PHYSICAL REVIEW LETTERS

Likelihood-Free Methods for Quantum Parameter Estimation

Christopher Ferrie[1] and Christopher E. Granade[2]

[1]*Center for Quantum Information and Control, University of New Mexico, Albuquerque, New Mexico 87131-0001, USA*
[2]*Institute for Quantum Computing and Department of Physics, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada*

In this Letter, we strengthen and extend the connection between simulation and estimation to exploit simulation routines that do not exactly compute the probability of experimental data, known as the likelihood function. Rather, we provide an explicit algorithm for estimating parameters of physical models given access to a simulator which is only capable of producing sample outcomes. Since our algorithm does not require that a simulator be able to efficiently compute exact probabilities, it is able to exponentially outperform standard algorithms based on exact computation. In this way, our algorithm opens the door for the application of new insights and resources to the problem of characterizing large quantum systems, which is exponentially intractable using standard simulation resources.

Much of physics is concerned with modeling complex behavior such that we can simulate systems of interest and can infer properties of those systems. On one hand, estimating the parameters of physical models given experimental data is critical to many practical objectives, such as precision metrology for frequency standards [1,2], and to probing fundamental questions, such as gravitational wave detection [3]. On the other hand, by simulating physical models, we can understand properties of the systems that follow those models. That is, by using simulation to reason about the probabilities of experimental data produced by physical models, we can expose how experimental observations will depend on properties of interest.

Thus, these two concerns are not independent, such that parameter estimation can be broadly thought of as choosing as our estimated model parameters those for which simulations predict the highest probability of obtaining data that agree with the observed experimental data. Once we have estimated parameters for a model, we can use those parameters to predict the future behavior of an experimental system by simulating according to those parameters. In this way, simulation and statistical estimation are seen to be intimately related.

In this work, we present evidence of this relationship in the case of weak simulation, in which one has access only to samples from a simulator rather than the explicit distributions. This is in contrast to a strong simulator, which produces the exact probabilities of each possible outcome of an experiment (see Fig. 1). The task of estimation is a statistical one and, in the language of statistics, strong simulation is equivalent to explicitly calculating the likelihood function. Many common estimation algorithms rely on explicit calculations of likelihood function and, hence, on strong simulation. Here, we rectify the situation by providing a method to perform statistical estimation of parameters given access to only a weak simulator [4]. In addition to being generally applicable in the estimation of physical parameters, our approach is necessary in quantum certification protocols making use of quantum resources [5,6].

The distinction between strong and weak simulation is particularly important when considering quantum mechanical models, where we are only beginning to broadly appreciate the difference [7]. In particular, it has been shown that many quantum mechanical models admit efficient weak simulation on a classical computer where strong simulation is exponentially more difficult [8–11]. Thus, a characterization method that depends only on weak simulation can exhibit a large advantage over strong-simulation characterization methods.

This advantage is especially imperative in the case of quantum information, as the number of parameters that must be measured in a tomographic experiment grows exponentially with the number of qubits. Though tomographic experiments have been carried out in systems as large as several qubits [12,13], the exponential nature of the problem prevents the extension of tomographic methods to large-scale quantum information processing devices, such as those currently being proposed [14,15]. Thus, in order to
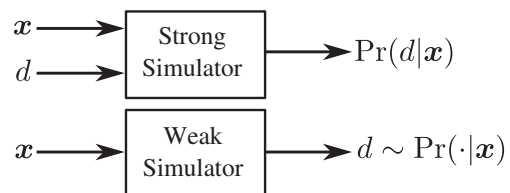

FIG. 1. Strong and weak simulators. A strong simulator computes the value of the likelihood function $\Pr(\cdot|\boldsymbol{x})$, given a set of parameters $\boldsymbol{x}$ and data d. By contrast, a weak simulator produces sample data d, drawn from the likelihood function, given only $\boldsymbol{x}$.

develop useful quantum information processing devices, it is necessary to develop novel and efficient statistical inference methods that can exploit prior information, reductions in model dimension and weak simulation. Since our algorithm needs only a weak simulator, and does not require calculation of the likelihood function itself, we call our algorithm the "likelihood-free parameter estimation" (LFPE) algorithm [16].

