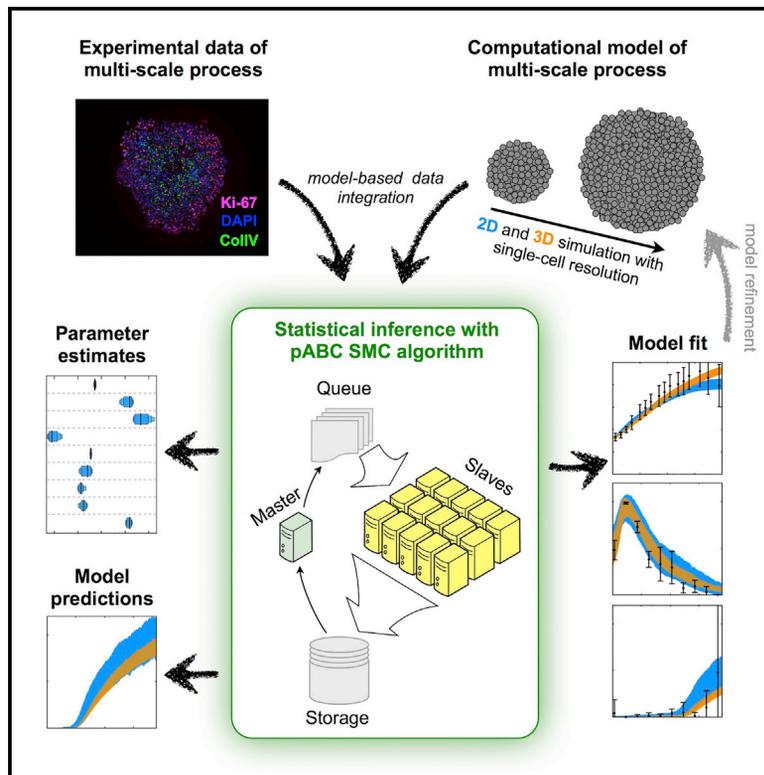


# Cell Systems

## Parallelization and High-Performance Computing Enables Automated Statistical Inference of Multi-scale Models

### Graphical Abstract



### Authors

Nick Jagiella, Dennis Rickert,  
Fabian J. Theis, Jan Hasenauer

### Correspondence

jan.hasenauer@helmholtz-muenchen.de

### In Brief

A new parallel approximate Bayesian computation sequential Monte Carlo (pABC SMC) algorithm allows for robust, data-driven modeling of multi-scale biological systems and demonstrates the feasibility of multi-scale model parameterization through statistical inference.

### Highlights

- Statistical inference for multi-scale models using high-performance computing
- Parallel implementation of the ABC SMC algorithm
- Study of tumor spheroid growth in droplets using growth curves and histological data
- Proof of principle for fitting of mechanistic model with  $10^6$  single cells

# Parallelization and High-Performance Computing Enables Automated Statistical Inference of Multi-scale Models

Nick Jagiella,<sup>1</sup> Dennis Rickert,<sup>1</sup> Fabian J. Theis,<sup>1,2</sup> and Jan Hasenauer<sup>1,2,3,\*</sup>

<sup>1</sup>Institute of Computational Biology, Helmholtz Zentrum München, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany

<sup>2</sup>Chair of Mathematical Modeling of Biological Systems, Center for Mathematics, Technische Universität München, Boltzmannstraße 3, 85748 Garching, Germany

<sup>3</sup>Lead Contact

\*Correspondence: [jan.hasenauer@helmholtz-muenchen.de](mailto:jan.hasenauer@helmholtz-muenchen.de)

<http://dx.doi.org/10.1016/j.cels.2016.12.002>

## SUMMARY

Mechanistic understanding of multi-scale biological processes, such as cell proliferation in a changing biological tissue, is readily facilitated by computational models. While tools exist to construct and simulate multi-scale models, the statistical inference of the unknown model parameters remains an open problem. Here, we present and benchmark a parallel approximate Bayesian computation sequential Monte Carlo (pABC SMC) algorithm, tailored for high-performance computing clusters. pABC SMC is fully automated and returns reliable parameter estimates and confidence intervals. By running the pABC SMC algorithm for  $\sim 10^6$  hr, we parameterize multi-scale models that accurately describe quantitative growth curves and histological data obtained in vivo from individual tumor spheroid growth in media droplets. The models capture the hybrid deterministic-stochastic behaviors of  $10^5$ – $10^6$  of cells growing in a 3D dynamically changing nutrient environment. The pABC SMC algorithm reliably converges to a consistent set of parameters. Our study demonstrates a proof of principle for robust, data-driven modeling of multi-scale biological systems and the feasibility of multi-scale model parameterization through statistical inference.

## INTRODUCTION

Systems and computational biology aims at a mechanistic understanding of complex biological behavior. To achieve this, biological processes on a wide range of time and length scales have to be captured (Hunter and Borg, 2003). To integrate these diverse data into a coherent view of how biological systems may work, multi-scale models of biological processes are needed. Interdisciplinary initiatives have been formed to develop multi-scale models and modeling approaches for basic research, diagnosis, and therapy (see Hunter and Borg, 2003; Karr et al., 2012; Noble, 2002; Tomita et al., 1999; Trayanova, 2011; and ref-

erences therein). Platforms for multi-scale modeling of individual cells (Schaff et al., 1997; Stiles and Bartol, 2001), tissues (Richmond et al., 2010; Starruß et al., 2014; Swat et al., 2012), and organs (Mirams et al., 2013) have also been implemented and popularized. These technological advances have resulted in a tremendous increase of the availability and popularity of multi-scale models. However, one problem remains largely unsolved: how can these models be parameterized in a consistent and rigorous way? Most model parameters cannot be measured directly. To enable truly quantitative predictions, the parameters of multi-scale models have to be inferred from experimental data.

For deterministic multi-scale models obtained by coupling ordinary differential equations (ODEs) and partial differential equations (PDEs), promising successes have been achieved. For example, an integrated, physiologically based, whole-body model of the glucose-insulin-glucagon regulatory system has been developed and parameterized in an automated way for individual patients to improve the understanding of type 1 diabetes (Schaller et al., 2013). Similarly, whole-heart models could be used to infer ischemic regions from body surface potential maps to provide an early diagnosis of heart infarction (Nielsen et al., 2013). These and other applications demonstrate that the automated parameterization of multi-scale models from experimental data using parameter estimation methods is feasible. However, parameter estimation is mostly limited to deterministic multi-scale models because they allow for efficient, gradient-based optimization. In gradient-based optimization, the local change of the likelihood function—a statistical measure for the goodness of fit—is evaluated to determine the direction in parameter space in which the fit improves most rapidly. This facilitates substantial improvements of the fit within a few iterations of the optimizer and frequently produces a good model with limited computational effort.

The parameterization of computationally demanding stochastic and hybrid stochastic-deterministic models is more challenging (Adra et al., 2011; Karr et al., 2015). However, to understand biological processes on the smaller scale, stochastic, and hybrid multi-scale models have to be considered (Dada and Mendes, 2011; Hasenauer et al., 2015; Walpole et al., 2013). Molecular processes such as gene expression (Eldar and Elowitz, 2010; Elowitz et al., 2002) and signal transduction (Klann et al., 2009; Niepel et al., 2009) are partially

stochastic, influencing cell division (Huh and Paulsson, 2011) and cell movement (Anderson and Quaranta, 2008; Graner and Glazier, 1992). The stochasticity of processes like these presents two key challenges to the analysis and parameterization. First, the simulation of stochastic models is often computationally demanding, especially when compared to similar deterministic models. Second, for stochastic models, the likelihood function and its gradients cannot be assessed in closed form.

To see these challenges in action, consider the sophisticated agent-based models of liver regeneration (Hoehme et al., 2010) and tumor growth (Anderson and Quaranta, 2008; Jagiella, 2012). These agent-based models provide hybrid stochastic-deterministic descriptions of the biological processes, and a single stochastic simulation takes days to months. To assess the average behavior of models, many such stochastic simulations are necessary. Even worse, the rigorous evaluation of the likelihood function of the data given the model—that is, the objective function for parameter optimization—requires the integration over all possible trajectories of the systems being modeled. This is already infeasible for simple models. In practice, approximations of the likelihood are computed, usually based on a few realizations of the processes. For this reason, they are easily corrupted by large statistical noise. This noise is further amplified during gradient calculation using methods like finite differences. Statistical noise renders the reliable calculation mostly infeasible and prevents the use of scalable gradient-based optimization methods in most cases (Raue et al., 2013). Instead, simple manual line search methods are used in practice (see, e.g., Jagiella, 2012; and Karr et al., 2012). These methods are known to be inefficient, do not reliably converge to the best solutions, and do not provide reliable information about the parameter uncertainty.

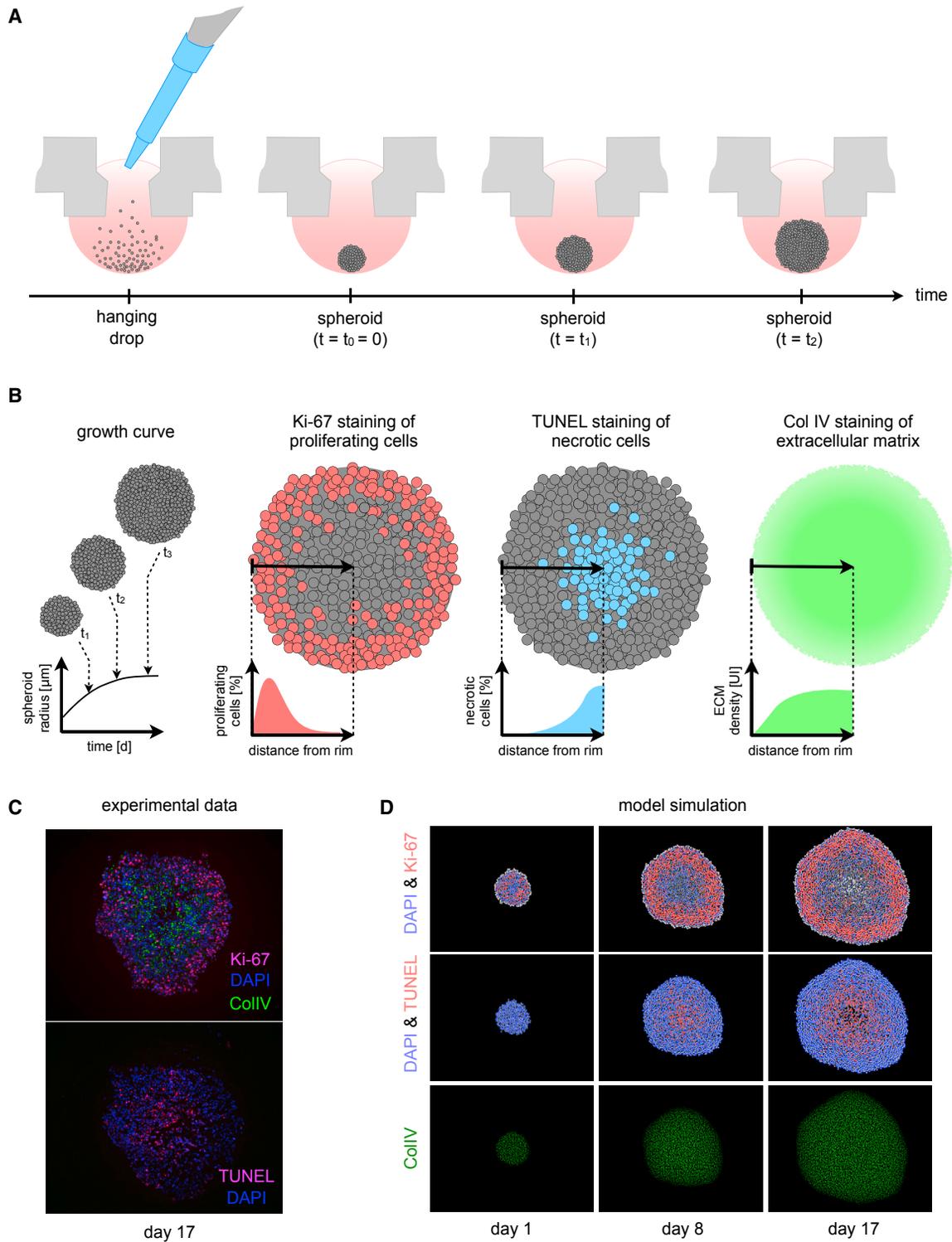
To infer parameters of stochastic processes, approximate Bayesian computation (ABC) algorithms have been developed (Beaumont et al., 2002). These ABC algorithms circumvent the evaluation of the likelihood function by assessing the distance between summary statistics of measured and simulated data. If the distance measure exceeds a threshold, the parameter values used to simulate data are rejected; otherwise, they are accepted. This concept can be used in rejection sampling (Beaumont et al., 2002), but as the acceptance rates are generally low, Markov chain Monte Carlo sampling (Marjoram et al., 2003; Sisson and Fan, 2011) and sequential Monte Carlo methods (Sisson et al., 2007; Toni and Stumpf, 2010; Toni et al., 2009) are usually more efficient. If the summary statistics are informative enough, samples obtained using ABC algorithms converge to the true posterior as the threshold approaches zero (Marin et al., 2014). A key advantage of ABC methods is that, in contrast to other search strategies (Adra et al., 2011; Karr et al., 2015), information about parameter and prediction uncertainties is obtained along with the calculation of good parameter estimates.

