

# Inheritance and variability of kinetic gene expression parameters in microbial cells: Modelling and inference from lineage tree data

## Supplementary material

Aline Marguet<sup>1</sup>, Marc Lavielle<sup>2</sup> and Eugenio Cinquemani<sup>1,\*</sup>

<sup>1</sup>Univ. Grenoble Alpes, Inria, 38000 Grenoble, France

<sup>2</sup>Inria Saclay & Ecole Polytechnique, Palaiseau, France

## Contents

<b>S1 Methodological details</b>	<b>2</b>
S1.1 Computation of sufficient statistics for the ARME identification algorithm . . . . .	2
S1.2 Details on the implementation of the maximization step for the ARME identification algorithm	2
S1.3 Single-cell parameter estimates from an identified ARME model . . . . .	2
S1.4 Computation of confidence intervals for the estimates of $\theta$ . . . . .	3
S1.5 Parametrization of the ARME identification algorithm . . . . .	3
S1.5.1 Estimation of unknown initial conditions for the ancestor cell . . . . .	4
S1.5.2 Combination of the proposal laws in the MCMC approach . . . . .	4
S1.6 Calculation of estimates and confidence intervals for matrix $A$ from standard ME identification	4
<b>S2 Additional simulation results</b>	<b>6</b>
S2.1 Validation for different noise levels and number of generations . . . . .	6
S2.2 Performance gain in inheritance estimation with different noise levels and number of generations	9
<b>S3 Evaluation of alternative experiment designs</b>	<b>11</b>
<b>S4 Estimation performance in presence of intrinsic noise</b>	<b>13</b>
<b>S5 Validation of the ARME predictive power</b>	<b>16</b>
S5.1 Validation on simulated data . . . . .	16
S5.2 Validation on real data . . . . .	18
<b>S6 Additional result on yeast osmotic shock gene expression data</b>	<b>19</b>

---

\*to whom correspondence should be addressed

## S1 Methodological details

### S1.1 Computation of sufficient statistics for the ARME identification algorithm

As in standard ME, we rewrite E and M steps of the SAEM algorithm in a more explicit form using the fact that the considered probability distributions belong to the curved exponential family (Lavielle, 2015, Section 9.2.3). Indeed, we have

$$\begin{aligned} \log(p(Y, \varphi|W, \theta)) &= -\frac{N_{\text{tot}}}{2} \log(2\pi h^2) - \sum_{v \in V} \sum_{j=1}^{n_v} \frac{(y_j^v - Cx^v(t_j^v))^2}{2h^2} \\ &\quad - \frac{N}{2} \log(2\pi|\Omega|) - \frac{1}{2} \sum_{v \in V^*} \eta(\varphi^v, \varphi^{v^-})^T \Omega^{-1} \eta(\varphi^v, \varphi^{v^-}) - \frac{1}{2} \log(2\pi|\Sigma|) - \frac{1}{2} (\varphi^\emptyset - \mu)^T \Sigma^{-1} (\varphi^\emptyset - \mu), \end{aligned}$$

where  $\eta(\varphi^v, \varphi^{v^-}) = \varphi^v - (A\varphi^{v^-} + (I - A)b)$ ,  $V$  denotes the set of individual in the population,  $\emptyset$  is the ancestor of the population,  $V^* = V \setminus \{\emptyset\}$ ,  $N_{\text{tot}} = \sum_{v \in V} n^v$  is the total number of observations. Rearranging the terms, we have

$$\log(p(Y, \varphi|W, \theta)) = -\Psi(\theta) + \langle \tilde{S}(Y, \varphi), \phi(\theta) \rangle,$$

where  $\langle A, B \rangle = \text{Tr}(A^T B)$  if  $A, B \in \mathbb{R}^{d \times d}$  and  $\langle x, y \rangle = x^T y$  if  $x, y \in \mathbb{R}^d$ ,  $\Psi, \phi$  are explicit functions of  $\theta$  and  $\tilde{S}(Y, \varphi)$  are the minimal sufficient statistics given by

$$\begin{aligned} \tilde{S}(Y, \varphi) &= \left[ \sum_{v \in V} \sum_{j=1}^{n_v} (y_j^v - Cx^v(t_j^v))^2, \sum_{v \in V^*} \varphi^v (\varphi^v)^T, \sum_{v \in V^*} \varphi^v (\varphi^{v^-})^T, \sum_{v \in V^*} \varphi^v, \sum_{v \in V^*} \varphi^{v^-} (\varphi^{v^-})^T, \right. \\ &\quad \left. \sum_{v \in V^*} \varphi^{v^-}, \varphi^\emptyset (\varphi^\emptyset)^T, \varphi^\emptyset \right]^T. \end{aligned}$$

Note that the first component of  $\tilde{S}$  depends on  $\varphi^v$  through  $x^v(t_j^v)$  for each cell  $v$ . Using this formalism, the E and M steps given in Section 3.1 in main text reduce to

- E-step: update  $s_{k+1} = s_k + \gamma_k \left( \tilde{S}(Y, \varphi^{(k+1)}, W) - s_k \right)$ ,
- M-step: update  $\hat{\theta}_{k+1} = \text{argmax}_\theta H(s_{k+1}, \theta)$ ,

where  $H(s, \theta) = -\Psi(\theta) + \langle s, \phi(\theta) \rangle$ .

### S1.2 Details on the implementation of the maximization step for the ARME identification algorithm

The M-step (maximization with respect to  $\theta = (A, b, \Omega, h)$  with  $A$  diagonal) is implemented using the package Optim (Mogensen and Riseth, 2018) in Julia with an Interior Point Newton algorithm, under the constraints that  $\text{Diag}(A) \in [0, 1]$  elementwise and  $h \geq 0$ . To ensure that  $\Omega$  is searched in the domain of positive semidefinite matrices, we rather work with its Cholesky factorization  $\Omega = LL^T$  and operate maximization with respect to  $L$ .

### S1.3 Single-cell parameter estimates from an identified ARME model

Given the identified parameters  $\hat{\theta}$ , for any cell  $v$ , random individual cell parameter values can be sampled in accordance with the identified conditional posterior distribution  $p(\varphi^v | \hat{\theta}, Y, W)$  as follows. We run the S-step several times with  $\theta = \hat{\theta}$  and initial values for the individual parameters given by the last values obtained by the ARME identification algorithm. From the stationary chain given by the MH algorithm of the S-step, we draw a sample  $\left( \varphi_{(k)}^v, v \in V \right)_{k \in K}$  of the *a posteriori* conditional distribution of the individual parameters by

taking successive values (Lavielle, 2015, Section 7.3). In particular, from this random sampling procedure, we build the conditional mean estimates of the individual parameters given for each cell  $v$  by

$$\hat{\varphi}^v = \frac{1}{K} \sum_{k=1}^K \varphi_{(k)}^v.$$

## S1.4 Computation of confidence intervals for the estimates of $\theta$

Confidence intervals for the estimates of  $\theta$  are obtained from the Fisher Information matrix (FIM) as explained *e.g.* in Ashyraliyev *et al.* (2009). In the SAEM approach, it is possible to obtain an estimation of the FIM simultaneously with the maximum likelihood estimators of  $\theta$ . Indeed, Fisher's identity and Louis's missing information principle yield respectively

$$\begin{aligned} \partial_{\theta} L(\theta|Y, W) &= \mathbb{E}_{\varphi|Y, W, \theta} [\partial_{\theta} \log(p(Y, \varphi|W, \theta))], \\ \partial_{\theta}^2 L(\theta|Y, W) &= \mathbb{E}_{\varphi|Y, W, \theta} [\partial_{\theta}^2 \log(p(Y, \varphi|W, \theta))] + \text{Cov}_{\varphi|Y, W, \theta} [\partial_{\theta} \log(p(Y, \varphi|W, \theta))] \end{aligned}$$

where  $\partial_{\theta}$  denotes the differential with respect to  $\theta$ . Therefore, using the same stochastic approximation as before on  $\partial_{\theta} L(\theta|Y, W)$  and  $\partial_{\theta}^2 L(\theta|Y, W)$  (Kuhn and Lavielle, 2005, Section 2.2.5), we are able to estimate the observed Fisher Information matrix  $-\partial_{\theta}^2 L(\theta|Y, W)$ .

In order to find confidence intervals for transformations of parameters estimates (such as  $\exp(b)$  or  $\Sigma$ ), we used a random sampling approach. Suppose that we have an estimation of a parameter  $\hat{\alpha}$  and its associated FIM and that we want a confidence interval for  $f(\hat{\alpha})$ , for some function  $f$ . Then, from a gaussian sample with mean  $\hat{\alpha}$  and covariance given by the FIM, we obtain a sample of  $f(\hat{\alpha})$  by simply applying  $f$ . The confidence intervals in Figure 6 in main text are then drawn from the statistics of this sample where the dot corresponds to the mean of the sample and the intervals corresponds to 2 times the standard deviation of the sample.

## S1.5 Parametrization of the ARME identification algorithm

The ARME identification algorithm is implemented as follows:

1. if initial states are missing for some cells in the population, we estimate it using a MH approach. This is described in the next section.
2. The population is initialized with parameters given by

$$A = 0.2\text{I}, b = [-2.4, -0.1, -5], \Omega = 0.5\text{I}, h = 50$$

for the applications to *in silico* experiments and by

$$A = 0.5\text{I}, b = [\log(0.294), \log(0.947), \log(4e - 03)], \Omega = \begin{bmatrix} 0.1 & 0.07 & 0.06 \\ 0.07 & 0.1 & 0.08 \\ 0.06 & 0.08 & 0.1 \end{bmatrix}, h = 500$$

for real data application. The initial individual parameters are sampled according to the hierarchical ARME model.