Parameter estimation problems can be phrased in the following general terms. To each physical model is associated a probability distribution $\Pr(d|\mathbf{x})$, where d is the data obtained and where $\mathbf{x}$ is a vector parameterizing the system of interest. In statistical parlance, this distribution is called the "likelihood function."

Now, suppose we have performed experiments and obtained a data set $D:=\{d_1, d_2, \ldots, d_N\}$. We assume that experiments are statistically independent so that the likelihood function becomes

$$\Pr(D|\mathbf{x}) = \prod_{k=1}^{N} \Pr(d_k|\mathbf{x}). \tag{1}$$

However, we are ultimately interested in $\Pr(\mathbf{x}|D)$, the probability distribution of the model parameters $\mathbf{x}$ given the experimental data. We obtain this using use Bayes' rule,

$$\Pr(\mathbf{x}|D) = \frac{\Pr(D|\mathbf{x})\Pr(\mathbf{x})}{\Pr(D)}, \tag{2}$$

where $\Pr(\mathbf{x})$ is the *prior*, which encodes any *a priori* knowledge of the model parameters. The final term $\Pr(D)$ can simply be found implicitly by normalizing the posterior. Since each measurement is statistically independent given $\mathbf{x}$, the processing of the data can be done on- or off-line. That is, we can sequentially update the probability distribution as the data arrive or postprocess it afterward. After all the data have been taken, we report the mean of the posterior distribution as our estimate of the parameters:

$$\hat{\mathbf{x}}(D) = \mathbb{E}_{\mathbf{x}|D}[\mathbf{x}] = \int \mathbf{x}\,\Pr(\mathbf{x}|D)d\mathbf{x}. \tag{3}$$

This method of parameter estimation is called Bayesian learning, and has been shown to be the optimal approach in a more general decision theoretic framework [17]. The meaning of this optimality is precisely that Eq. (3) minimizes the mean squared error (MSE) figure of merit: $\mathrm{MSE}(\hat{\mathbf{x}}) = \mathbb{E}_{\mathbf{x},D}[(\mathbf{x} - \hat{\mathbf{x}}(D))^2]$.

In order to efficiently compute the integral expectation in Eq. (3), we employ the *sequential Monte Carlo* (SMC) method, which has been used for the purpose of Hamiltonian learning [18] and in the tomographic estimation of one and two qubit states [19], and in the continuous measurement of a qubit [20].

The SMC method prescribes that we should approximate a distribution over model parameters with a distribution that has support only over a finite number of points (often referred to as "particles"). Each particle is assigned a weight, informally thought of as its relative plausibility. More concretely, we approximate the posterior distribution at the Nth measurement by

$$\Pr(\mathbf{x}|D) \approx \sum_{k=1}^{n} w_k(d_N)\delta(\mathbf{x} - \mathbf{x}_k), \tag{4}$$

where the weights at each step are iteratively calculated from the previous step via

$$w_k(d_{j+1}) = \Pr(d_{j+1}|\mathbf{x}_k)w_k(d_j)/\mathcal{N}, \tag{5}$$

where $\mathcal{N}$ is found implicitly by imposing the normalization condition $\sum_k w_k(d_{j+1}) = 1$. The positions $\{\mathbf{x}_i\}$ of each particle are sampled according to the prior $\Pr(\mathbf{x})$. The particle approximation can be made arbitrarily accurate by increasing the number of particles. The initial weights, when no data (denoted d_0) has been observed, are given by $w_k(d_0) = 1/n$ for all k. This choice is made to ensure that the effective sample size $n_{\mathrm{ess}}:=1/\sum_i w_i^2$ is initially n. As $n_{\mathrm{ess}} \to 0$, the algorithm becomes numerically unstable and fails to explore the parameter space; this may be recovered by a resampling step [21]. We explored some variants of this algorithm and presented it in much greater detail in Ref. [18].

Equation (5) suggests that we require a full specification of the likelihood function $\Pr(d|\mathbf{x})$. Suppose, however, we have access to only a weak simulator, which produces outcomes $d \sim \Pr(\cdot|\mathbf{x})$ [22]. One extreme is to run the simulator many times and reconstruct $\Pr(d|\mathbf{x})$ from the simulated data—a metaestimation problem. At the other extreme is to perform estimation with only one sample per SMC particle. The method truly becomes "likelihood free" as we could not even hope to guess the functional form of the likelihood function from a single sample.