ABC algorithms have been used in a multitude of systems biology applications for the analysis of intra-cellular processes, e.g., gene expression and signal transduction (Liepe et al., 2013; Lillacci and Khammash, 2013; Loos et al., 2015; Toni et al., 2011, 2009). Furthermore, a few studies considered cell proliferation and cell movement using cellular Potts models (Sottoriva et al., 2015; Sottoriva and Tavaré, 2010) or agent-based

models (Johnston et al., 2014). In a recent study, ABC methods have even been used for the model-based analysis of intra-tumoral heterogeneity in colorectal cancer (Sottoriva et al., 2015). However, the inference of the hybrid stochastic-deterministic models of multi-scale processes has, to the best of our knowledge, not been reported. This may be because the number of necessary simulations is large, as is the computation time for individual simulations. For computationally less intensive problems, parallelization on small computing clusters (Feng et al., 2003; Jabot et al., 2013) and graphical processing units (GPUs) (Liepe et al., 2010) has been used to address such computational bottlenecks. Here, we move one step further—namely, to high-performance computing.

In this article, we introduce a parallel approximate Bayesian computation sequential Monte Carlo (pABC SMC) algorithm. This extension of the ABC SMC method facilitates the use of a broad spectrum of multi-core systems and computing clusters, thereby enabling the analysis of computationally demanding stochastic multi-scale models, including hybrid discrete-continuum models. Convergence of the pABC SMC sampling to the posterior distribution is ensured by sample sequence preservation. A crucial reduction of computation time is achieved using early rejection, a method implemented in several available ABC algorithms (see, e.g., Liepe et al., 2010). The pABC SMC algorithm facilitates parameter inference for the widely used class of hybrid discrete-continuum models. Hybrid discrete-continuum models are highly flexible, as they combine discrete agent-based descriptions of individual cells with continuous PDE-based description of extracellular substances.

We use the algorithm to analyze tumor spheroid growth in droplets (Figure 1A), an increasingly popular experimental model for anti-cancer drug screening (Carver et al., 2014; Kwapiszewska et al., 2014; Lemmo et al., 2014). The variability and morphology of tumor spheroids depend on various factors, including nutrition concentrations, and can be assessed using growth curves and immunostaining data (Figure 1B). Immunostaining data revealed that tumor spheroids usually consist of proliferating, quiescent, and necrotic cells. The cell fate depends on the microenvironment and intra-cellular processes, such as energy metabolism. Accordingly, multi-scale models describing the time-dependent spatial structure as well as properties of individual cells are required, which renders this an ideal test case for the pABC SMC algorithm. We consider a hybrid discrete-continuous model (Jagiella, 2012) for describing tumor spheroid growth. This model simulates up to  $10^6$  cancer cells on a growing three-dimensional domain. The individual cancer cells are modeled as discrete, interacting agents with intra-cellular information processing. The dynamics of extracellular substances, such as nutrition and extracellular matrix, are captured by reaction-diffusion equations. These reaction-diffusion equations are coupled with the agent dynamics. Experimental data and model simulations are illustrated in Figures 1C and 1D. In contrast to previous publications relying on tedious manual parameter tuning (Jagiella, 2012; Jagiella et al., 2016), the fully automated pABC SMC algorithm provides both parameter and prediction confidence bounds. Our study provides a proof-of-principle that the parameter inference for computationally demanding stochastic models of multi-cellular processes is feasible, using tailored, scalable estimation methods.

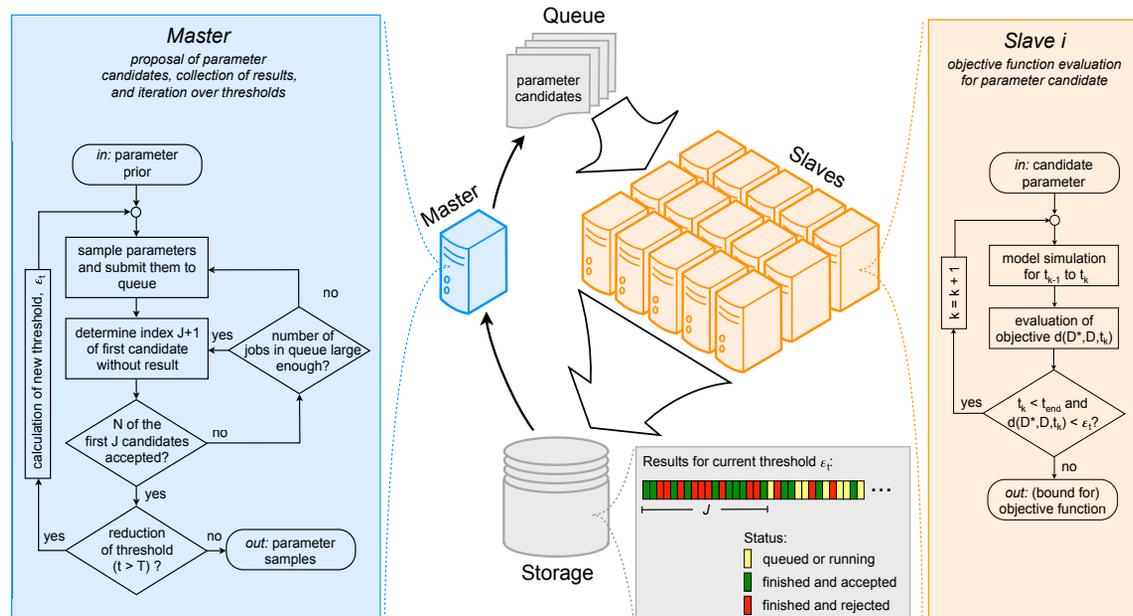


**Figure 1. Experimental Analysis and Modeling of Tumor Spheroid Growth**

(A) Schematic of 3D tumor spheroid culturing in hanging drops. Individual points indicate cells.

(B) Illustration of measurement data available for tumor spheroids: growth curves and marker staining. The imaging data are preprocessed, and the average staining for different distances from the spheroid rim is quantified.

(C and D) Shown here are (C) a representative imaging dataset (collected in Jagiella, 2012) and (D) illustrative model simulation for a glucose concentration ( $G$ ) of 25 mM and an oxygen concentration ( $O_2$ ) of 0.28 mM.



**Figure 2. Illustration of pABC SMC Methods**

The pABC SMC method uses a master/slave structure. The master node generates the parameter candidates, submits the jobs, collects the results, and proceeds to the next generation. Slave nodes simulate the model for different parameter values, evaluate the distance measure, and return the results. The results for individual simulations are stored in the order they have been submitted.

## RESULTS

### Implementation of pABC SMC Algorithms

To facilitate parameter estimation for computationally demanding hybrid discrete-continuum models, we implemented the pABC SMC algorithm illustrated in Figure 2. ABC methods rely on Bayes's theorem and approximate the posterior distribution  $p(\theta | \mathcal{D}) \propto p(\mathcal{D} | \theta)p(\theta)$  of the parameter  $\theta$  given the data  $\mathcal{D}$ . To circumvent the evaluation of the likelihood  $p(\mathcal{D} | \theta)$ , measured and simulated data are compared directly using distance measures  $d(\cdot, \cdot)$ . A parameter value  $\theta$  is accepted if the distance between a corresponding stochastic simulation and the data does not exceed a threshold  $\epsilon$ ; otherwise, the parameter vector  $\theta$  is rejected. To capture the posterior distribution, stochastic simulations for many proposed parameter values  $\theta$  have to be performed, yielding a sample of accepted parameters  $\{\theta^{(i)}\}_{i=1}^N$ . Straightforward but slow approaches sample the parameter values  $\theta$  from the prior  $p(\theta)$ . To accelerate convergence, the ABC SMC algorithm constructs a series of distributions for decreasing threshold  $\epsilon_t$ , with  $\epsilon_0 > \epsilon_1 > \dots > \epsilon_{T-1}$ . The sample  $\{\theta_t^{(i)}\}_{i=1}^N$  obtained for the threshold  $\epsilon_t$  is called generation  $t$ . For  $\epsilon_{T-1} \rightarrow 0$ , the final sample resembles the posterior distribution.

We parallelized the ABC SMC methods (Toni and Stumpf, 2010; Toni et al., 2009) by performing the simulation of the current generation  $t$  in parallel. For each threshold  $\epsilon_t$ , a sample of at least  $N$  accepted parameter values is required. To obtain this sample, the pABC SMC algorithm draws parameter candidates from the distribution approximation obtained for generation  $t - 1$ , simulates the hybrid discrete-continuum model, and evaluates the distance between simulation and data. The computationally inexpensive generation of parameter candi-

dates is performed in the master node, while simulation and objective function evaluation is parallelized using a large number of slave nodes. To accelerate the parameter estimation further, we intertwined simulation and distance measure evaluation. We used sums of weighted least-squares type distance measures, which strictly increase over time. If the objective function threshold  $\epsilon_t$  was already reached for the data points up to the current simulation time, the simulation was stopped, and the corresponding parameter vector was rejected. This early rejection procedure reduced the computation time by avoiding unnecessary calculations.