3. We run the S-step 10 times without updating parameter  $\theta$  (burn-in iterations) in order to obtain reasonable individual parameters from the data. The S-step is detailed in Section S1.5.2
4. We run the sequence of S-step, E-step and M-step 100 times, computing in the meantime the FIM, as explained above.
5. Convergence of SAEM requires that  $\sum_{k \geq 1} \gamma_k = \infty$  and  $\sum_{k \geq 1} \gamma_k^2 < \infty$  (see Delyon *et al.* (1999)). Hence, we set  $\gamma_k = 1$  for  $1 \leq k \leq 80$  and  $\gamma_k = 1/k$  for  $k \geq 80$ .

### S1.5.1 Estimation of unknown initial conditions for the ancestor cell

The initial states of the cells are sometimes unknown, as it is the case for some individual cells in data of yeast osmotic shock response. In that case, we need to estimate those states. The estimation of those initial states is important because it not only determines the beginning of the individual cell protein concentration trajectory but also affects the trajectories of its progeny.

We use a MCMC approach similar to the one used in the S-step of the ARME identification algorithm in order to initialize both the initial conditions and the individual parameters using the data. As initial conditions are needed only for the ancestor of the population, we do not take into account the correlations due to inheritance in this estimation procedure. The proposal distributions that we consider are  $q_2$  and the equivalent of  $q_3$  mentioned in main text on an augmented state space in order to allow an update of initial conditions. We use a uniform proposal law for the initial condition on a compact set chosen according to the fluorescence measurements given in the data.

### S1.5.2 Combination of the proposal laws in the MCMC approach

As described previously, we use three different kernels  $q_1, (q_2^{(v)}, v \in V)$  and  $(q_3^{(v)}, v \in V)$  for the simulation of  $\varphi$  according to the conditional distribution in the S-step. We also use a slight modification of the kernel  $q_3$  denoted by  $q_4$  which is its componentwise version: for  $i = 1, 2, 3$ ,

$$q_{4,i}^{(v)}(\varphi^v, \tilde{\varphi}^v) \propto \exp\left(-\frac{(\varphi_i^v - \tilde{\varphi}_i^v)^2}{2\sigma^2}\right).$$

The mixing of several kernels in MH is standard in ME (Lavielle, 2015, Section 9.3). We choose the following sequence of proposals to constitute one S-step:

1. update of  $\varphi$  at the population level using  $N_1$  times  $q_1$ ,
2. from top (ancestor cell) to bottom (leaves of the genealogical tree), in each generation, for each individual  $v$ ,
  - (a) update  $\varphi^v$  using  $N_2$  times  $q_2^{(v)}$ , then  $N_3$  times  $q_3^{(v)}$ , then  $N_3$  times  $q_4^{(v)}$  on each coordinate  $i$  of  $\varphi^v$  (computation can be carried out in parallel for the different cells  $v$ ),
  - (b) update the protein concentrations (in particular the initial conditions) of all individuals in the next generations.

We choose  $N_1 = N_2 = N_3 = 10$ . This intuitively appealing choice of the order of application of the different kernels has been validated in simulation. The idea is to update the parameters from the coarsest scale (the whole population) to the finest scale (individual components of the parameter vector of one individual).

## S1.6 Calculation of estimates and confidence intervals for matrix $A$ from standard ME identification

We detail here the computation of an estimate of the inheritance from standard ME estimates of the population parameters. The statistic presented in the boxplots of Figure 4 in main text are computed from 20 estimates of the correlation from 20 different datasets. For each dataset, we compute a family of individual-cell estimates  $\hat{\Phi} = \{\varphi^v, v \in V\}$  from a sample of size  $K = 30$ , as explained in Section S1.3. Then, estimates of the diagonal inheritance matrix parameters  $A_{i,i}$  is given for  $i \in \{1, 2, 3\}$  by  $\hat{A}_{i,i} = \hat{\xi}_i \hat{\sigma}_i^{-2}$ , with

$$\hat{\sigma}^2 = \frac{1}{|V|} \sum_{v \in V} (\hat{e}_i^v)^2, \quad \hat{\xi} = \frac{1}{|W|} \sum_{(v^-, v) \in W} \hat{e}_i^v \hat{e}_i^{v^-}, \quad (\text{S1})$$

where  $\hat{e}^v = \varphi^v - \hat{b}$  and  $\hat{b}$  is the empirical mean of the individual-cell parameter estimates  $\hat{\Phi}$ .

We use the same approach on real data, replacing the different datasets by more samples. More precisely, we pick  $n = 20$  samples of size  $K = 30$  and from (S1), we obtain  $n$  estimates  $\{(\hat{A}_{i,i}^{(\ell)})_{1 \leq i \leq 3}, 1 \leq \ell \leq n\}$  of

the diagonal inheritance matrix  $A$  from individual-cell estimates. Then, in Figure 6 in main text, for the estimation of  $A$  in orange, the dot corresponds to the mean of the sample and the intervals corresponds to 2 times the standard deviation of the sample.

## S2 Additional simulation results

### S2.1 Validation for different noise levels and number of generations

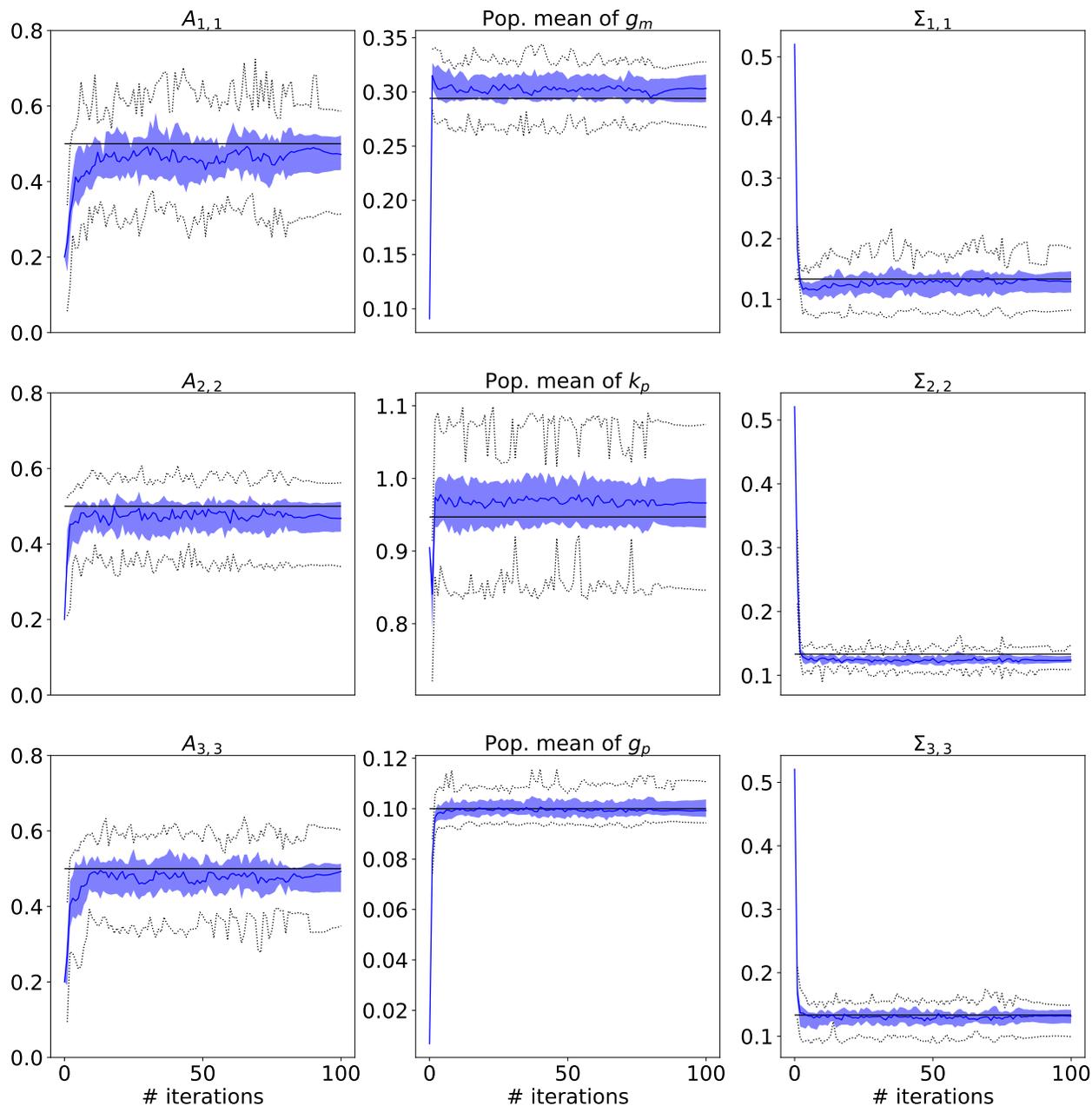


Figure S1: Iterative ARME identification of parameters  $\text{Diag}(A) = (A_{1,1}, A_{2,2}, A_{3,3})$ ,  $b$ ,  $\text{Diag}(\Sigma) = (\Sigma_{1,1}, \Sigma_{2,2}, \Sigma_{3,3})$  from the application of generalized SAEM to 20 simulated datasets  $Y$  for 80 search iterations plus 20 stabilizing iterations (100 iterations total). Identification is based on data simulated over 7 generations with measurement noise level  $h = 10$ . Horizontal black lines: true parameters; Blue lines: median of the iterative estimation profiles; Shaded blue region, and dashed lines: at every iteration, 25% and 75% quantiles of the estimates over the 20 datasets, and extension of corresponding whiskers, as computed in all boxplots of main text.