In the extreme case where the weak simulator is used to very accurately compute the likelihood function via repeated sampling, the SMC algorithm does not change. At the opposite extreme, when only a single sample is generated from the simulator per particle, we must modify the algorithm. To this end, suppose we have obtained data d from the experiment. For each SMC particle, $\mathbf{x}_k$, we request a single sample d'_k from our simulator and update the weight as follows:

$$w_k = \begin{cases} 1 & \text{if } d = d'_k \\ 0 & \text{otherwise} \end{cases}. \tag{6}$$

Between the two extremes of a single simulator sample per particle and enough to compute the likelihood function nearly exactly, we can approximately reconstruct the

likelihood function sets of simulated data. In particular, for each datum d and particle $\mathbf{x}_k$, we draw a set of samples D'_k from our simulator. We then update the weights according to Eq. (5) with estimated likelihood function given by the naive maximum likelihood estimator

$$\Pr(d|\mathbf{x}_k) \approx \frac{|\{d' \in D'_k : d' = d\}|}{|\{D'_k\}|}. \qquad (7)$$

As mentioned above, we will measure the performance of our algorithms with the mean squared error. In our Bayesian setting, this is also the variance of the posterior distribution. Since the experiments are assumed independent and identically distributed, the posterior variance will decrease as $O(1/N)$, where N is the total number of measurements. We appeal to standard Monte Carlo analyses which suggest that the SMC algorithm will increase this variance by at least $O(1/n)$, where n is the number of SMC particles. Now, if we use a weak simulator with a fixed experiment and particle number, the same statistical argument suggests that the variance will scale as $O(1/m)$, where m is the number of simulator calls we use (per particle) to estimate the likelihood function. Since the total number of samples is nm, we expect the mean square error to scale as

$$\mathrm{MSE}(\hat{\mathbf{x}}) \sim \frac{a}{N} + \frac{b}{nm}, \qquad (8)$$

for constants a and b depending only on the parameters of the problem.

To verify these claims, we perform numerics. Our example is that of a noisy photodetector where the efficiency of the photon source is p, which we would like to estimate. This value is equivalent to the probability for the detector to click in the presence of no noise. In reality, dark counts register clicks when no photon is present and

losses register no clicks when a photon is present. Let these happen with probability α and β, respectively. Then, given p, the probability for a click to actually happen is $\Pr(\text{click}|p) = p(1-\beta) + (1-p)\alpha$. From these clicks, our task is to estimate p.

Aysmptotically, the posterior variance is given by the inverse of the Fisher information evaluated at, for example, the maximum likelihood estimate [23]. The details of this calculation are presented in the Supplemental Material [24]. The result is the asymptotic bound

$$\mathrm{MSE}(p) \geq \frac{1}{6(1-\alpha-\beta)^2 N}, \qquad (9)$$

which we will use to verify our algorithm is near optimal. In practice, p will be a function of some parameters of interest, $p = p(\mathbf{x})$. We restrict ourselves to this example in order to illustrate the effects on inference due to weak simulation.

First, we verify that given a fixed number of experiments, the MSE scales as $O(1/n)$ (where n is again the number of SMC particles) for both the strong simulating SMC algorithm and likelihood-free weak simulation. The data, plotted in Fig. 2 (left), bears out our expectations quite convincingly; even in the case of a single sample from the simulator, the accuracy can be increased (at the expected $O(1/n)$ rate) until it reaches the bound given by Eq. (9). Next, in Fig. 2 (middle), we show that fixing the number of particles and varying the number of simulations per particle, m, results in an MSE that scales as $O(1/m)$. Thus, as expected, the more accurately we can compute the likelihood function, the better our accuracy will be—but only up to a certain point. That is, it is not advantageous to continue improving the accuracy of the estimate of the likelihood function beyond roughly $1/N + 1/n$ since the errors from finite particles and samples will begin to dominate.
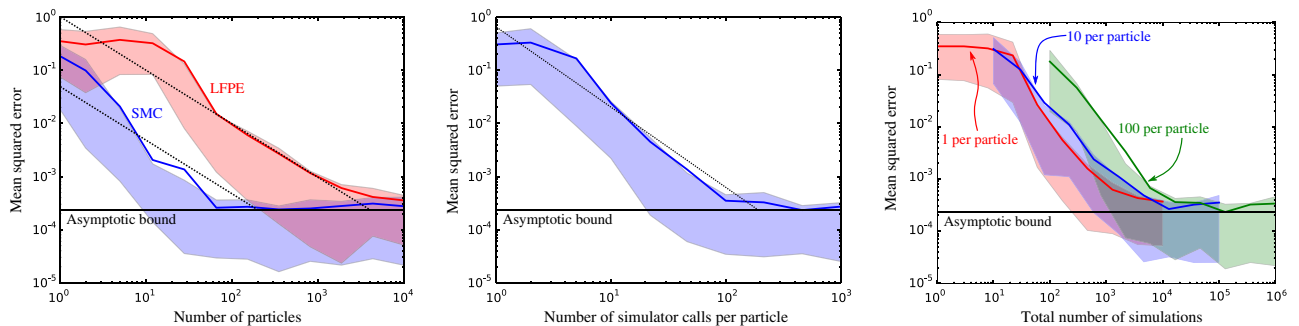