The proposed algorithm is suited for a large number of infrastructures (multi-core, GPU, cluster, etc.). We implemented it on a queue-mediated cluster architecture with over 1000 cores. A master is running the ABC SMC routine and is outsourcing the computation time and memory-consuming model simulation and distance evaluation to slave nodes. The work distribution is handled by a queue (Univa Grid Engine). The number of queued model evaluations is kept constant at  $m$ ; i.e., finished jobs are immediately replaced by new jobs. The evaluation results are stored in the same order as the corresponding jobs are submitted. As soon as the first  $J$  jobs are finished containing  $N$  accepted parameters, the master stops all still-running/queued evaluations and continues with the next generation. We note that it was important to not simply wait for  $N$  samples to be accepted, but we had to use  $N$  in the first  $J$  finished jobs. Otherwise, the parameter samples would have been biased toward regimes for which the computation time was lower. For details regarding the ABC SMC method and our parallel implementation, we refer to the STAR Methods.

### Model and Experimental Data of Tumor Spheroid Growth

To study the capabilities of the parallelized ABC SMC methods, we exploited it for the data-driven modeling of tumor spheroids formed by SK-MES-1 cells. In droplets, SK-MES-1 cells form spheroids with a rich spatial structure, including a proliferative rim and necrotic core, which resemble avascular tumors. These tumor spheroids are more suited for the analysis of drug delivery and drug response than mono-layer cultures (Carver et al., 2014; Kwapiszewska et al., 2014; Lemmo et al., 2014). However, an understanding of the underlying mechanisms requires quantitative mechanistic models. In the following, we consider 2D and 3D hybrid discrete-continuum models, which we developed previously (Jagiella, 2012). These models exploit an agent-based description for individual cells and a PDE-based description for extracellular metabolites and extracellular matrix (ECM) components. The intra-cellular regulation of cell division and of cell death is captured by a combination of continuous-time Markov chains and simple decision rules. The trajectories of the tumor growth models are subject to stochastic fluctuations. In particular, during the initial growth phase, which is marked by low cell numbers, stochastic simulations differ greatly. During later phases with higher cell numbers, a self-averaging effect occurs. Detailed descriptions of the models are provided in the STAR Methods.

We considered experimental data for tumor spheroids collected and processed by Jagiella et al. (2016). These experimental data provide the fraction of proliferation and necrotic cells, the relative ECM abundance, and the time-dependent spheroid radius (Figure 1B) under up to four experimental conditions, i.e., different oxygen and glucose concentrations (see STAR Methods). The data reveal that proliferation is limited to an outer rim, while cells further in the interior are mostly quiescent (Figure 1C). Furthermore, ECM abundance increases from the outer border toward the interior. For details regarding the experimental data and their evaluation, we refer to the original publication (Jagiella et al., 2016).

For evaluation purposes, we also consider artificial data obtained by simulating the model for the known parameter values (STAR Methods). Figure 1D depicts a sequence of snapshots, illustrating the time evolution of the model. The artificial data closely resemble the aforementioned properties of the experimental observations. Furthermore, we observe substantial stochastic variability between realizations. This stochastic variability poses challenges and renders this model ideal for the evaluation of our pABC SMC algorithm.

### Performance and Reliability of the pABC SMC Algorithm

Given the challenges of statistical inference for stochastic models, we asked whether the pABC SMC algorithm can fit hybrid discrete-continuum models and whether it provides reliable parameter estimates. To address this, we used the 2D model and the corresponding artificial dataset. A single experimental condition without nutrition limitation was considered, implying that cell proliferation depends exclusively on the available space and the ECM abundance. Parameters used to simulate the artificial data and to specify of the experimental condition are provided in the STAR Methods. For the estimation, the parameters  $\theta_i$  were restricted to the range  $10^{-5}$ – $10^0$  to resemble the common lack of prior information. The sum of weighted

least-squares was used to measure the distance between measured data and simulation, using the SD of each data point as weighting.

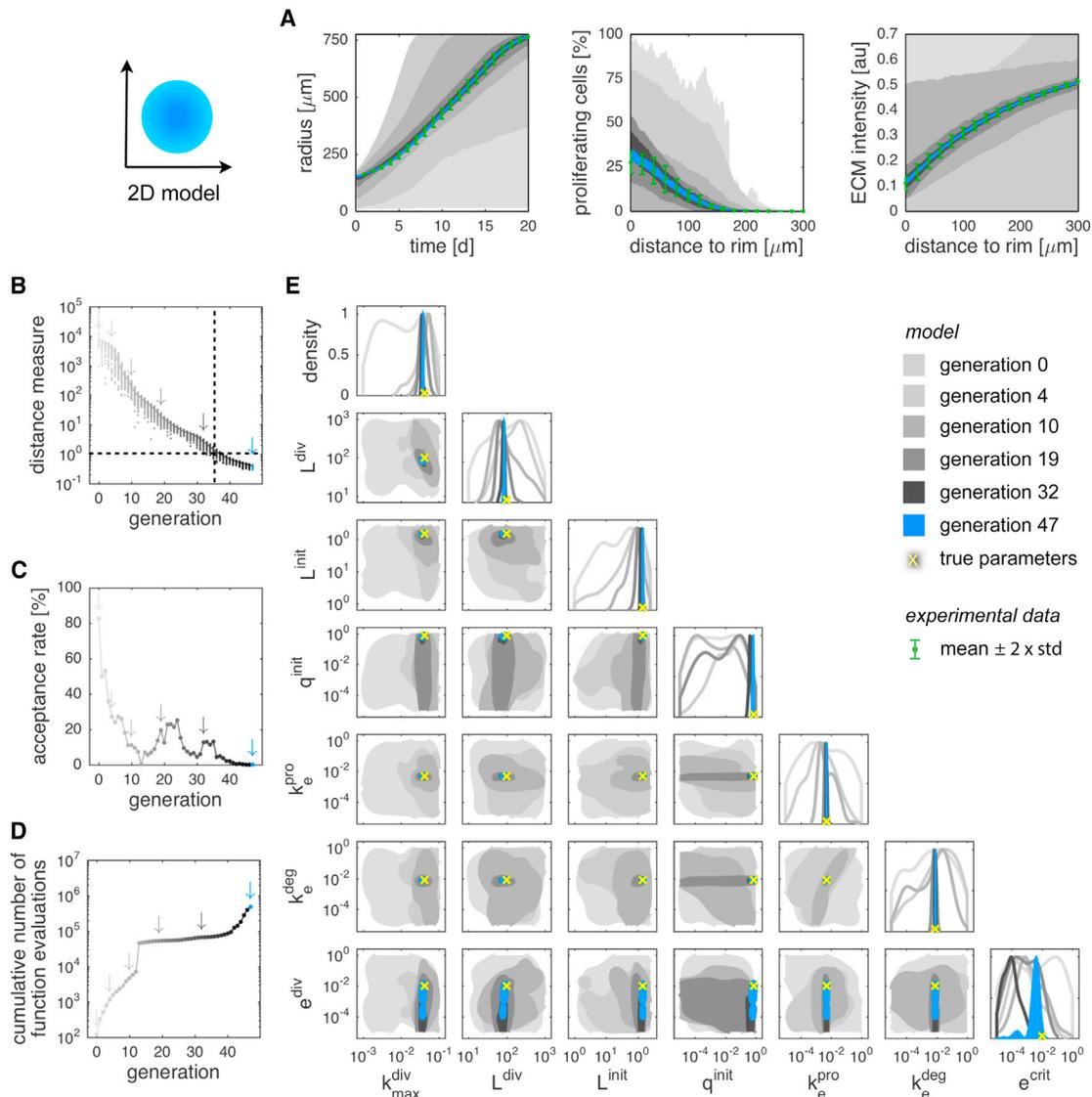
A visualization of the behavior of the pABC SMC algorithm is provided in Figure 3. We found that the pABC SMC algorithm yielded excellent fits to the artificial experimental data (Figure 3A). Although not a single member of the first generation of the sequential scheme provided a satisfactory fit, after 35 generations, the model simulations closely resembled the observed data. After 35 generations, the normalized fitting error per data point was below 1, which is what we expect for the true parameters (Figure 3B). For the subsequent generations, we observed an acceptance rate for new parameter candidates below 5% (Figure 3C), resulting in a rapid increase of the cumulative number of function evaluations (Figure 3D). This was not surprising, as we found in an independent evaluation that, even for simulations with the true parameter values, a small fraction of the stochastic simulations was accepted. Over the different generations, the parameter sample successively contracted around the true parameter used to generate the artificial data (Figure 3E). Hence, we concluded that the pABC SMC algorithm worked. While the final confidence intervals for most parameters were narrow, for the critical ECM concentration,  $e^{\text{div}}$ , we observed a relatively large uncertainty. This indicated a weaker dependence of the observables on the critical ECM concentration than on the other parameters. All these findings were reproducible across several runs of the method.

In total, for parameter estimation, we used a queue with  $C = 100$  cores and required  $N = 100$  accepted samples per generation. An individual simulation of the 2D model took, on average, about 0.1 min, resulting in an overall computation time of roughly  $10^4$  CPU hr. Accordingly, parallelization was essential for obtaining results in a reasonable amount of time. As the sample size  $N$  influences the convergence of the estimators, as well as the computation time, we studied its impact on the approximation of the posterior distribution  $p(\theta | \mathcal{D})$ . We found that, for this estimation problem,  $N = 100$  is sufficient, as similar results were observed for large sample sizes, e.g.,  $N = 1,000$ . A significant decrease of the sample size below  $N = 100$  resulted in convergence problems and biased results. Potential causes are the limited coverage of the distribution and degeneracy of the perturbation kernel (see STAR Methods). The computation time increased linearly with  $N$ , which was expected.

Our analysis of artificial data verified that the pABC SMC algorithm facilitates the reliable inference of hybrid discrete-continuum models. The algorithm worked robustly despite the stochastic nature of the problems and parallelization rendered its application tractable for complex simulation models.

### Consistency of Parameter Estimates for 2D and 3D Models

The positive results for the artificial data suggested that the pABC SMC algorithm might be suited for the application to experimental data. To evaluate this, we considered the aforementioned published experimental data for SK-MES-1 cells (Jagiella et al., 2016). These data were already modeled using the hybrid discrete-continuum model that we considered in the previously published article. However, in that previous work, parameters were determined using a combination of



**Figure 3. Evaluation of pABC SMC for Artificial Data**

(A) Artificial data and fits for generations 0, 4, 10, 19, 32, and 47. For the fit, the 90% confidence intervals of the accepted stochastic simulations are depicted. std, SD.

(B) Distance between simulation and data for accepted samples of different generations. The line of medians is provided as reference.

(C) Acceptance rate for different generations. The seemingly low acceptance rate for generation 13 is caused by a single stochastic simulation that took very long, delaying the progression to the next generation.