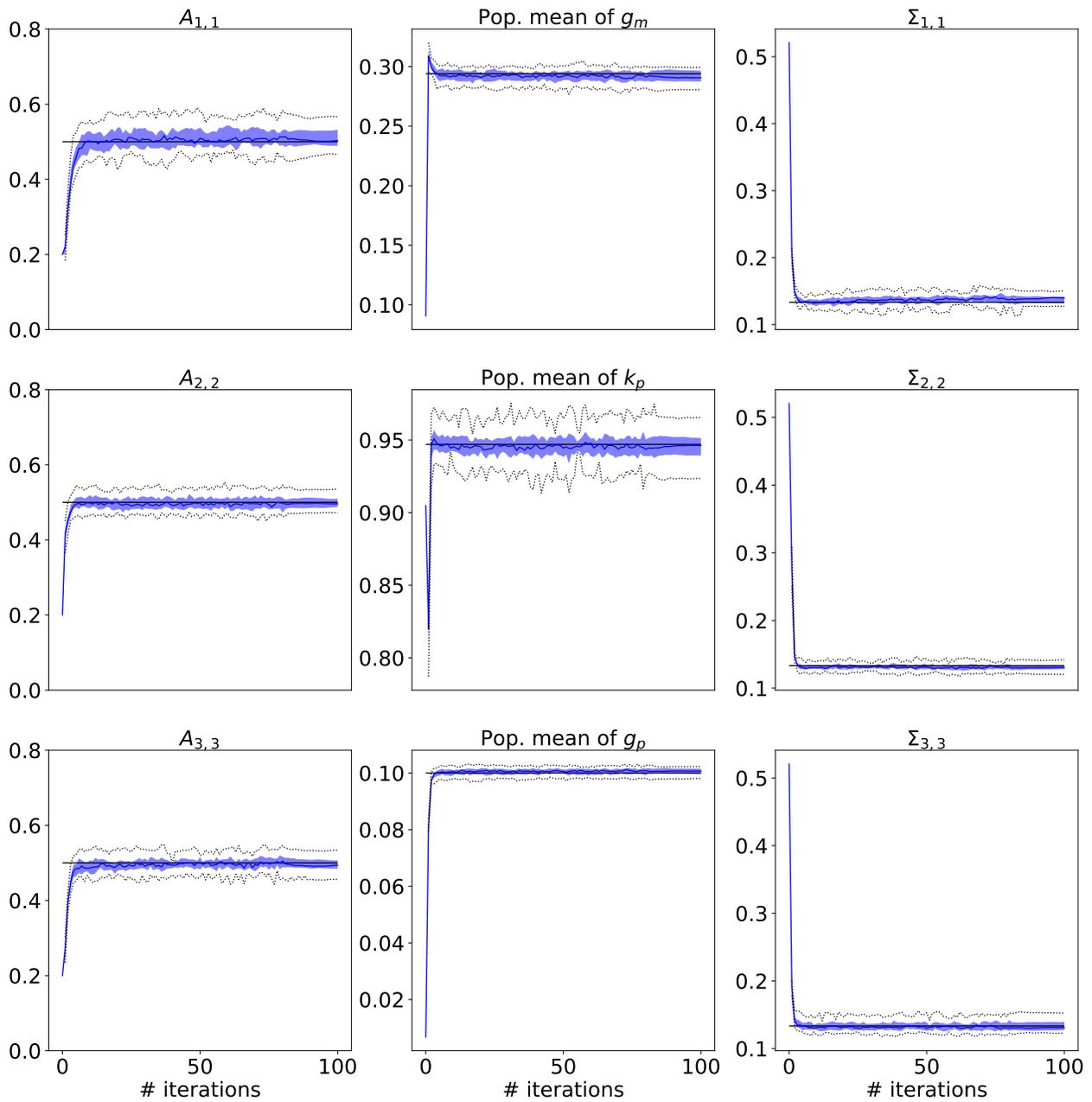


Figure S2: Same as for Fig. S1 but with identification based on data simulated over 11 generations with measurement noise level  $h = 20$ .

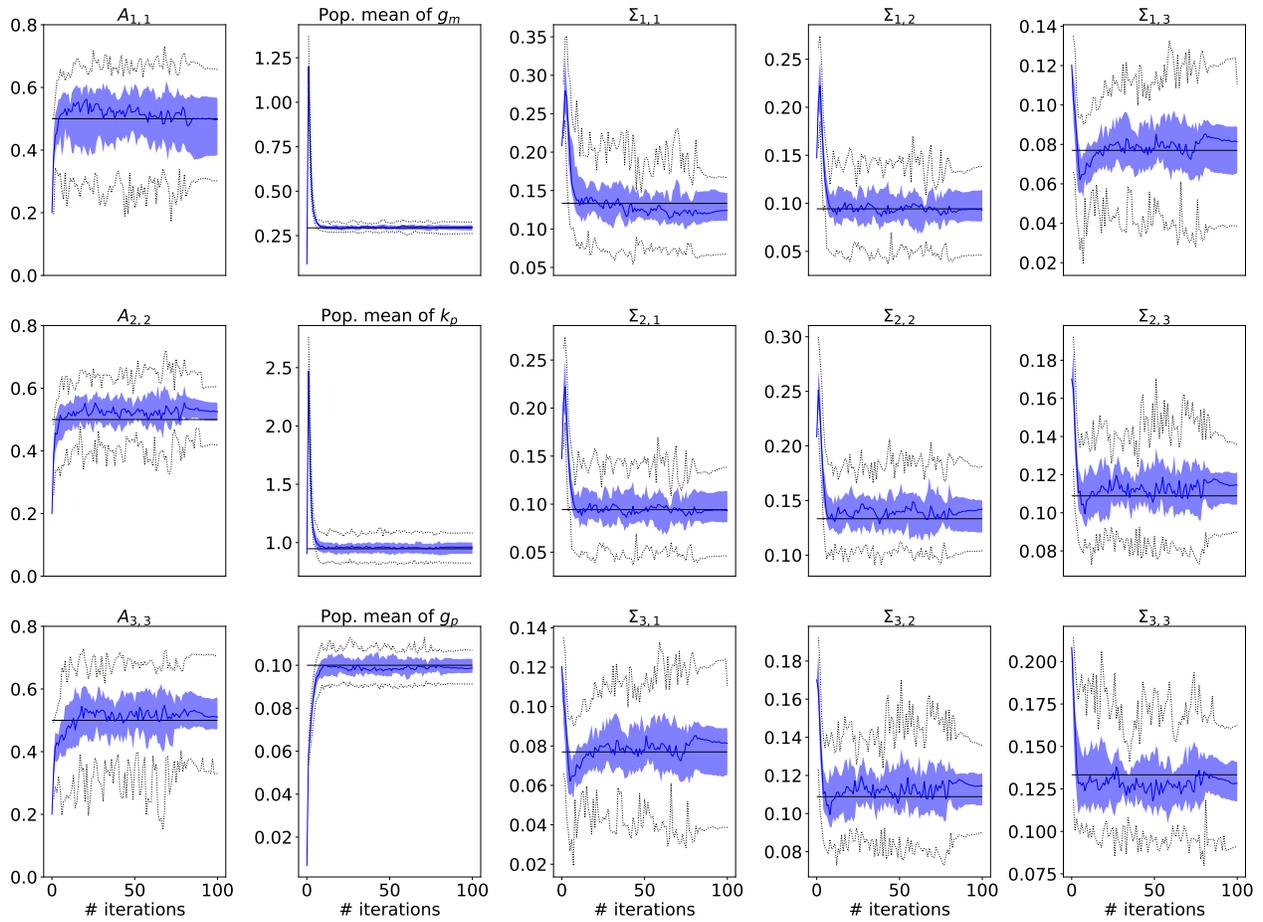


Figure S3: Same as for Fig. S1, but with  $h = 20$  and a full (nondiagonal) matrix  $\Sigma$ .