FIG. 2 (color online). Left: the MSE using a strong simulator (SMC) and weak simulator (LFPE) as a function of the number of particles. For the LFPE algorithm, a single sample from the simulator per particle is used to perform the inference. The dotted line gives the conjectured $O(1/n)$ scaling. Middle: the MSE using the LFPE algorithm as a function of the number of simulator calls per particles. The number of particles is fixed at $n = 100$. The dotted lines give the claimed $O(1/m)$ scaling. Right: the MSE of the LFPE algorithm as a function of the total number of simulator calls for varying number of simulator calls per particle. Here, $\alpha = 0.9$ and $\beta = 0.05$ and the number of measurements is fixed at $N = 1000$. The solid lines indicate the mean over 100 trials while the shaded areas represent the interquartile range (where the middle half of data lie). The black solid line is the asymptotic bound given in Eq. (9).

On the other hand, the strategy of estimating the likelihood via samples ignores the cost of simulation. The total number of simulations is nm, the number of particles times the number of samples per particle. In Fig. 2 (right), we plot the MSE against the total number of simulator calls and find that, perhaps surprisingly, the likelihood-free approach of using a single simulator sample is best.

In the above arguments, the total number of measurements was held fixed to verify the performance as a function of the algorithmic parameters. If, on the other hand, simulations are relatively cheap compared to obtaining experimental samples, we would like to optimize performance by finding the appropriate number of simulated experiments without going beyond the redundancy noted above. However, in most cases, the limit of accuracy is not *a priori* known. In such cases, we have devised an algorithm we call "adaptive likelihood estimation" (ALE). Essentially, our algorithm adaptively calls the simulator until we deem the accuracy in our estimate sufficient.

For brevity, we will discuss the binary case with outcomes labeled 0 and 1. The unknown probability $p_0 := \Pr(0|\mathbf{x})$ can be treated as a parameter to be estimated. In particular, since we have assumed that the data are conditionally independent given the model, repeatedly sampling the likelihood function will produce data that follows a binomial distribution with parameter p_0. Estimating the parameter of a binomial distribution from sample data is a well-understood statistical problem. Supposing k 0 s were observed in m trials, a typical estimator is

$$\hat{p}_{0\gamma}(k) = \frac{k+\gamma}{m+2\gamma}, \tag{10}$$

where γ is a free parameter. These are called "linear" or "add-γ" estimators [25]. The latter phrase is due to the equivalence to standard maximum likelihood estimation when adding γ fictitious observations—also termed "hedging" [26]. These estimators can also be understood to arise from a Bayesian approach as well. In particular, the estimator in Eq. (10) is the posterior mean when using the following Beta distribution as a prior [25]:

$$\Pr(p_0) \propto p_0^{\gamma-1}(1-p_0)^{\gamma-1}. \tag{11}$$

The posterior variance of this distribution can also be calculated as

$$\hat{\sigma}_{0\gamma}^2(k) = \frac{(k+\gamma)(m-k+\gamma)}{(m+2\gamma)^2(m+2\gamma+1)} = \frac{\hat{p}_{0\gamma}(1-\hat{p}_{0\gamma})}{m+2\gamma+1}. \tag{12}$$

Here we will use the value $\gamma = 1$, as it corresponds to a uniform prior distribution. We leave the optimization of this algorithmic parameter for future work.

