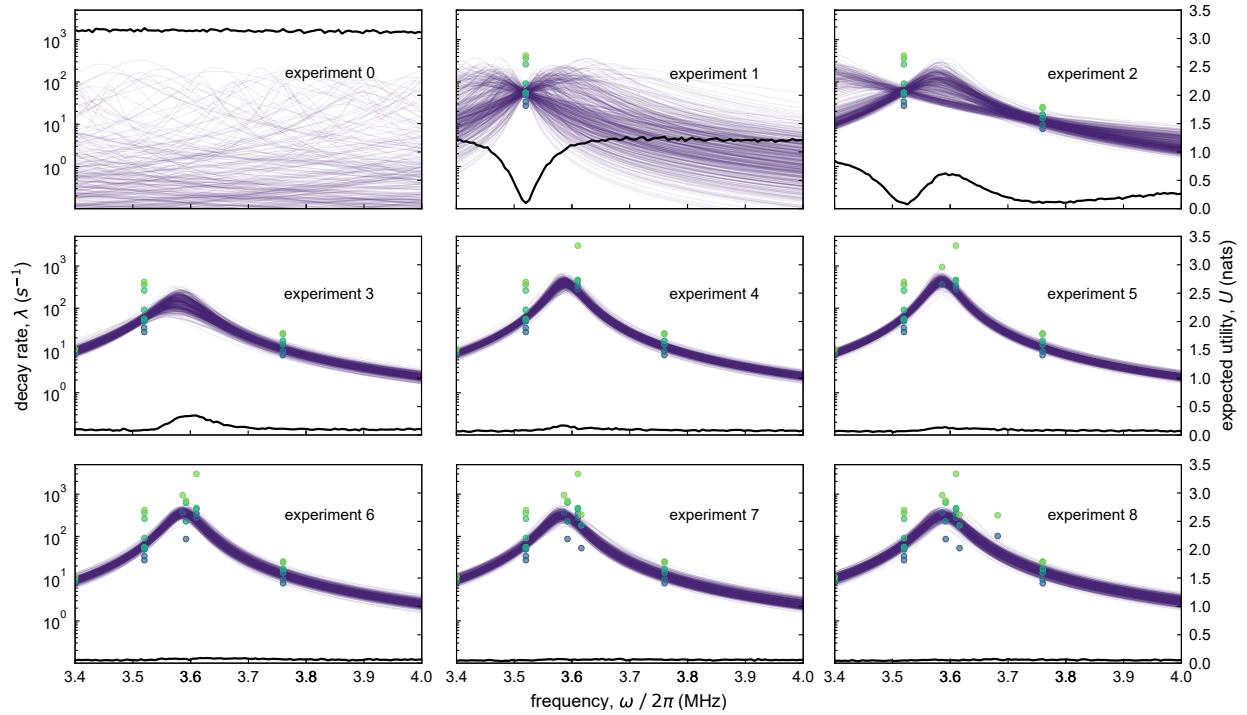

Figure S8: Inference and design in an automated series of eight optimal experiments (corresponding to those in Fig. S7 and Figs. 5 & 6). Predicted rate parameter $\lambda(\omega)$ as a function of frequency ω based on the posterior distribution for each experiment. Each purple curve is evaluated using equation (5) with parameter values α, β, γ drawn from the posterior distribution. Markers show estimates of the rate parameter λ_{ij} for each experiment i and particle j. The solid black curve represents the expected utility U of a subsequent experiment as a function of the driving frequency.

References

1. Castro, S. & Loukianov, A. *Python wrapper for Cubature: adaptive multidimensional integration, Version 0.14.3* 2020.

2. Heavens, A., Fantaye, Y., Mootoovaloo, A., Eggers, H., Hosenie, Z., Kroon, S. & Sellentin, E. Marginal Likelihoods from Monte Carlo Markov Chains, arXiv:1704.03472 (2017).