## S2.2 Performance gain in inheritance estimation with different noise levels and number of generations

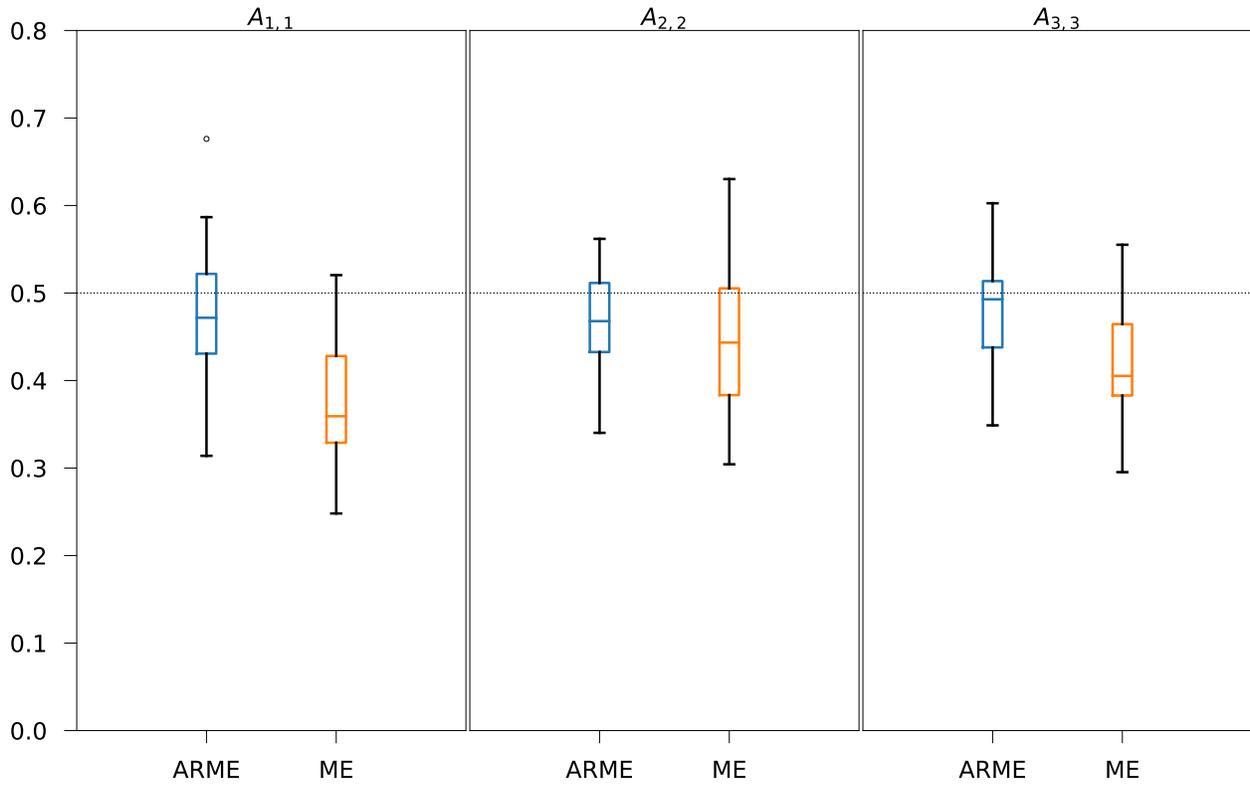


Figure S4: Statistics of identification of inheritance parameters  $\text{Diag}(A) = (A_{1,1}, A_{2,2}, A_{3,3})$  over 20 simulated datasets. Identification is based on data simulated over 7 generations with measurement noise level  $h = 10$ . For each parameter, we compare the results from ARME identification and from indirect estimation based on ME.

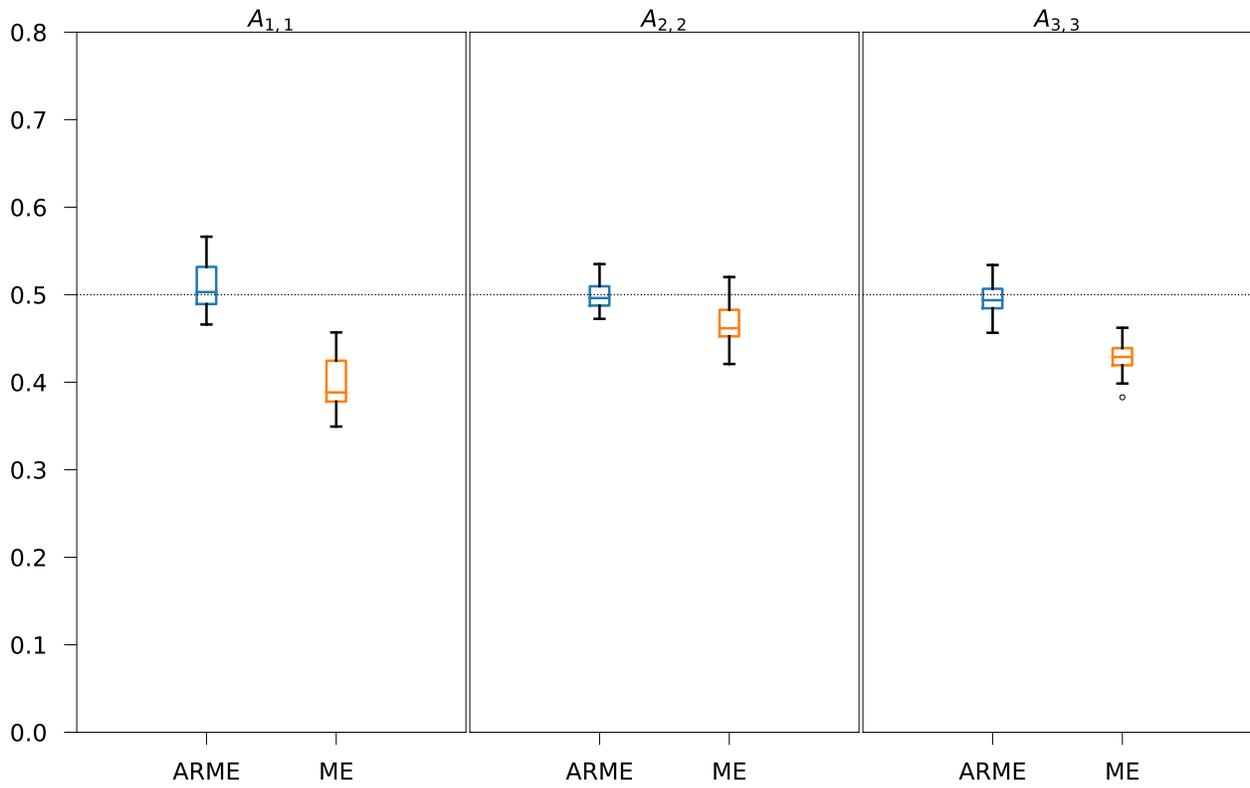


Figure S5: Same as for Fig. S4 but with identification based on data simulated over 11 generations with measurement noise level  $h = 20$ .

### S3 Evaluation of alternative experiment designs

In view of the results of Sec. 4 in main text, whereby estimation of  $A$  is intimately related with exploitation of mother-daughter relationships, a natural question is how estimation performance depends on the structure of  $W$ . When using indirect methods based on standard ME estimates,  $|W|$  (the number of mother-daughter pairs in  $W$ ) clearly determines the accuracy of estimates (see *e.g.* Eq. (7) in main text). This prompts us to examine how the structure of  $W$  influences estimation of  $A$  (more generally, of  $\theta$ ) in the ARME approach.

We look at three cases, reflecting different possible experimental configurations. In Case I, we look at a full tree of 7 generations from a single cell at generation 1 (as per full monitoring on a growing colony *e.g.* on a plate). In Case II, starting from a single cell, only one daughter is kept track of at every division over 127 generations (a single branch of a longer tree is observed, as *e.g.* by so-called mother machines (Izard *et al.*, 2015)). For both I and II,  $|V| = 127$  and  $|W| = 126$ . Finally, in Case III, we consider observation of  $N = 42$  mutually independent cherry-like triplets with one mother related with two daughter cells. Here  $|V| = 126$  and  $|W| = |V| - N = 84$ . Observe that  $|V|$  is essentially the same in all cases. For every case, we simulated 20 datasets with the same parameters  $\theta$  as in Sec. 4.1 of the main text, and ran ARME identification on each dataset.

Estimation statistics for Case I, II and III are compared in Fig. S6 below. In all cases I, II and III and for both  $A$  and  $b$ , estimation is essentially unbiased (results for  $\Omega$  are not shown since essentially identical in all cases). Performance is very similar in Case I and II also in terms of estimation uncertainty. This suggests that estimation performance depends primarily on  $|W|$ . Indeed, different results are obtained in Case III. Compared to I and II, larger uncertainties are observed in the estimation of the inheritance factor  $A$ , whereas smaller uncertainties are observed for the estimation of the mean parameter vector  $b$ . We explain this as follows. The decrease of  $|W|$  from 126 (Case I and II) to 84 (Case III) worsens the accuracy of  $\hat{A}$ , in accordance with the fact that  $A$  is intimately related with mother-daughter cell dependencies. On the other hand  $b$  is not related with these dependencies since it pertains individual cell statistics. Rather, the decrease of  $|W|$  from Case I and II to Case III improves estimation performance of  $b$ , since it increases the statistical diversity of the sample itself. This is analogous to the fact that precision in the estimation of a population mean increases with the number of independent individuals observed. Note that the observed differences are not marginal in view of the fact that the change in  $|W|$  is of only about 30%. In sums, from the experimental viewpoint, this says that the observation pattern to be sought in a population of cells depends on whether focus is on the estimation of inheritance gene expression parameters or on their distribution in the population.