If we are willing to tolerate an error ϵ in our reconstruction of the likelihood, then we can check after each sample if $\hat{\sigma}_{0\gamma} < \epsilon$. If not, we collect more samples until the condition is met. We therefore have a single quality parameter for this adaptive protocol: ϵ. Since this is our estimate of the variance in the estimate of the likelihood function, the MSE is expected to scale as $O(\epsilon)$ (for fixed measurement and particle number). Thus, as discussed above, the optimal choice will be $\epsilon \approx 1/N + 1/n$ since anything smaller will fast result in diminished returns—the MSE will be limited by either the number of measurements N or particles n, depending on which is smaller. We illustrate this with our example in Fig. 3.

In this work, we have demonstrated an improvement of the sequential Monte Carlo parameter estimation algorithm that allows for its extension to the case of weak (sampling) simulators. For models with fast weak simulation available, our algorithm can be seen to provide dramatic advantages in terms of classical computing costs over sequential Monte Carlo alone and at minimal cost in estimation performance. This extension allows for us to perform inference in subtheories of quantum mechanics that admit a large separation between the tractability of strong and weak simulation.

We have necessarily demonstrated these improvements for an example model in which the analytical solution was tractable. In practice, if only weak simulation is available, then standard approaches to parameter estimation making use of calculations of the likelihood function do not apply and our method is necessary. Within the confines of quantum theory, an ever growing class of weak simulation schemes have been proposed which have been proven to have an exponential separation in computational complexity between weak and strong simulation. In addition to the
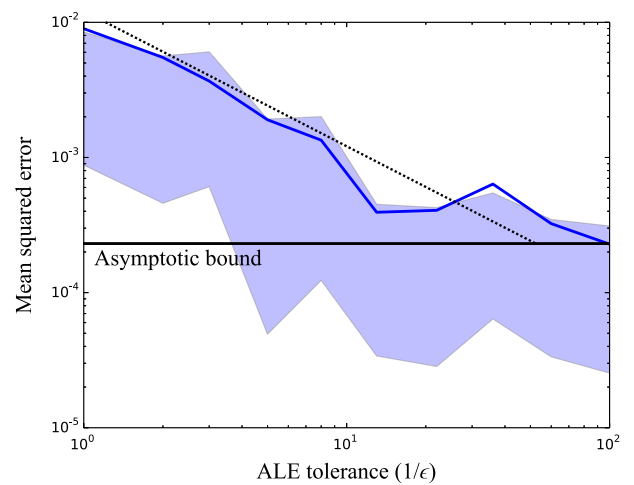

FIG. 3 (color online). The MSE using LFPE with adaptive likelihood estimation (ALE) as a function of the (inverse of the) ALE tolerance ϵ. As claimed, the MSE scales as $O(\epsilon)$ until it reaches the bound. The parameters of the problem are as in Fig. 2.

large class of circuits identified by van den Nest *et al.*, others include simulating the evolution of states with positive Wigner function [27–29]. In such cases, LFPE provides an exponential improvement in accuracy for a fixed amount of computational resources. More recently, there have been proposals for the use of quantum resources (necessarily weak simulators) to aid in overcoming the complexity in simulating the physical model [5,6]. Such ideas could mitigate the need for classical simulators to certify near-future quantum devices which go beyond the classical regime, such as BosonSamplers [7]. As the complexity of candidate quantum information processors grows, our algorithm provides a way forward to estimating properties of very large systems by exploiting the deep connection between simulation and estimation.

———————

[1] T. Udem, R. Holzwarth, and T. W. Hansch, Nature (London) **416**, 233 (2002).

[2] N. Hinkley, J. A. Sherman, N. B. Phillips, M. Schioppo, N. D. Lemke, K. Beloy, M. Pizzocaro, C. W. Oates, and A. D. Ludlow, Science **341**, 1215 (2013).

[3] J. Aasi, J. Abadie, B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, C. Adams, T. Adams, P. Addesso, R. X. Adhikari *et al.*, Nat. Photonics **7**, 613 (2013).

[4] We note that validation techniques (estimating "closeness" to some target as opposed to identifying an unknown set of parameters) have recently been proposed which take advantage of a similar notion of weak simulation [30–34]

[5] N. Wiebe, C. Granade, C. Ferrie, and D. G. Cory, arXiv:1309.0876.