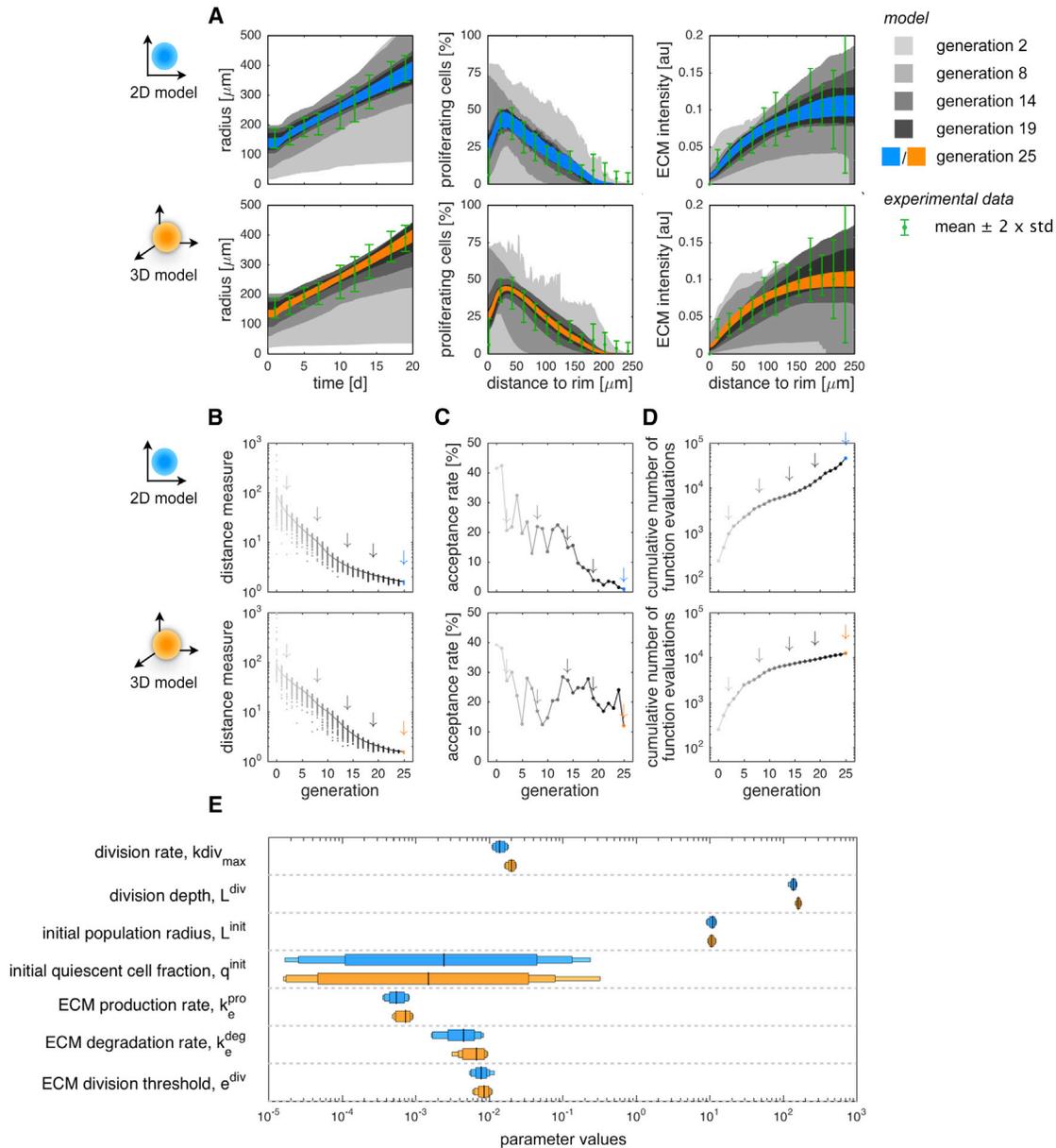
(D) Cumulative number of function evaluations for the different generations of the pABC SCM algorithm.

(E) 2D scatterplots of parameter samples for different generations and true parameter. For all parameter pairs, the 90% confidence regions are depicted. The colors in the different subplots are matched, and the corresponding generations are indicated by arrows.

manual search and parameter sweeps. Although neither optimization nor uncertainty analysis had been performed, we considered the parameters derived in Jagiella et al. (2016) as reference parameters,  $\theta^{\text{ref}}$ , and restricted our search domain to  $\theta \in [10^{-2} \cdot \theta^{\text{ref}}, 10^2 \cdot \theta^{\text{ref}}]$ .

The 3D model captured the dynamics of up to  $10^6$  cells and required the simulation of a 3D system of coupled PDEs. A single simulation of the 3D model at the reference parameters for all four experimental conditions required 3–4 CPU days. This computation time posed a serious challenge for parameter esti-

mation and rendered parallelization essential. To assess the feasibility of inference using the 3D model, we first considered only the experimental condition without nutrition limitations (25 mM glucose and 0.28 mM oxygen). In this condition, the model simplified as the PDEs for glucose and oxygen concentrations could be disregarded. This reduced the computation time for the 3D model for this condition to roughly 1 CPU hr. We used the pABC SMC algorithm to estimate the parameters of the 3D model in the reduced setting. In addition, we estimated the parameter of the 2D model, for which simulation required



**Figure 4. Comparison of Inferences Using 2D and 3D Models for Experimental Data**

(A) Experimental data and fits for the 2D and 3D models for generations 2, 8, 14, 19, and 25. For the fit, the 90% confidence intervals of the accepted stochastic simulations are depicted. std, SD.

(B) Distance between simulation and data for accepted samples for different generation. The median is provided as reference.

(C) Acceptance rate for different generations.

(D) Cumulative number of function evaluations.

(E) Confidence intervals for parameters of the 2D model and the 3D model for the final generation. The horizontal bars represent the confidence intervals corresponding to different confidence levels (80%, 95%, and 99%), and the line indicates the median.

The colors in the different subplots are matched and the corresponding generations indicated by arrows.

roughly 0.1 CPU min, and asked how similar the estimation results obtained using 2D and 3D models are for this setting. The estimation results are summarized in Figure 4.

The evaluation of the estimation results revealed that the 2D model and the 3D model could be fitted to the experimental data using our pABC SMC algorithm (Figure 4). This verified the practical applicability of the method and the feasibility of sta-

tistical inference for computationally intensive multi-scale models. Both the 2D and 3D models allowed for a good description of the experimental data (Figure 4A). Furthermore, the convergence properties for both models were compatible (Figure 4B), while the acceptance rates and the cumulative number of function evaluations were slightly better for the 3D model (Figures 4C and 4D). As the simulation of the 2D model was,

however, almost two orders of magnitude faster than for the 3D model, the parameter estimation for the 2D model was substantially faster. The difference in computation time appeared, although the computationally most intensive simulations of the 3D model were avoided by the early rejection methods.

While the 3D model described a spheroid, the 2D model essentially assumed symmetry in the third direction and, instead, described a cylinder. Given the difference, we were surprised that the parameter estimates were in good agreement. The posterior medians, as well as the confidence intervals, are similar (Figure 4E). This implied that, for high nutrition concentrations, the parameters of the 3D biological process could be inferred using a 2D model.

### Multi-experiment Data Integration

Given the feasibility of parameter estimation for single experimental conditions, we considered the problem of model-based data integration across experimental conditions. We used previously measured growth curves and histological information (Jagiella et al., 2016) for up to four experimental conditions with differing glucose and oxygen concentrations. For the lower glucose and oxygen concentrations, cells in the core of the spheroid might suffer nutrition limitations. Therefore, we used the hybrid discrete-continuum model, which captures the local glucose, oxygen, lactate, and cell debris concentrations. In line with the results presented in the previous section, we used the 2D model to reduce the computational complexity. This complexity, however, remained substantial as (1) the simulation of the 2D model for all four conditions under the altered setting takes hours and as (2) the number of unknown parameters increases from 7 to 18. The latter required an increased sample size,  $N = 1000$  as found by preliminary evaluations.

We performed the parameter estimation using our pABC SMC algorithm on a cluster with over 1000 cores. The calculation ran for roughly 1 month, corresponding to an overall computation time of almost  $10^6$  CPU hr. Accordingly, parameter estimation for this multi-scale and multi-cellular model would not have been possible without massive parallelization. The fit achieved using the Big Computing approach closely resembled the measured growth curves (Figure 5A) and immunostaining data (Figure 5B) for all experimental conditions. Among others, the slow spheroid growth under low glucose or oxygen concentrations (conditions III and IV) (Figure 5A) and the altered necrosis profile (conditions II versus III) on day 17 (Figure 5B) and day 24 (Figure S1) were captured. The predictions for proliferation, necrosis, and ECM profiles for conditions under which they have not been measured (conditions III and IV) appeared plausible.

Our results showed that the 2D model can resemble the data measured in the 3D system under four different experimental conditions. Previously, however, we only verified the consistency of the 2D and 3D models under high nutrition concentrations. To assess whether the results also hold in this more complex scenario, we subsampled the parameter sample obtained using the 2D model and used the subsample obtained to simulate the 3D model. The simulation results for the 3D model, indeed, closely resembled the experimental data and the fitting results of the 2D model. Only the saturated growth observed under conditions II and III were mis-matched. Notably,

however, the measurement uncertainty in this regime was high, and the experimental data showed, counterintuitively, stronger growth under lower glucose (condition I versus condition II) concentrations after 30 days. This suggests that the mis-match between model and experiment likely reflects the fact that the experiment was conducted in an atypical biological regime rather than a problem with the model per se.

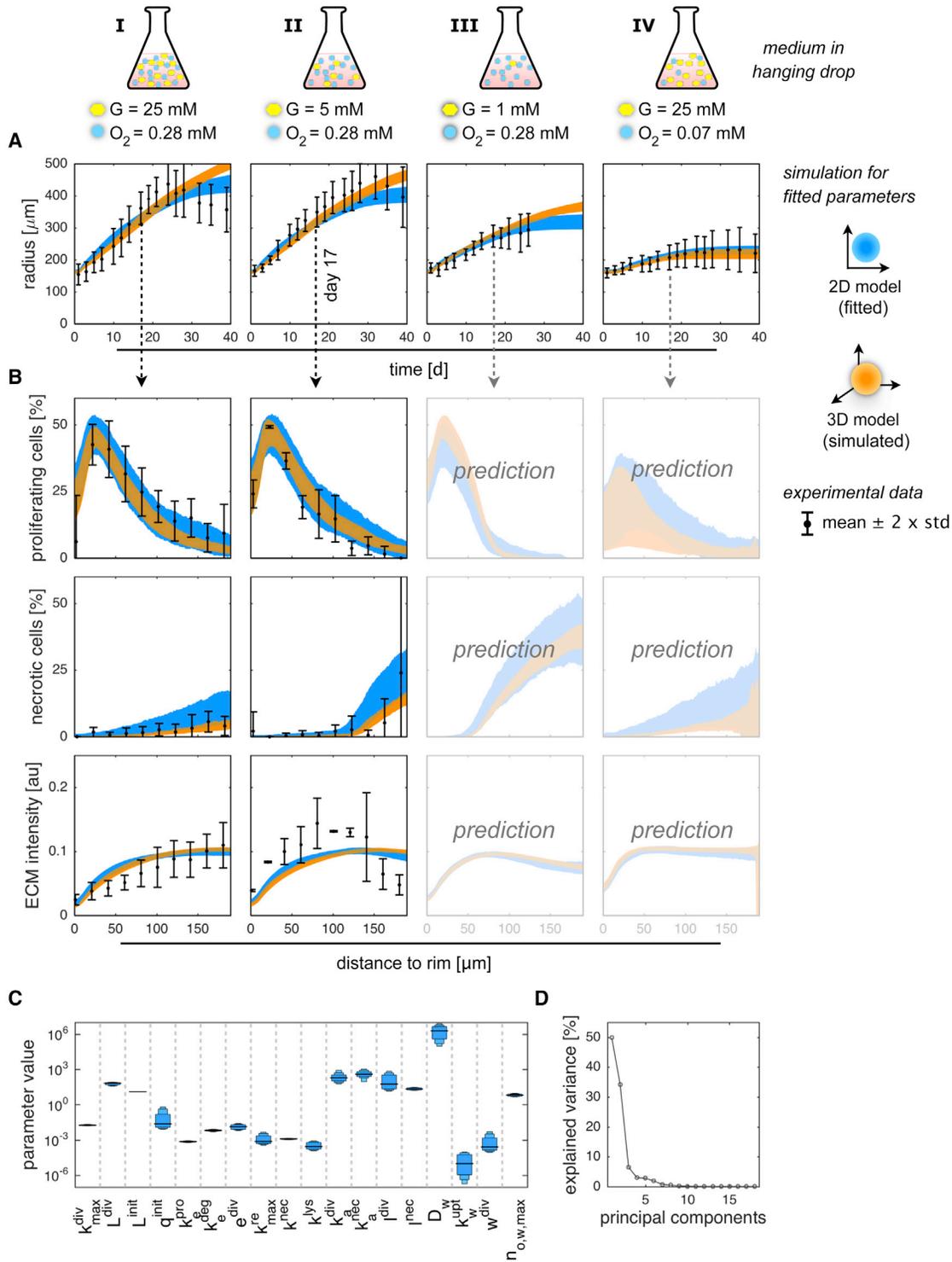
To assess the uncertainty of the individual model parameters, we analyzed the final parameter sample. Although the parameter dimension increased, the parameter uncertainties are comparatively small (Figure 5C). In addition, the first two principal components of the parameter sample capture most of the variability (Figure 5D), implying that all but two directions in parameter space are well determined. The good parameter identifiability was achieved by integrating multiple experimental conditions and data types. We evaluated how the parameter identifiability depends on the availability of individual readouts, e.g., the fraction of necrotic cells. To achieve this, we re-ran the pABC SMC algorithm for the 2D model presented in the previous section with different reduced datasets. The analysis revealed that, already, the removal of a single readout would result in large parameter and prediction uncertainties (Figure S2).