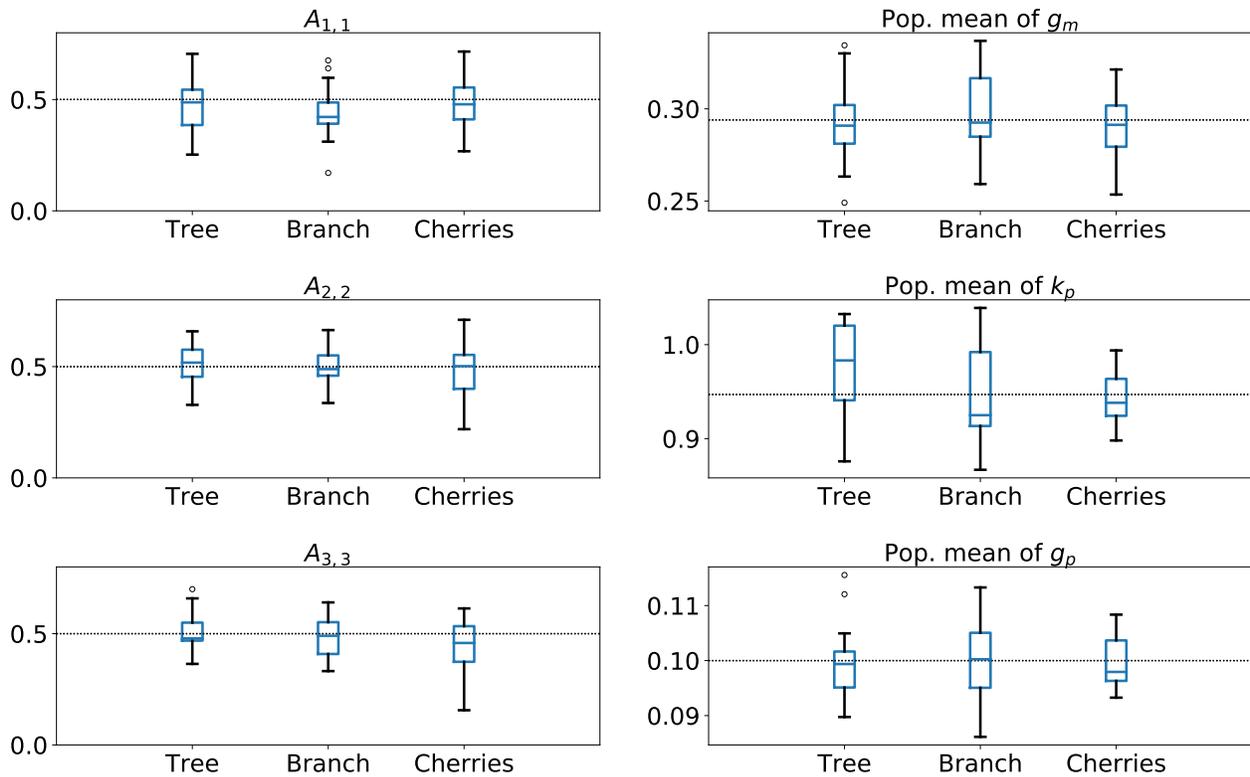
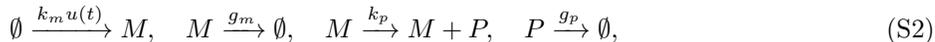


Figure S6: Statistics of ARME identification of inheritance parameters  $\text{Diag}(A) = (A_{1,1}, A_{2,2}, A_{3,3})$  and mean parameters  $(\exp(b_1), \exp(b_2), \exp(b_3))$  over 20 simulated datasets. Simulated noise level is  $h = 20$ . For each parameter, compared are the results for the cases of a full a population tree of 127 cells, of a single branch of 127 cells, and of 42 groups of cherry-like trees of 3 cells each (respectively, Case I, II and III in the text).

## S4 Estimation performance in presence of intrinsic noise

Although our modelling and inference methods do not take into account intrinsic noise, in this section we investigate how they perform in terms of accuracy of estimation if the observed protein dynamics are affected by different levels of intrinsic noise.

We simulated new datasets with gene expression dynamics given by the Gillespie-type stochastic implementation of the biochemical reactions



where  $u(t)$  is the strength of promoter activation at time  $t$ . The mean value of the resulting process is given by the solution of the differential equation system (1) in the main text. We consider two different levels of expression by choosing different values for  $k_m$  and  $k_p$ . More precisely, we set  $(k_m, k_p) = (10, 0.947)$ , which corresponds to strong mRNA transcription, or  $(k_m, k_p) = (1, 9.47)$ , for more moderate mRNA transcription. Notice that fixing the product  $k_m k_p$  to the same value yields the same mean amount of proteins. However, the reduced number of mRNA molecules in the first case makes intrinsic noise effects more pronounced. In Fig. S7, we show the difference in noise level for our two choices of parameters.

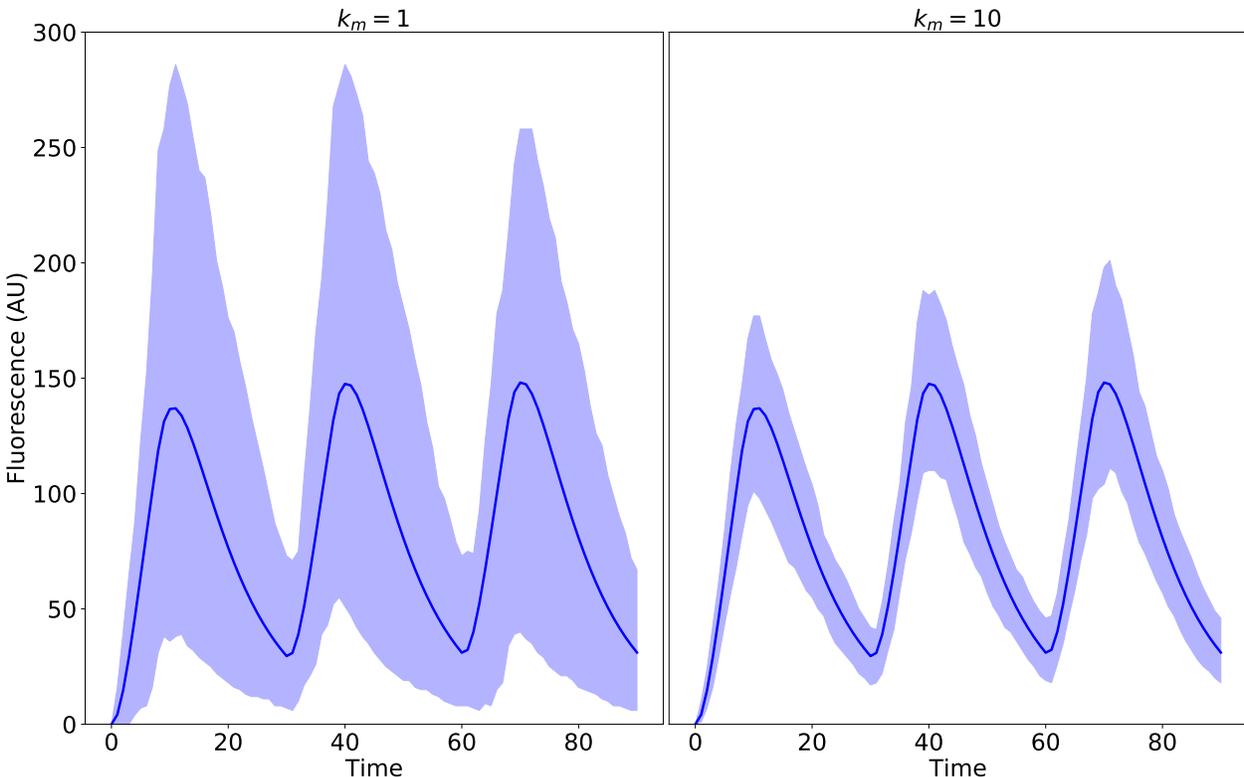


Figure S7: Comparison of the noise levels for the stochastic biochemical process corresponding to Eq. (S2) for two different parameter choices (left:  $(k_m, k_p) = (1, 9.47)$ ; right:  $(k_m, k_p) = (10, 0.947)$ ). Blue curves represent the mean profile of the fluorescent reporter protein abundance and are equal in the two cases. Shaded areas represent the 95% confidence bands of the protein abundance process.

For both parameter choices, we simulated 20 datasets from populations of 127 cells generated in accordance with Section 4 of the main text. That is, starting from a single uninduced cell at time 0 (generation 1), we simulate division of every existing cell into two daughter cells every 90 minutes over 7 generations, thus obtaining a full cell tree. Single-cell parameters are simulated on the basis of model (2) in the main text. Gene expression dynamics of every cell are simulated in accordance with the stochastic model (S2) instead of deterministic model (1) in the main text. The following table summarizes the parameter values used in the two cases (non-identical parameters are emphasized in bold).

Strong mRNA transcription	Moderate mRNA transcription
$\mathbf{k}_m = \mathbf{1}$ $A = \text{Diag}(0.5, 0.5, 0.5)$ $b = [\log(0.294), \log(\mathbf{9.47}), \log(0.1)]^T$ $\Omega = \text{Diag}(0.1, 0.1, 0.1)$	$\mathbf{k}_m = \mathbf{10}$ $A = \text{Diag}(0.5, 0.5, 0.5)$ $b = [\log(0.294), \log(\mathbf{0.947}), \log(0.1)]^T$ $\Omega = \text{Diag}(0.1, 0.1, 0.1)$

Next, we used the ARME identification method (assuming the deterministic gene expression dynamics (1) in main text) on the two collections of 20 datasets. We focus on the performance in the estimation of the inheritance factor  $A$ . These are displayed in Fig. S8. In the case of small intrinsic noise (right column), our identification method still provides good results, although they are less precise than the results obtained in Fig. 3 in the main text. This is because the model of the gene expression dynamics used in identification is different from the data-generating process. As expected, we observe that the precision of the identification results deteriorates in the case of larger intrinsic noise (left column). In conclusion, our method is robust to small amounts of intrinsic noise, while performance deteriorates in proportion to the intrinsic noise level.