[6] N. Wiebe, C. Granade, C. Ferrie, and D. G. Cory, arXiv:1311.5269.

[7] S. Aaronson and A. Arkhipov, arXiv:1011.3245.

[8] R. Jozsa and A. Miyake, Proc. R. Soc. A **464**, 3089 (2008).

[9] M. Van den Nest, Quantum Inf. Comput. **10**, 258 (2010).

[10] M. Van den Nest, Quantum Inf. Comput. **11**, 784 (2011).

[11] R. Jozsa and M. Van den Nest, arXiv:1305.6190.

[12] Y. S. Weinstein, S. Lloyd, J. Emerson, and D. G. Cory, Phys. Rev. Lett. **89**, 157902 (2002).

[13] H. Haffner, W. Hansel, C. F. Roos, J. Benhelm, D. Chek-al kar, M. Chwalla, T. Korber, U. D. Rapol, M. Riebe, P. O. Schmidt *et al.*, Nature (London) **438**, 643 (2005).

[14] N. C. Jones, R. Van Meter, A. G. Fowler, P. L. McMahon, J. Kim, T. D. Ladd, and Y. Yamamoto, Phys. Rev. X **2**, 031007 (2012).

[15] M. H. Devoret and R. J. Schoelkopf, Science **339**, 1169 (2013).

[16] We note that there exists a large class of "likelihood–free" methods, which has found use mostly in biological applications. Our approach seems to be the closest to that in Ref. [35]. However, contrary to our approach, all of the current likelihood-free methods available are based on rejection sampling.

[17] R. Blume-Kohout, New J. Phys.. **12**, 043034 (2010).

[18] C. E. Granade, C. Ferrie, N. Wiebe, and D. G. Cory, New J. Phys. **14**, 103013 (2012).

[19] F. Huszár and N. M. T. Houlsby, Phys. Rev. A **85**, 052120 (2012).

[20] B. A. Chase and J. M. Geremia, Phys. Rev. A **79**, 022314 (2009).

[21] J. Liu and M. West, *Combined Parameter and State Estimation in Simulation-Based Filtering* (Springer-Verlag, Berlin, 2000).

[22] Note that we will overload the notation $\sim$. Here we use $x \sim y$ to mean x is random variable distributed according to y. Later we use $x \sim y$ to mean x is on the order of y, which is standard asymptotic notation formally defined to mean $x/y \to 1$. The difference should be clear from context.

[23] B. S. Clarke, IEEE Trans. Inf. Theory **45**, 165 (1999).

[24] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.112.130402 for a derivation of Eq. (9).

[25] E. L. Lehmann and G. Casella, *Theory of Point Estimation* (Springer, New York, 1998), 2nd ed.

[26] R. Blume-Kohout, Phys. Rev. Lett. **105**, 200504 (2010).

[27] V. Veitch, C. Ferrie, D. Gross, and J. Emerson, New J. Phys. **14**, 113011 (2012).

[28] V. Veitch, N. Wiebe, C. Ferrie, and J. Emerson, New J. Phys. **15**, 013037 (2013).

[29] A. Mari and J. Eisert, Phys. Rev. Lett. **109**, 230503 (2012).

[30] M. P. da Silva, O. Landon-Cardinal, and D. Poulin, Phys. Rev. Lett. **107**, 210404 (2011).

[31] J. Emerson, M. Silva, O. Moussa, C. Ryan, M. Laforest, J. Baugh, D. G. Cory, and R. Laflamme, Science **317**, 1893 (2007).

[32] E. Knill, D. Leibfried, R. Reichle, J. Britton, R. B. Blakestad, J. D. Jost, C. Langer, R. Ozeri, S. Seidelin, and D. J. Wineland, Phys. Rev. A **77**, 012307 (2008).

[33] E. Magesan, J. M. Gambetta, and J. Emerson, Phys. Rev. Lett. **106**, 180504 (2011).

[34] S. T. Flammia and Y. K. Liu, Phys. Rev. Lett. **106**, 230501 (2011).

[35] S. A. Sisson, Y. Fan and M. M. Tanaka, Proc. Natl. Acad. Sci. U.S.A. **104**, 1760 (2007).