### Uncertainty-Aware Prediction of Tumor Spheroid Growth

Beyond the integration of experimental data for measured experimental conditions, statistical inference of mechanistic models facilitates uncertainty-aware predictions. To illustrate this, we studied tumor spheroid growth behavior for a wide range of glucose and oxygen concentrations using the 2D model. Among others, we considered the depth of the proliferating zone, the depth of the viable zone, and the initial growth rate. To account for stochasticity and parameter uncertainties, stochastic simulations are performed for the parameter sample obtained by the pABC SMC algorithm.

The analysis of stochastic simulations for a broad spectrum of nutrition concentrations indicated the existence of three growth regimes. For glucose concentrations  $< 0.1$  mM, no growth is observed. The depth of the proliferating zone and the initial growth rate were both zero (Figures 6A and 6B), and cells were undergoing necrosis. For glucose concentrations  $> 0.1$  mM and oxygen concentrations  $< 0.1$  mM, the model predicted an initial spheroid growth rate of  $2 - 5$   $\mu\text{m}/\text{d}$ . The initial growth rate and the depth of the proliferating zone slightly increased with the glucose concentration but were essentially independent of the oxygen concentration, indicating anaerobic growth. For glucose concentrations  $> 0.1$  mM and oxygen concentrations  $> 0.1$  mM, the model predicted initial growth rates of up to  $15$   $\mu\text{m}/\text{d}$ . In this aerobic growth regime, the initial growth rate and the depth of the proliferating zone depended strongly on the glucose concentration but were again almost independent of the oxygen concentration. Accordingly, the oxygen concentration only controls the switch between anaerobic and aerobic growth, a result of the metabolic model embedded in the individual cells.

To assess the reliability of these predictions, we evaluated the SD of the growth properties considered. We found that the variability of the model predictions—this considered stochasticity and parameter uncertainty—was small compared to the

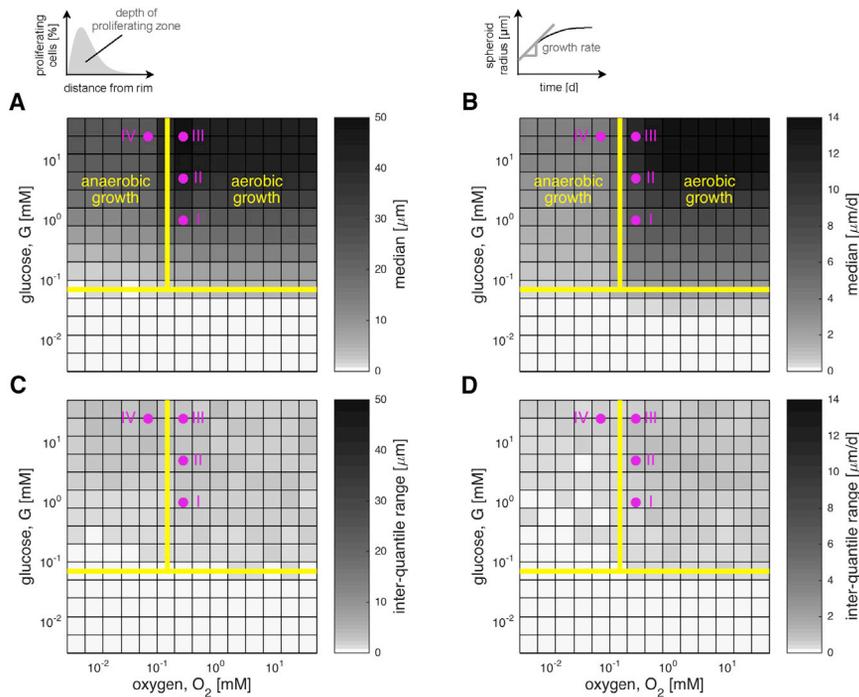


**Figure 5. Multi-experiment Data Integration**

(A and B) Shown here are (A) growth curves and (B) immunostainings on day 17. Experimental data, the fitting result for the 2D model, and simulation results for the 3D model are depicted. The simulation results for the 3D model were obtained using the parameter sample determined by fitting the 2D model. For the 2D and 3D models, the 90% percentile intervals of the fitting/simulation results are depicted. G, glucose. std, SD.

(C) Confidence intervals for parameters of the 2D model for the final generation. The vertical bars represent the confidence intervals corresponding to different confidence levels (80%, 95% and 99%), while the line indicates the median.

(D) Contribution of principal components to the overall variance in the parameter sample.



**Figure 6. Model-Based Prediction of Growth Behavior for Different Nutrient Conditions**

(A–D) In (A and B), the median of the simulation results are shown, providing a prediction. (C and D) Inter-quantile range of simulation results, providing the prediction uncertainty resulting from parameter uncertainty and stochastic variability. The prediction and prediction uncertainties are visualized for (A and C) depth of proliferating zone on day 17 and (B and D) median growth rate in the linear regime. The shading indicates the values of the median and inter-quantile range obtained from 50 simulation runs of the 2D models for parameters sampled from the final generation. The dots indicate the nutrition combinations of the experimental data used for fitting.

changes observed across the studied range of nutrition conditions (Figures 6C and 6D). This was also the case for nutrition conditions that were far from the conditions for which experimental data were collected. This analysis demonstrates that not only are our model's parameters defined with high confidence, but its predictions are also. In addition to the dependence of the growth behavior on the oxygen concentration, we found several interesting features that are predicted with similar exactitude. For example, in the anaerobic regime, increasing the glucose concentration results in an increase of the depth of the proliferating zone before the depth of the viable zone increases (Figures S3A and S3B). Thus, the fitted model provided testable predictions (with uncertainty bounds) for model validation *in vivo*.

## DISCUSSION

In the past, quantitative multi-scale models have mostly been obtained by data-driven modeling of individual scales and subsequent coupling (Chew et al., 2014; Hayenga et al., 2011; ten Tusscher et al., 2004). While this approach is usually computationally less demanding than parameter estimation for multi-scale models, for certain classes of multi-scale couplings, it is not applicable, and consistency as well as optimality cannot be ensured (Hasenauer et al., 2015). In addition, in many studies, experimental data for different submodels have been collected under different experimental conditions, raising questions of model validity. To overcome these limitations, methods for integrated statistical inference need to be adapted for the challenges faced in multi-scale modeling. In this article, we propose a pABC SMC algorithm that provides reliable confidence intervals in agreement with theory on ABC (see, e.g., Marjoram et al., 2003; Sisson et al., 2007; Toni et al., 2009 and references therein). The application of the method to 2D and 3D hybrid

computationally demanding stochastic multi-scale models in systems biology.

The pABC SMC algorithms that we implemented worked efficiently for the examples considered; however, a variety of aspects might be improved. Sophisticated local perturbation kernels (Filippi et al., 2013) and optimized threshold schedules (Silk et al., 2013) can reduce the required number of function evaluations and improve the convergence. Moreover, methods to adjust the effective sample size online might improve the robustness of the methods. For the considered inference problems, surprisingly low sample sizes proved to be sufficient. For problems with higher dimensional parameter spaces and posterior distribution with complex shapes, including multiple modes, a substantially larger number of samples will be required. These improvements will facilitate the analysis of even larger multi-scale models, e.g., models for the study of intra-tumor heterogeneity in large lesions (Waclaw et al., 2015).

Beyond parameter estimation, many applications require the comparison of competing hypotheses, also known as model selection. Similar to the standard ABC SMC algorithm (Toni and Stumpf, 2010), pABC SMC can be used for model selection by including the model index as an additional (discrete) variable. While this does not require any changes to the implementation, the choice of appropriate distance measures and summary statistics becomes even more critical (Robert et al., 2011). As for multi-scale models, the selection of important features of the data and their weighting is non-trivial; methods for the optimal selection of summary statistics might be used (Nunes and Balding, 2010). The evaluation of the method on the experimental data revealed that the weighted least-squares method, with weights determined from the SDs of experimental replicates, does not work reliably, as the number of replicates is usually too small to obtain robust estimates of the SDs. Results obtained

using the dynamic range of the signal turned out to be more robust. The improvements on the methodological side need to be complemented by the development of software packages and standards to improve the reproducibility, the transparency, and the exchange of models further.

In basic research, as well as clinical applications, a multitude of tissue samples are collected and analyzed. This provides a wealth of experimental data, which is mostly analyzed using statistical tools. The measured data are, however, associated to mixtures of different, interacting cell types arranged in complex morphologies. This renders a simple analysis of the resulting averages problematic and, in some situations, even misleading (Altschuler and Wu, 2010; Hasenauer et al., 2014; Intosalmi et al., 2016). Multi-scale models combined with advanced statistical inference methods can contribute to the deconvolution and subsequent mechanistic interpretation of the data. They allow for the integration of prior knowledge on intra- and inter-cellular processes from available databases, such as STRING (Franceschini et al., 2013), KEGG (Kanehisa and Goto, 2000), and Reactome (Croft et al., 2011), as well as the integration of multiple data sources. In addition, mechanistic/first-principles modeling on different scales can effectively reduce the number of parameters, as macroscopic properties usually originate directly from microscopic properties (Kevrekidis and Samaey, 2009). This turns data-driven, multi-scale modeling into an enabling technology.