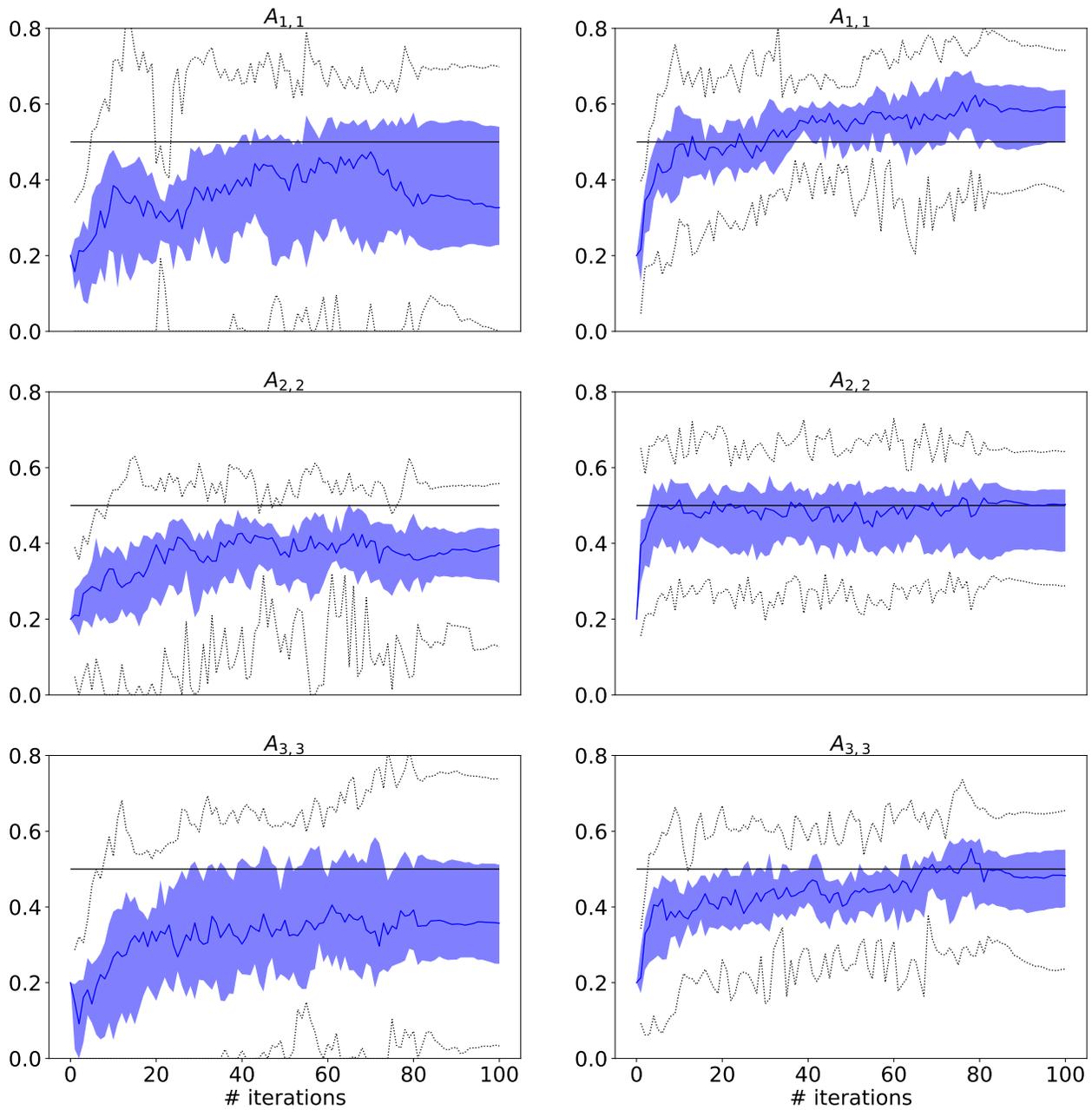


Figure S8: Each column is the same as the first column for Fig. 3 (main text). First column: lowly expressed gene datasets, second column: highly expressed gene datasets.

## S5 Validation of the ARME predictive power

### S5.1 Validation on simulated data

To assess the predictive power of the ARME inheritance model and compare it with that of a model inferred by ME (that is, neglecting inheritance), we generated 20 new validation datasets (full trees with 7 generations) using the same parameters as in Section 4.1 in the main text, that is

$$A = \text{Diag}(0.5, 0.5, 0.5), \quad b = [\log(0.294), \log(0.947), \log(0.1)]^T, \quad \Omega = \text{Diag}(0.1, 0.1, 0.1).$$

Then, on the basis of the parameters  $\hat{\theta}^{\text{ARME}} = (\hat{A}, \hat{b}, \hat{\Omega})$  inferred on the original identification dataset, we computed an estimation of the individual cell parameters (posterior mean) for each validation dataset using the method described in S1.3. For each dataset, we computed the normalized distance

$$d_i^{\text{ARME}} := \sqrt{\frac{1}{b_i^2} \sum_{v \in V} (\hat{\varphi}_i^{v, \text{ARME}} - \varphi_i^v)^2}$$

for  $i = 1, 2, 3$ , corresponding to parameters  $g_m, k_p$  and  $g_p$ , in the same order. Repeating the same protocol but using parameters  $\hat{\theta}^{\text{ME}}$  inferred with standard ME (from the same identification dataset) for the estimation of the individual cell parameters  $\{\hat{\varphi}_i^{v, \text{ME}}\}$  (on the same validation datasets), we obtain the corresponding values  $d_i^{\text{ME}}$ , with  $i = 1, 2, 3$ , for standard ME.

In the top plots of Fig. S9, for each individual cell parameter, we show the statistics of  $d^{\text{ARME}}$  obtained with the 20 validation datasets. The dotted line is the corresponding value of  $d^{\text{ARME}}$  obtained on the identification dataset. The individual cell parameter prediction errors  $d^{\text{ARME}}$  are equivalent to those observed on the identification dataset, reconfirming the absence of overfit of the identification data.

In the bottom plots of Fig. S9, for each individual cell parameter, we show the statistics of the difference  $d^{\text{ME}} - d^{\text{ARME}}$  obtained with the 20 validation datasets. The dotted line is the baseline value 0 which would correspond to an equal predictive power of the two methods tested. The positive values of the difference (specifically, for  $g_m, k_p$  and  $g_p$ , the difference is positive on 18, 15 and 18 validation datasets, respectively) witnesses the better prediction power of the model identified by ARME versus the model identified by the ME-based method.

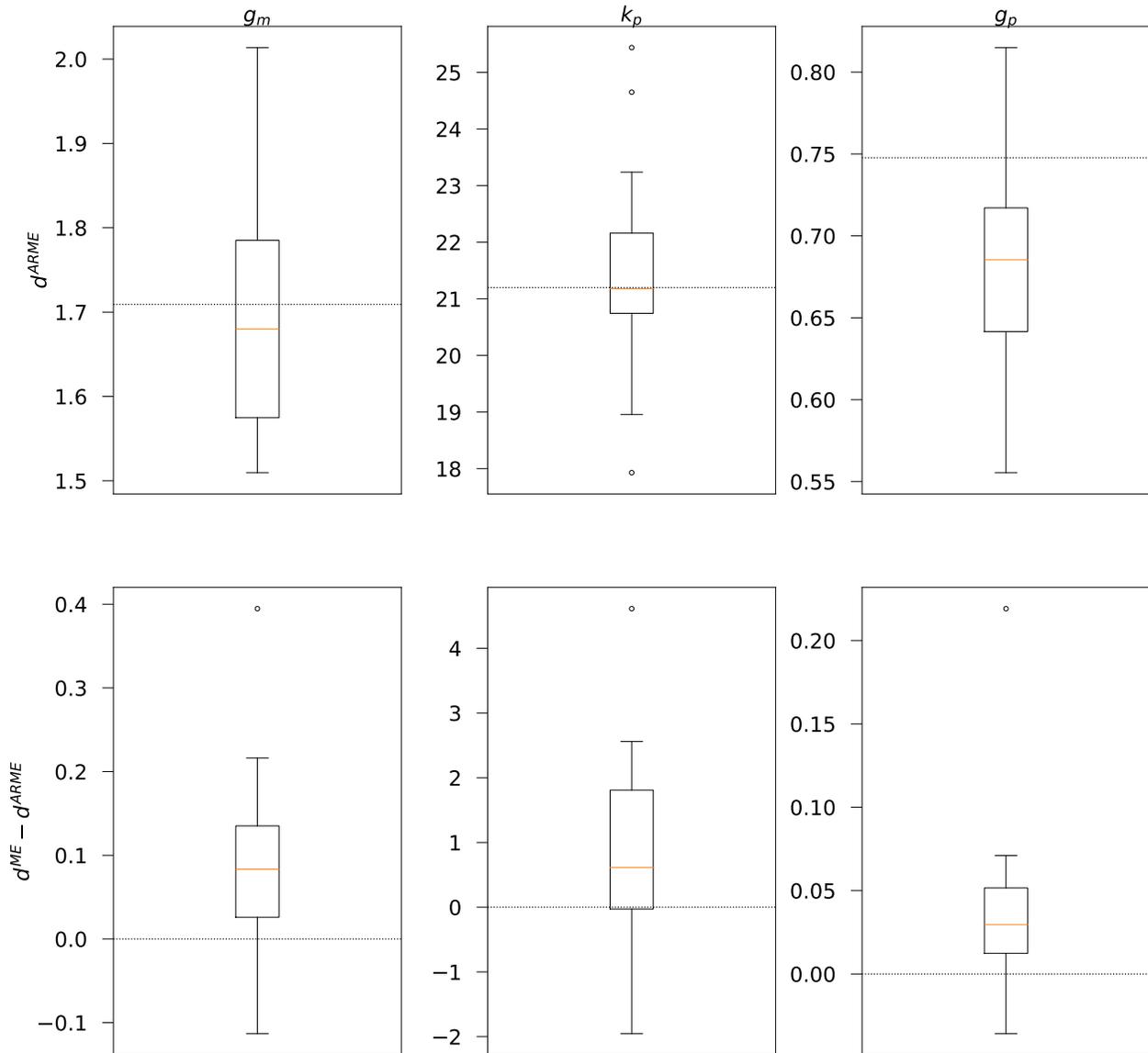


Figure S9: First row: Boxplots for the values of  $d_i^{\text{ARME}}$  obtained on 20 validation datasets. The dotted lines correspond to the values of the  $d_i^{\text{ARME}}$  obtained on the dataset used for inference. Second row: Boxplots of the differences  $d_i^{\text{ME}} - d_i^{\text{ARME}}$  for 20 validation datasets. From left to right:  $i = 1, 2, 3$ , corresponding to the statistics for  $g_m$ ,  $k_p$  and  $g_p$ , in the same order.