The relevance of multi-scale and multi-cellular models in systems biology is steadily increasing (Dada and Mendes, 2011; Hunter and Borg, 2003; Martins et al., 2010; Walpole et al., 2013); however, the methods for automated statistical inference are lagging behind. We introduced the pABC SMC algorithm, the first and only method to allow parameter estimation for detailed stochastic multi-scale models. The pABC SMC algorithm is not only an improvement over existing ABC methods but it also actually renders a new class of problems solvable by exploiting high-performance computing. We demonstrated this for a hybrid discrete-continuum model of tumor spheroids with single-cell resolution. The pABC SMC algorithm is applicable to a broad classes of multi-scale models and provides novel insights via the consistent integration of data from multiple experiments and measurement devices. In addition, by eliminating the need for error-prone manual parameter tuning and the bias of individual researchers, the proposed method will improve the reproducibility of multi-scale modeling studies. This renders the pABC SMC algorithm and the extension of it valuable for the analysis of a broad class of modeling projects in quantitative biology. This can result in a paradigm shift toward data-driven multi-scale modeling and could have a considerable impact on computational modeling.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Growth curves and histological imaging data

## ● METHOD DETAILS

- Hybrid discrete-continuum model for tumor spheroid growth
- Individual-based description of single-cell dynamics
- Continuum-based description of the dynamics of extracellular substances
- Numerical simulation

## ● QUANTIFICATION AND STATISTICAL ANALYSIS

- Parallel Approximate Bayesian Computing Sequential Monte Carlo method
- Distance measure
- Adaptation of perturbation kernel and threshold
- Parameterization, prior distribution and parameter bounds
- Population size and analysis of convergence
- Analysis of parameter and prediction uncertainties
- Prediction of spheroid growth characteristics
- Assessment of the importance of individual data types
- Implementation of the statistical analysis

## ● DATA AND SOFTWARE AVAILABILITY

- Data resources
- Software resources

## SUPPLEMENTAL INFORMATION

Supplemental Information includes three figures and can be found with this article online at <http://dx.doi.org/10.1016/j.cels.2016.12.002>.

## AUTHOR CONTRIBUTIONS

Conceptualization, F.J.T. and J.H.; Methodology, N.J., D.R., F.J.T., and J.H.; Investigation, N.J., D.R., and J.H.; Writing, N.J. and J.H.; Funding Acquisition, F.J.T. and J.H.; Resources, N.J., F.J.T., and J.H.; Supervision, J.H.

## ACKNOWLEDGMENTS

The authors acknowledge financial support from the German Federal Ministry of Education and Research (BMBF) within the SYS-Stomach project (Grant No. 01ZX1310B) and the Postdoctoral Fellowship Program (PFP) of the Helmholtz Zentrum München.

Received: July 12, 2016

Revised: September 14, 2016

Accepted: November 30, 2016

Published January 11, 2017

## REFERENCES

- Adra, S.F., Kiran, M., McMinn, P., and Walkinshaw, N. (2011). A multiobjective optimisation approach for the dynamic inference and refinement of agent-based model specifications. In Proceedings of the IEEE Congress on Evolutionary Computation (CEC) (New Orleans, LA: IEEE), pp. 2237–2244.
- Altschuler, S.J., and Wu, L.F. (2010). Cellular heterogeneity: do differences make a difference? *Cell* 141, 559–563.
- Anderson, A.R.A., and Quaranta, V. (2008). Integrative mathematical oncology. *Nat. Rev. Cancer* 8, 227–234.
- Baumont, M.A., Zhang, W., and Balding, D.J. (2002). Approximate Bayesian computation in population genetics. *Genetics* 162, 2025–2035.
- Carver, K., Ming, X., and Juliano, R.L. (2014). Multicellular tumor spheroids as a model for assessing delivery of oligonucleotides in three dimensions. *Mol. Ther. Nucleic Acids* 3, e153.
- Chew, Y.H., Wenden, B., Flis, A., Mengin, V., Taylor, J., Davey, C.L., Tindal, C., Thomas, H., Ougham, H.J., de Reffye, P., et al. (2014). Multiscale digital

- Arabidopsis predicts individual organ and whole-organism growth. *Proc. Natl. Acad. Sci. USA* **111**, E4127–E4136.
- Croft, D., O’Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., et al. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **39**, D691–D697.
- Dada, J.O., and Mendes, P. (2011). Multi-scale modelling and simulation in systems biology. *Integr. Biol.* **3**, 86–96.
- Eldar, A., and Elowitz, M.B. (2010). Functional roles for noise in genetic circuits. *Nature* **467**, 167–173.
- Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic gene expression in a single cell. *Science* **297**, 1183–1186.
- Feng, X., Rose, D.B.J.R., and Waddell, P.J. (2003). Parallel algorithms for Bayesian phylogenetic inference. *J. Parallel Distrib. Comput.* **63**, 707–718.
- Filippi, S., Barnes, C.P., Cornibise, J., and Stumpf, M.P. (2013). On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. *Stat. Appl. Genet. Mol. Biol.* **12**, 87–107.
- Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., and Jensen, L.J. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41**, D808–D815.
- Gillespie, D.T. (1977). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361.
- Graner, F., and Glazier, J.A. (1992). Simulation of biological cell sorting using a two-dimensional extended Potts model. *Phys. Rev. Lett.* **69**, 2013–2016.
- Hasenauer, J., Hasenauer, C., Hucho, T., and Theis, F.J. (2014). ODE constrained mixture modelling: a method for unraveling subpopulation structures and dynamics. *PLoS Comput. Biol.* **10**, e1003686.
- Hasenauer, J., Jagiella, N., Hross, S., and Theis, F.J. (2015). Data-driven modelling of biological multi-scale processes. *J. Coupled Syst. Multiscale Dyn.* **3**, 101–121.
- Hayenga, H.N., Thorne, B.C., Peirce, S.M., and Humphrey, J.D. (2011). Ensuring congruency in multiscale modeling: towards linking agent based and continuum biomechanical models of arterial adaptation. *Ann. Biomed. Eng.* **39**, 2669–2682.
- Hoehme, S., Brulport, M., Bauer, A., Bedawy, E., Schormann, W., Hermes, M., Puppe, V., Gebhardt, R., Zellmer, S., Schwarz, M., et al. (2010). Prediction and validation of cell alignment along microvessels as order principle to restore tissue architecture in liver regeneration. *Proc. Natl. Acad. Sci. USA* **107**, 10371–10376.
- Hug, S., Raue, A., Hasenauer, J., Bachmann, J., Klingmüller, U., Timmer, J., and Theis, F.J. (2013). High-dimensional Bayesian parameter estimation: case study for a model of JAK2/STAT5 signaling. *Math. Biosci.* **246**, 293–304.
- Huh, D., and Paulsson, J. (2011). Non-genetic heterogeneity from stochastic partitioning at cell division. *Nat. Genet.* **43**, 95–100.
- Hunter, P.J., and Borg, T.K. (2003). Integration from proteins to organs: the Physiome Project. *Nat. Rev. Mol. Cell Biol.* **4**, 237–243.
- Intosalmi, J., Nousiainen, K., Ahlfors, H., and Lähdesmäki, H. (2016). Data-driven mechanistic analysis method to reveal dynamically evolving regulatory networks. *Bioinformatics* **32**, i288–i296.
- Jabot, F., Faure, T., and Dumoulin, N. (2013). EasyABC: performing efficient approximate Bayesian computation sampling schemes using R. *Methods Ecol. Evol.* **4**, 684–687.
- Jagiella, N. (2012). Parameterization of lattice-based tumor models from data. PhD thesis (Université Pierre et Marie Curie, Paris, France).
- Jagiella, N., Müller, B., Müller, M., Vignon-Clementel, I.E., and Drasdo, D. (2016). Inferring growth control mechanisms in growing multi-cellular spheroids of NSCLC cells from spatial-temporal image data. *PLoS Comput. Biol.* **12**, e1004412.
- Johnston, S.T., Simpson, M.J., McElwain, D.L., Binder, B.J., and Ross, J.V. (2014). Interpreting scratch assays using pair density dynamics and approximate Bayesian computation. *Open Biol.* **4**, 140097.
- Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30.
- Karr, J.R., Sanghvi, J.C., Macklin, D.N., Gutschow, M.V., Jacobs, J.M., Bolival, B., Jr., Assad-Garcia, N., Glass, J.I., and Covert, M.W. (2012). A whole-cell computational model predicts phenotype from genotype. *Cell* **150**, 389–401.
- Karr, J.R., Williams, A.H., Zucker, J.D., Raue, A., Steiert, B., Timmer, J., Kreutz, C., Wilkinson, S., Allgood, B.A., Bot, B.M., et al.; DREAM8 Parameter Estimation Challenge Consortium (2015). Summary of the DREAM8 parameter estimation challenge: Toward parameter identification for whole-cell models. *PLoS Comput. Biol.* **11**, e1004096.
- Kevrekidis, I.G., and Samaey, G. (2009). Equation-free multiscale computation: algorithms and applications. *Annu. Rev. Phys. Chem.* **60**, 321–344.
- Klann, M.T., Lapin, A., and Reuss, M. (2009). Stochastic simulation of signal transduction: impact of the cellular architecture on diffusion. *Biophys. J.* **96**, 5122–5129.
- Kong, A., Liu, J.S., and Wong, W.H. (1994). Sequential imputations and Bayesian missing data problems. *J. Am. Stat. Assoc.* **89**, 278–288.
- Kwapiszewska, K., Michalczyk, A., Rybka, M., Kwapiszewski, R., and Brzózka, Z. (2014). A microfluidic-based platform for tumour spheroid culture, monitoring and drug screening. *Lab Chip* **14**, 2096–2104.
- L’Ecuyer, P., and Simard, R. (2007). TestU01: A C library for empirical testing of random number generators. *ACM Trans. Math. Softw.* **33**, Article 22.
- Lemmo, S., Atefi, E., Luker, G.D., and Tavana, H. (2014). Optimization of aqueous biphasic tumor spheroid microtechnology for anti-cancer drug testing in 3D culture. *Cell. Mol. Bioeng.* **7**, 344–354.
- Liepe, J., Barnes, C., Cule, E., Erguler, K., Kirk, P., Toni, T., and Stumpf, M.P.H. (2010). ABC-SysBio—approximate Bayesian computation in Python with GPU support. *Bioinformatics* **26**, 1797–1799.
- Liepe, J., Filippi, S., Komorowski, M., and Stumpf, M.P.H. (2013). Maximizing the information content of experiments in systems biology. *PLoS Comput. Biol.* **9**, e1002888.
- Lillacci, G., and Khamash, M. (2013). The signal within the noise: efficient inference of stochastic gene regulation models using fluorescence histograms and stochastic simulations. *Bioinformatics* **29**, 2311–2319.
- Liu, J.S. (1996). Metropolisized independent sampling with comparisons to rejection sampling and importance sampling. *Stat. Comput.* **6**, 113–119.
- Loos, C., Marr, C., Theis, F.J., and Hasenauer, J. (2015). Approximate Bayesian Computation for stochastic single-cell time-lapse data using multivariate test statistics. In *Computational Methods in Systems Biology*, O. Roux and J. Bourdon, eds. (Springer International Publishing), pp. 52–63.
- Marin, J.-M., Pillai, N.S., Robert, C.P., and Rousseau, J. (2014). Relevant statistics for Bayesian model choice. *J. R. Stat. Soc. B* **76**, 833–859.
- Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **100**, 15324–15328.
- Martins, M.L., Ferreira, S.C., Jr., and Vilela, M.J. (2010). Multiscale models for biological systems. *Curr. Opin. Colloid Interface Sci.* **15**, 18–23.
- Matsumoto, M., and Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput. Simul.* **8**, 3–30.
- Mirams, G.R., Arthurs, C.J., Bernabeu, M.O., Bordas, R., Cooper, J., Corrias, A., Davit, Y., Dunn, S.-J., Fletcher, A.G., Harvey, D.G., et al. (2013). Chaste: an open source C++ library for computational physiology and biology. *PLoS Comput. Biol.* **9**, e1002970.
- Nielsen, B.F., Lysaker, M., and Grøtting, P. (2013). Computing ischemic regions in the heart with the bidomain model—first steps towards validation. *IEEE Trans. Med. Imaging* **32**, 1085–1096.
- Niepel, M., Spencer, S.L., and Sorger, P.K. (2009). Non-genetic cell-to-cell variability and the consequences for pharmacology. *Curr. Opin. Chem. Biol.* **13**, 556–561.
- Noble, D. (2002). Modeling the heart—from genes to cells to the whole organ. *Science* **295**, 1678–1682.

- Nunes, M.A., and Balding, D.J. (2010). On optimal selection of summary statistics for approximate Bayesian computation. *Stat. Appl. Genet. Mol. Biol.* 9, Article 34.
- Raue, A., Schilling, M., Bachmann, J., Matteson, A., Schelker, M., Kaschek, D., Hug, S., Kreutz, C., Harms, B.D., Theis, F.J., et al. (2013). Lessons learned from quantitative dynamical modeling in systems biology. *PLoS ONE* 8, e74335.
- Richmond, P., Walker, D., Coakley, S., and Romano, D. (2010). High performance cellular level agent-based simulation with FLAME for the GPU. *Brief. Bioinform.* 11, 334–347.
- Robert, C.P., Cornuet, J.-M., Marin, J.-M., and Pillai, N.S. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proc. Natl. Acad. Sci. USA* 108, 15112–15117.
- Rong, Z., Leitao, E., Popplewell, J., Alp, B., and Vadgama, P. (2008). Needle enzyme electrode for lactate measurement in vivo. *IEEE Sens. J.* 8, 113–120.
- Salmon, J.K., Moraes, M.A., Dror, R.O., and Shaw, D.E. (2011). Parallel random numbers—as easy as 1, 2, 3. In *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis (SC '11)* (New York, NY: ACM Press), pp. 16:1–16:12.
- Schaff, J., Fink, C.C., Slepchenko, B., Carson, J.H., and Loew, L.M. (1997). A general computational framework for modeling cellular structure and function. *Biophys. J.* 73, 1135–1146.
- Schaller, G., and Meyer-Hermann, M. (2005). Multicellular tumor spheroid in an off-lattice Voronoi-Delaunay cell model. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 71, 051910.
- Schaller, S., Willmann, S., Lippert, J., Schaupp, L., Pieber, T.R., Schuppert, A., and Eissing, T. (2013). A generic integrated physiologically based whole-body model of the glucose-insulin-glucagon regulatory system. *CPT Pharmacometrics Syst. Pharmacol.* 2, e65.
- Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization* (New York, NY: John Wiley & Sons).
- Silk, D., Filippi, S., and Stumpf, M.P.H. (2013). Optimizing threshold-schedules for sequential approximate Bayesian computation: applications to molecular systems. *Stat. Appl. Genet. Mol. Biol.* 12, 603–618.
- Sisson, S.A., and Fan, Y. (2011). Likelihood-free Markov chain Monte Carlo. In *Handbook of Markov Chain Monte Carlo*, S.P. Brooks, A. Gelman, G. Jones, and X.-L. Meng, eds. (Chapman & Hall/CRC), pp. 319–341.
- Sisson, S.A., Fan, Y., and Tanaka, M.M. (2007). Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* 104, 1760–1765.
- Sottoriva, A., and Tavaré, S. (2010). Integrating approximate Bayesian computation with complex agent-based models for cancer research. In *Proceedings of COMPSTAT 2010*, Y. Lechevallier and G. Saporta, eds. (Physica-Verlag HD), pp. 57–66.
- Sottoriva, A., Kang, H., Ma, Z., Graham, T.A., Salomon, M.P., Zhao, J., Marjoram, P., Siegmund, K., Press, M.F., Shibata, D., and Curtis, C. (2015). A Big Bang model of human colorectal tumor growth. *Nat. Genet.* 47, 209–216.
- Starruß, J., de Back, W., Brusch, L., and Deutsch, A. (2014). Morpheus: a user-friendly modeling environment for multiscale and multicellular systems biology. *Bioinformatics* 30, 1331–1332.
- Stiles, J.R., and Bartol, T.M. (2001). Monte Carlo methods for simulating realistic synaptic microphysiology using MCell. In *Computational Neuroscience: Realistic Modeling for Experimentalists*, E. De Schutter, ed. (Boca Raton, FL: CRC Press), pp. 87–127.
- Swat, M.H., Thomas, G.L., Belmonte, J.M., Shirinifard, A., Hmeljak, D., and Glazier, J.A. (2012). Multi-scale modeling of tissues using CompuCell3D. *Methods Cell Biol.* 110, 325–366.
- ten Tusscher, K.H.W.J., Noble, D., Noble, P.J., and Panfilov, A.V. (2004). A model for human ventricular tissue. *Am. J. Physiol. Heart Circ. Physiol.* 286, H1573–H1589.
- Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T.S., Matsuzaki, Y., Miyoshi, F., Saito, K., Tanida, S., Yugi, K., Venter, J.C., and Hutchison, C.A., 3rd (1999). E-CELL: software environment for whole-cell simulation. *Bioinformatics* 15, 72–84.
- Toni, T., and Stumpf, M.P.H. (2010). Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics* 26, 104–110.
- Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M.P.H. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* 6, 187–202.
- Toni, T., Jovanovic, G., Huvet, M., Buck, M., and Stumpf, M.P.H. (2011). From qualitative data to quantitative models: analysis of the phage shock protein stress response in *Escherichia coli*. *BMC Syst. Biol.* 5, 69.
- Trayanova, N.A. (2011). Whole-heart modeling: applications to cardiac electrophysiology and electromechanics. *Circ. Res.* 108, 113–128.
- Waclaw, B., Bozic, I., Pittman, M.E., Hruban, R.H., Vogelstein, B., and Nowak, M.A. (2015). A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. *Nature* 525, 261–264.
- Walpole, J., Papin, J.A., and Peirce, S.M. (2013). Multiscale computational models of complex biological systems. *Annu. Rev. Biomed. Eng.* 15, 137–154.

## STAR★METHODS

### KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Growth curves and radial profiles of histological stainings	Jagiella et al., 2016	<a href="https://github.com/ICB-DCM/pABC-SMC">https://github.com/ICB-DCM/pABC-SMC</a>
Software and Algorithms		
MATLAB (including the Statistics Toolbox)	Mathworks	<a href="https://www.mathworks.com/">https://www.mathworks.com/</a>
Implementation of the 2D and 3D agent-based model of tumor spheroid growth	Jagiella et al., 2016	<a href="https://github.com/ICB-DCM/pABC-SMC">https://github.com/ICB-DCM/pABC-SMC</a>
Grid-specific implementation of the parallel Approximate Bayesian Computing Sequential Monte Carlo (pABC SMC)	This paper	<a href="https://github.com/ICB-DCM/pABC-SMC">https://github.com/ICB-DCM/pABC-SMC</a>

### CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for software and algorithms should be directed to the Lead Contact Jan Hasenauer ([jan.hasenauer@helmholtz-muenchen.de](mailto:jan.hasenauer@helmholtz-muenchen.de)).

### EXPERIMENTAL MODEL AND SUBJECT DETAILS

#### Growth curves and histological imaging data

We study growth curves and histological imaging data collected under up to four experimental conditions with different glucose and oxygen concentrations. The data have been collected, processed and published by Jagiella et al. (2016).

The growth curves provide the measured radius of spheroids  $r^m(t_{g,k})$  at time points  $t_{g,1}, \dots, t_{g,n_g}$ . The histological imaging data provide the spatially resolved fraction of proliferating cells, necrotic cells and the extracellular matrix intensity. To obtain informative summary statistics, we computed the average fraction of proliferating and necrotic cells as well as the average extracellular matrix intensity at different distances  $d_1, \dots, d_{n_d}$  from the spheroid rim. This yields the fraction of proliferating and necrotic cells,  $p^m(t_{h,k}, d_l)$  and  $n^m(t_{h,k}, d_l)$ , as well as the extracellular matrix intensity,  $e^m(t_{h,k}, d_l)$ , at distances  $d_1, \dots, d_{n_d}$  and time points  $t_{h,1}, \dots, t_{h,n_h}$ . The superscript  $m$  indicates a measured value while the subscripts  $g$  and  $h$  indicates growth curve and histological data, respectively. Accordingly, the number of measured time points for the growth curves and the histological experiments are denoted by  $n_g$  and  $n_h$ . In addition, the number of distances is denoted by  $n_d$ .

For the histological imaging data at most two replicates were available. Accordingly, the estimates of the standard deviations included in the figures were unreliable and were not used for statistical inference.

### METHOD DETAILS

#### Hybrid discrete-continuum model for tumor spheroid growth

We consider a stochastic multi-scale model describing *in-vitro* tumor growth. The model exploits an individual-based description of tumor cells and a continuum-based description of key metabolites, extracellular matrix and waste material from cellular debris of necrotic cells. Individual cells are modeled by agents which can sense their environment, move, divide and die. Furthermore, these agents interact directly via cell-cell contact and indirectly via uptake/secretion of extracellular substances. The dynamics of extracellular matrix, the key metabolites and waste material are modeled using partial differential equations. The model we consider is based on our own previous work (Jagiella, 2012) and will be introduced in the following.

*Notation:*  $H[x]$  denotes the Heaviside step function evaluated at  $x$ , with

$$H[x] = \begin{cases} 0 & \text{for } x < 0, \\ 1 & \text{for } x \geq 0. \end{cases}$$

Furthermore, we denote the second derivative with respect to spatial coordinate  $x$  – the Laplace operator – by  $\nabla_x$ .

#### Individual-based description of single-cell dynamics

The agent-based model considers proliferating, quiescent and necrotic cells populating a static unstructured lattice. Each lattice site can be occupied by at most one cell. The behavior of a cell located at site  $x$  can depend on the time-dependent local concentrations of extracellular matrix  $e(t, x)$ , glucose  $g(t, x)$ , oxygen  $o(t, x)$ , lactate  $l(t, x)$ , adenosine triphosphate  $a(t, x)$  and waste material from debris of necrotic cells  $w(t, x)$  as well as the distance  $L(t, x)$  to the next vacant lattice site.