## S5.2 Validation on real data

In order to test the predictive power of our method on real data, where true individual cell parameter values are unknown, we rely on the predicted statistics of fluorescence distribution over the cells available at different experimental times. We split the experimental dataset used for inference in Sec. 5 (main text) in two parts:

- a reduced identification dataset with 76 individuals, composed of 15 pairs of one mother generating one daughter cells, 10 triplets of one mother generating two daughter cells and 3 quadruples of one mother generating three daughter cells,
- a validation dataset with 10 individuals divided in 2 triplets of one mother generating two daughter cells and 1 quadruple of one mother generating three daughter cells.

Granted a suitable proportion between identification and validation cells, this partitioning was determined at random. We refer the reader to Fig. 5 in the main text for an illustration of the different tree formats. The results of the identification of the ARME model parameters using the reduced identification dataset are close to the results obtained for the whole dataset (see Fig. 6 in the main text) and are given by:

$$\hat{A} = \text{Diag}(0.678, 0.608, 0.467), \quad \exp(\hat{b}) = [0.012, 1.266, 0.010], \quad \hat{\Sigma} = \begin{bmatrix} 0.354 & 0.138 & -0.169 \\ 0.138 & 0.385 & 0.183 \\ -0.169 & 0.183 & 0.287 \end{bmatrix}.$$

Using these parameter estimates, we predicted the dynamics of proteins for the subpopulations in the validation dataset. More precisely, given the lineage structure  $W_V$  of the validation dataset and the corresponding cell division times (which differ from the identification dataset), we drew 100 realisations of the individual parameters according to  $p(\cdot|\hat{\theta}, W_V)$ , where  $\hat{\theta} = (\hat{A}, \hat{b}, \hat{\Omega})$ , and simulated corresponding fluorescence level profiles. For comparison, we did the same for the reduced dataset on which identification was performed.

Fig. S10 reports the 95% confidence intervals (computed over the 100 simulations) for the dynamics of the population, by considering at each timestep the fluorescence reporter protein concentration of all individuals alive at the corresponding time. The confidence intervals are similar for the dataset used for inference and the validation dataset despite the different lineage structure (see Figure S10). Moreover, they agree with the identification and the validation data equally well. This reconfirms the absence of overfitting of the identification dataset and the equally good performance in predicting the statistics of a new population.

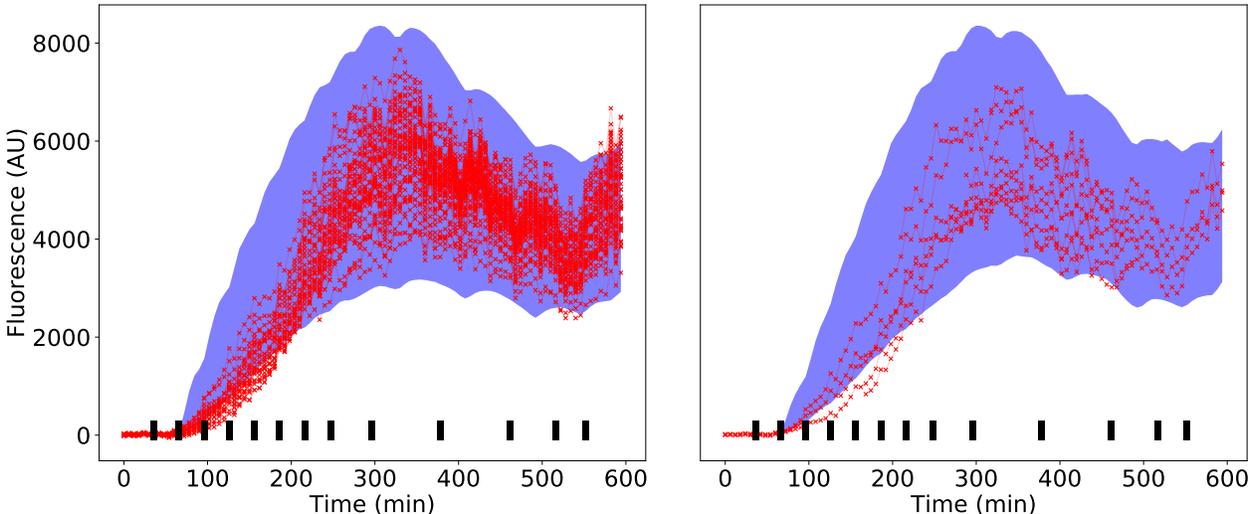


Figure S10: Confidence bands at 95% level (blue areas) and data (red crosses and connecting lines) for the reduced identification dataset (left) and the validation dataset (right).

## S6 Additional result on yeast osmotic shock gene expression data

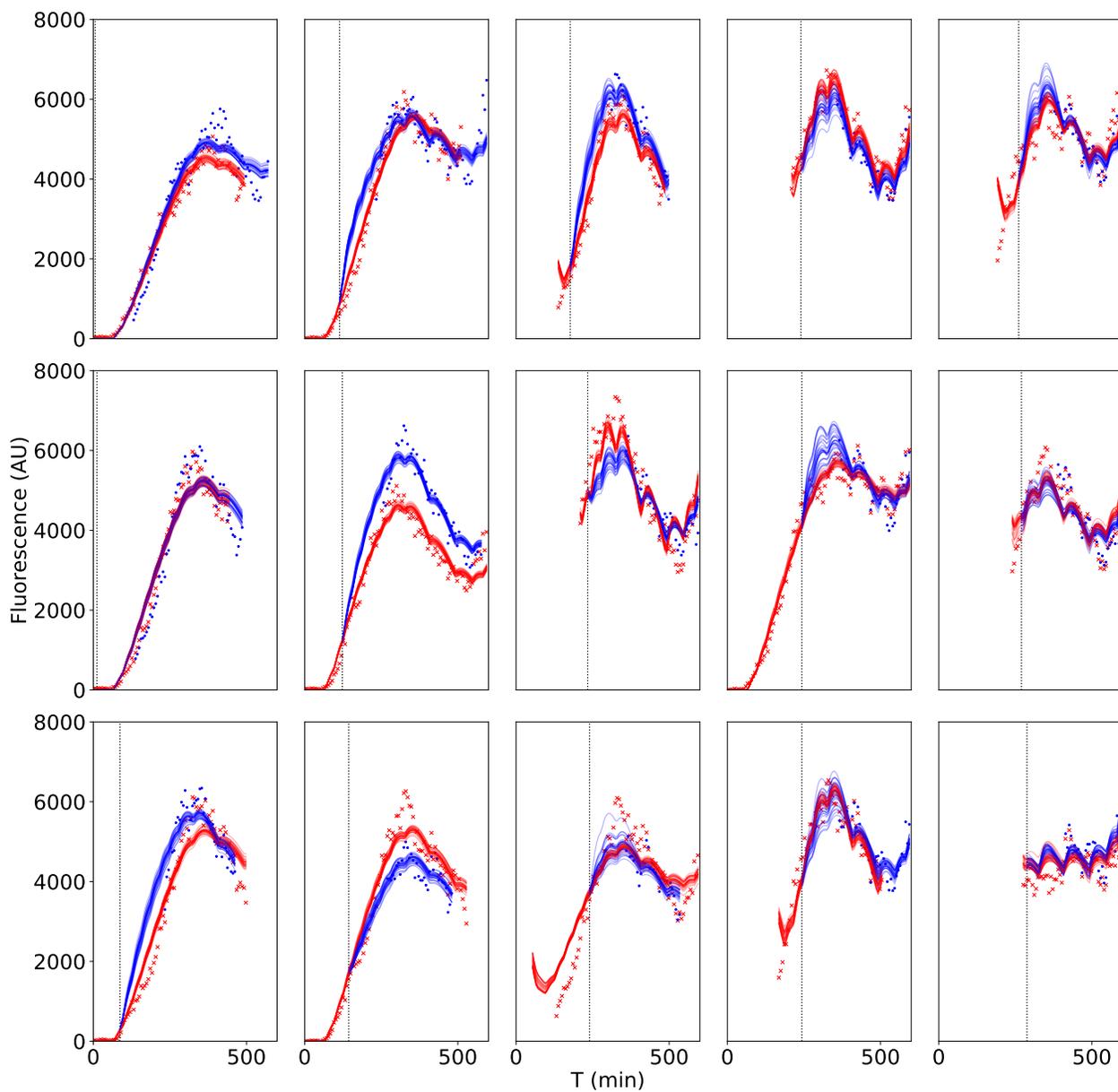


Figure S11: Single-cell fits to real data for the identified model, as in Fig. 6 of main text, for all observed pairs of one mother and one daughter cell.

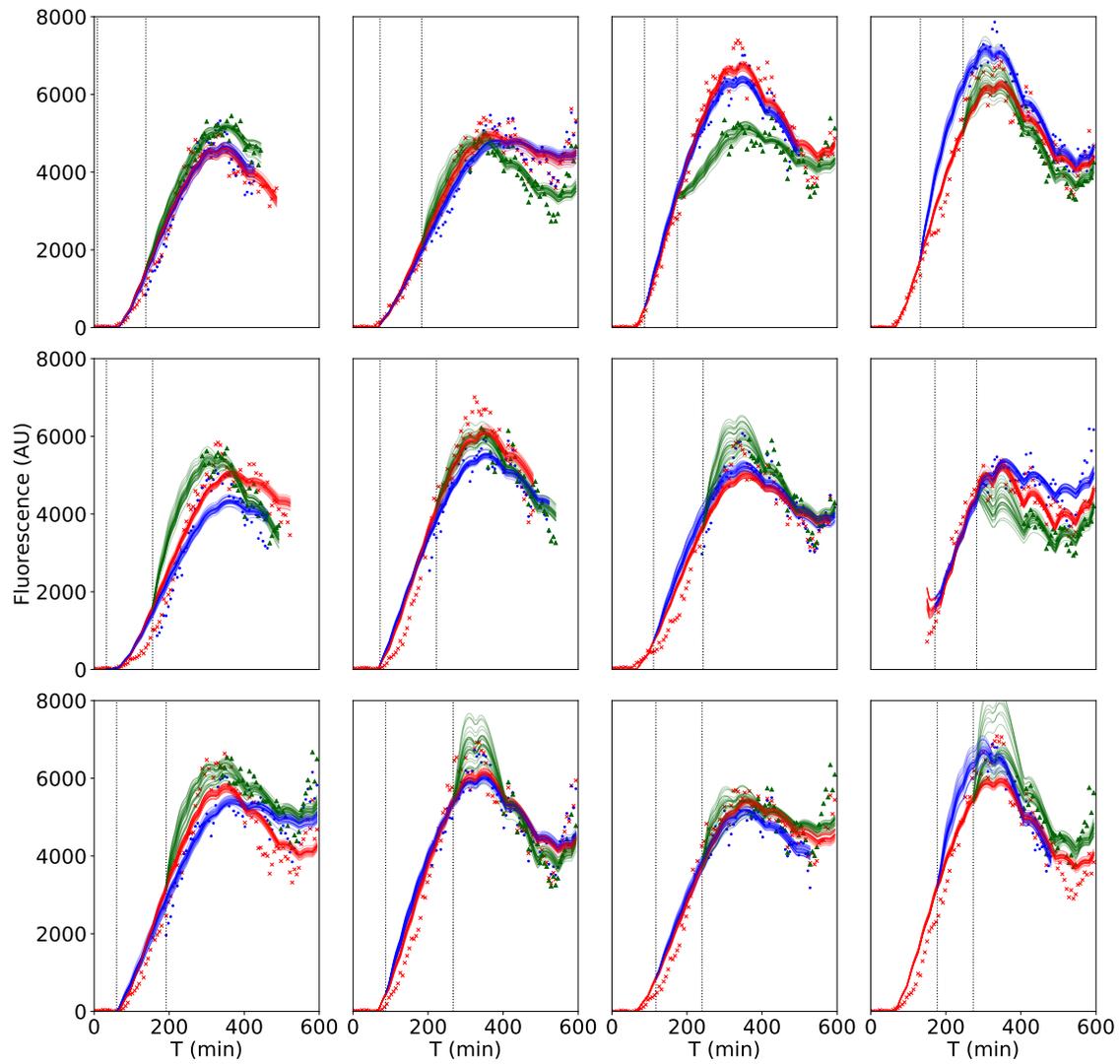


Figure S12: Single-cell fits to real data for the identified model, as in Fig. 6 of main text, for all observed triplets of one mother and two daughter cells.

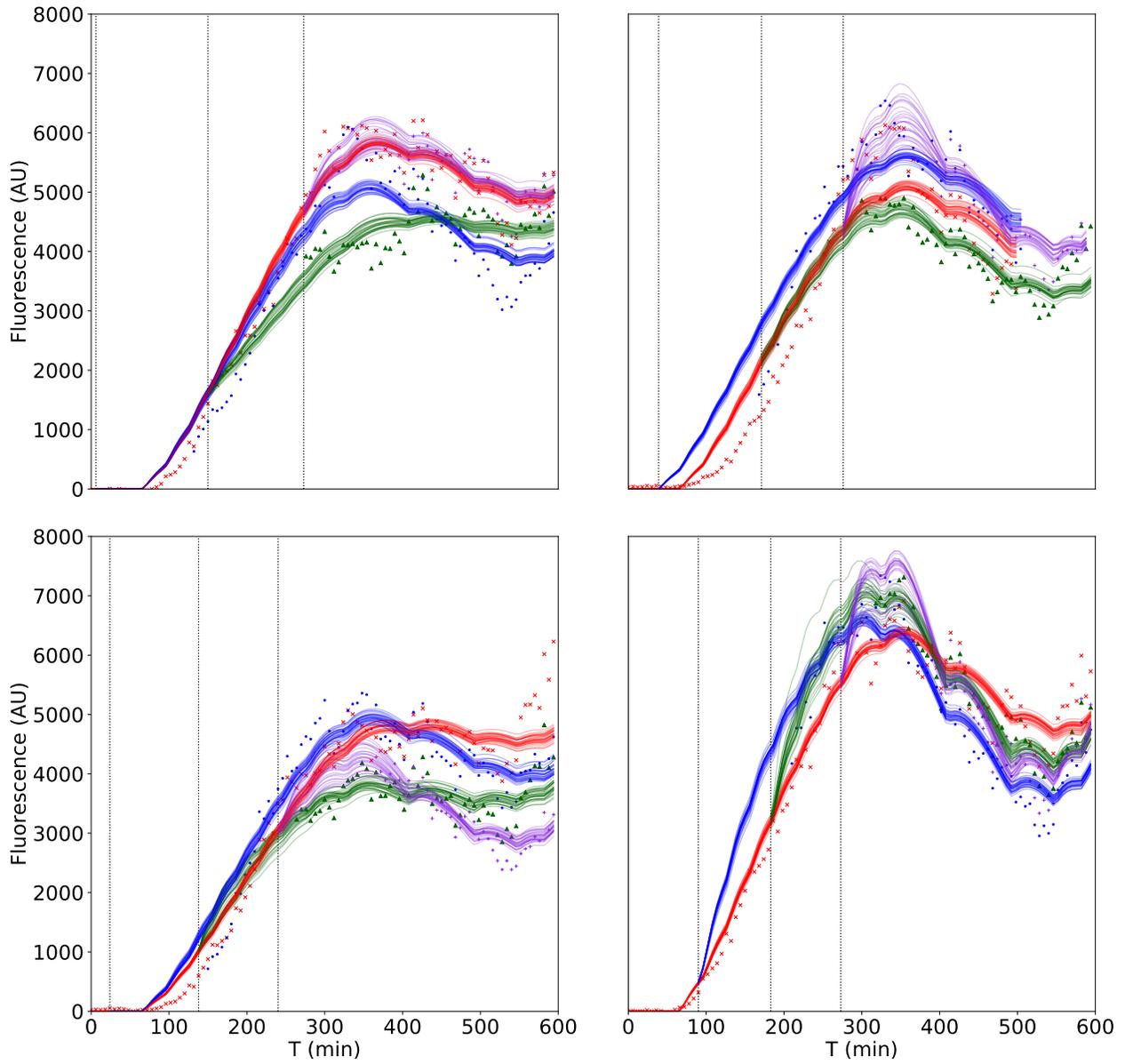


Figure S13: Single-cell fits to real data for the identified model, as in Fig. 6 of main text, for all independent quadruplets of one mother and three daughter cells.

## References

- Ashyraliyev, M., Fomekong-Nanfack, Y., Kaandorp, J., and Blom, J. (2009). Systems biology: Parameter estimation for biochemical models. *FEBS J.*, **276**, 886–902.
- Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.*, **27**(1), 94–128.
- Izard, J., Gomez Balderas, C. D., Ropers, D., Lacour, S., Song, X., Yang, Y., Lindner, A. B., Geiselmann, J., and de Jong, H. (2015). A synthetic growth switch based on controlled expression of rna polymerase. *Mol Syst Biol.*, **11**(11).
- Kuhn, E. and Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Comput. Statist. Data Anal.*, **49**(4), 1020–1038.
- Lavielle, M. (2015). *Mixed effects models for the population approach. Models, Tasks, Methods & Tools*. Chapman & Hall/CRC Biostatistics Series. CRC press.
- Mogensen, P. K. and Riseth, A. N. (2018). Optim: A mathematical optimization package for Julia. *J. Open Source Softw.*, **3**(24), 615.