**Proliferating cells** progress in a discretised cell cycle with  $m_d$  stages,  $m = 1, \dots, m_d$ . The transition from stage  $m$  to stage  $m + 1$  occurs with propensity

$$k_{\text{div},m}(t, x) = k_{\text{max}}^{\text{div}} m_d \left( 1 - \frac{1}{2} \text{H}[o^{\text{div}} - o(t, x)] \right) \left( 1 - \frac{1}{2} \text{H}[w(t, x) - w^{\text{div}}] \right),$$

with maximal division rate  $k_{\text{max}}^{\text{div}}$ , oxygen division threshold  $o^{\text{div}}$  and waste division threshold  $w^{\text{div}}$ . This transition propensity increases with the availability of oxygen and decreases in the presence of waste. As stage  $m = m_d$  is reached, the cell grows and occupies an adjacent lattice site. If no adjacent lattice site is vacant, the neighboring cells are pushed along the shortest path toward the closest vacant lattice site. For  $m = m_d$ , the cell divides into two daughter cells. An individual daughter cell decides to proliferate with probability.

An individual daughter cell decides to proliferate with probability

$$p_{\text{re}}(t, x) = \exp \left[ -\frac{L(t, x)}{L^{\text{div}}} \right] \text{H}[e(t, x) - e^{\text{div}}] \text{H}[k_a(t, x) - k_a^{\text{div}}] \text{H}[l(t, x) - l^{\text{div}}] \text{H}[n_{w,o,\text{max}} - n_{w,o}(t, x)]$$

and otherwise becomes quiescent. Proliferating daughter cells start in the first cell cycle stage,  $m = 1$ . The probability to become a proliferating cell depends on the distance to the next free lattice side  $L(t, x)$ , the local concentrations of extracellular matrix  $e(t, x)$ , the local concentration of lactate  $l(t, x)$ , the ATP synthesis rate  $k_a(t, x) = 2q_g(t, x) + (17/3)q_o(t, x)$  as well as on the time the cell was deprived of oxygen or exposed to waste material  $n_{w,o}(t, x)$ . Parameters are the division depth  $L^{\text{div}}$ , the ECM division threshold  $e^{\text{div}}$ , the lactate division threshold  $l^{\text{div}}$ , the ATP synthesis division threshold  $k_a^{\text{div}}$  as well as the maximum number of cell cycles under waste exposure/oxygen deprivation  $n_{w,o,\text{max}}$ . The time of oxygen deprivation and waste exposure is calculated as

$$n_{w,o}(t, x) = \int_0^t 1 - \text{H}[w^{\text{div}} - w(\tau, x(\tau))] \text{H}[o^{\text{div}} - o(\tau, x(\tau))] d\tau$$

in which  $x(\tau)$  denotes the time-dependent spatial location of the cell located at time  $t$  at position  $x$ . The ATP synthesis rate depends on the local glucose and oxygen consumptions,  $q_g(t, x)$  and  $q_o(t, x)$ , which are defined below.

**Quiescent cells** are arrested in cell cycle but can reenter cell cycle and become proliferating cells with stage  $m = 1$ . A quiescent cell attempts to reenter the cell cycle with propensity

$$k_{\text{re}}(t, x) = k_{\text{max}}^{\text{re}} \left( 1 - \frac{1}{2} \text{H}[w(t, x) - w^{\text{div}}] \right) \left( 1 - \frac{1}{2} \text{H}[o^{\text{div}} - o(t, x)] \right)$$

and succeed with probability  $p_{\text{re}}(t, x)$ . The maximal reentry rate is denoted by  $k_{\text{max}}^{\text{re}}$ .

**Necrotic cells** emerge from proliferating and quiescent cells with propensity

$$k_{\text{nec}}(t, x) = k_{\text{max}}^{\text{nec}} \text{H}[k_a^{\text{nec}} - k_a(t, x)] \frac{l(t, x)^2}{l(t, x)^2 + (l^{\text{nec}})^2}$$

which at low ATP synthesis levels increases with the local lactate concentration. The ATP synthesis necrosis threshold and lactate necrosis threshold are  $k_a^{\text{nec}}$  and  $l^{\text{nec}}$ , respectively. Necrotic cells are lysed with constant propensity  $k_{\text{lys}}$  and afterward removed from the corresponding lattice site.

The initial cell population at time point  $t = 0$  occupies all lattice sites within a sphere of radius  $L^{\text{init}}$  around the center of the unstructured lattice. The individual cells are quiescent with probability  $q^{\text{init}}$  and otherwise proliferating (with  $m = 1$ ).

A detailed discussion of the transition propensities and reentering probabilities is provided in (Jagiella, 2012; Jagiella et al., 2016). Precise numerical values for the thresholds ( $e^{\text{div}}$ ,  $k_a^{\text{div}}$ ,  $k_a^{\text{nec}}$ ,  $l^{\text{div}}$ ,  $l^{\text{nec}}$ ,  $w^{\text{div}}$ ,  $n_{w,o,\text{max}}$ , and  $L^{\text{div}}$ ) at which cells change their behavior as well as the properties of the initial cell population ( $L^{\text{init}}$  and  $q^{\text{init}}$ ) are mostly unknown. The considered parameter regimes are provided below.

### Continuum-based description of the dynamics of extracellular substances

The dynamics of the extracellular molecular species are governed by a system of partial differential equations (PDEs), accounting for different processes. In the following, we describe the models for the individual extracellular substances and the coupling to the single-cell dynamics.

**Glucose and oxygen**, the primary energy sources, are subject to diffusive transport and consumption,

$$\frac{\partial g(t, x)}{\partial t} = D_g \nabla_x g(t, x) - q_g(t, x) c(t, x), \quad \text{with } q_g(t, x) = V_{m,g}(t, x) \frac{g(t, x)}{g(t, x) + k_{m,g}},$$

$$\frac{\partial o(t, x)}{\partial t} = D_o \nabla_x o(t, x) - q_o(t, x) c(t, x), \quad \text{with } q_o(t, x) = V_{m,o}(t, x) \frac{o(t, x)}{o(t, x) + k_{m,o}},$$

with diffusion coefficients  $D_g$  and  $D_o$ , maximum consumption rates  $V_{m,g}(t, x)$  and  $V_{m,o}(t, x)$ , and Michaelis-Menten constants  $k_{m,g}$  and  $k_{m,o}$ . Cells lacking one of the metabolites, glucose or oxygen, were observed to compensate for it by upregulating the consumption rates of the other one in order to keep the net production of ATP molecules constant. The maximum consumption rates of glucose  $V_{m,g}(t, x)$  and oxygen  $V_{m,o}(t, x)$  account for these cross-dependencies,

$$V_{m,g}(t, x) = q_g^{\max} \left( 1 - \left( 1 - \frac{q_g^{\min}}{q_g^{\max}} \right) \frac{o(t, x)}{o(t, x) + k_o} \right),$$

$$V_{m,o}(t, x) = q_o^{\max} \left( 1 - \left( 1 - \frac{q_o^{\min}}{q_o^{\max}} \right) \frac{g(t, x)}{g(t, x) + k_g} \right),$$

with consumption parameters  $q_g^{\min}$ ,  $q_g^{\max}$ ,  $k_g$ ,  $q_o^{\min}$ ,  $q_o^{\max}$ , and  $k_o$ . As glucose and oxygen are merely consumed by proliferating and quiescent cells, we introduce the indicator function  $c(t, x)$  which is 1 if a proliferating or a quiescent cell occupies  $x$  at time point  $t$  and 0 otherwise. The Michaelis-Menten constants and the consumption parameters are available from the literature (see (Jagiella, 2012) and references therein) and listed below. Glucose and oxygen enter the simulation domain  $\Omega$  from the surrounding medium and we assume Dirichlet boundary condition,  $g(t, x) = g_0$  and  $o(t, x) = o_0$  for  $x \in \partial\Omega$ . Initially, glucose and oxygen concentrations are equivalent to this boundary conditions,  $g(0, x) = g_0$  and  $o(0, x) = o_0$  for  $x \in \Omega$ .

**Lactate** is a by-product of the anaerobic energy metabolism. It is produced by proliferating and quiescent cells with rate  $2(q_g(t, x) + \min\{q_g(t, x), 1/6q_o(t, x)\})$  and diffuses. This leads to the model

$$\frac{\partial l(t, x)}{\partial t} = D_l \nabla_x l(t, x) + 2 \left( q_g(t, x) + \min \left\{ q_g(t, x), \frac{1}{6} q_o(t, x) \right\} \right) c(t, x).$$

We assume that lactate dilutes and zero Dirichlet boundary conditions,  $l(t, x) = 0$  for  $x \in \partial\Omega$ . At the start of the experiment, the lactate concentration is zero everywhere,  $l(0, x) = 0$  for  $x \in \Omega$ .

**Extracellular matrix** is a collection of extracellular molecules. The extracellular matrix provides structural support for cells and is involved in cell adhesion as well as cell-to-cell communication. The components of the extracellular matrix are synthesized and secreted by cells and can be degraded. This yields the governing equations for the dynamics of the concentration of extracellular matrix,

$$\frac{\partial e(t, x)}{\partial t} = k_e^{\text{pro}} c(t, x) - k_e^{\text{deg}} e(t, x).$$

The production rate  $k_e^{\text{pro}}$  and degradation rate  $k_e^{\text{deg}}$  are assumed to be constant. Note that the production rate  $k_e^{\text{pro}}$  as well as the division threshold  $k_e^{\text{deg}}$  is in units of intensity, as the absolute extracellular matrix concentration cannot be assessed experimentally. The boundary and initial concentration of extracellular matrix are assumed to be zero,  $e(t, x) = 0$  for  $x \in \partial\Omega$  and  $e(0, x) = 0$  for  $x \in \Omega$ .

**Waste materials** are produced by necrotic cells and absorbed by living cells with a constant rate. Accordingly, the evolution equation for the waste concentration is

$$\frac{\partial w(t, x)}{\partial t} = k_w^{\text{pro}} c^{\text{nec}}(t, x) - k_w^{\text{upt}} w(t, x) c(t, x)$$

with the indicator function  $c^{\text{nec}}(t, x)$  being 1 if a necrotic cell occupies  $x$  at time point  $t$  and 0 otherwise. Waste production and uptake rates are denoted by  $k_w^{\text{pro}}$  and  $k_w^{\text{upt}}$ . As initially merely proliferation and quiescent cells are present, the initial waste concentration is zero,  $w(0, x) = 0$  for  $x \in \Omega$ . Furthermore, as waste is not transported and as no cells are at the boundary, we use zero Dirichlet boundary conditions,  $w(t, x) = 0$  for  $x \in \partial\Omega$ .

A detailed list of the boundary conditions for the different scenarios and experimental conditions is provided below.

### Numerical simulation

To simulate the individual scenarios we exploit a hybrid approach. The cellular dynamics are simulated using Gillespie's algorithm (Gillespie, 1977), which accounts for the stochasticity of cellular processes and decision making. The PDEs governing the spatio-temporal evolution of glucose, oxygen, lactate, extracellular matrix and waste concentration are discretised using finite differences and solved using an implicit scheme.

We use this hybrid simulation approach to study two scenarios:

**Scenario I – no nutrition limitation:** Glucose and oxygen are assumed to be available in excess. Hence, the propensities for the cellular dynamics simplify and neither lactate nor waste material is produced. Extracellular matrix dynamics are still modeled using the aforementioned PDE. Cells becoming quiescent are assumed to be permanently arrest in  $G_0$  phase. To capture this scenario, different parameters are set to zero or infinity, effectively reducing the dimensionality of the PDE system.

Scenario I is studied in the section *Performance and reliability of pABC SMC algorithm* and section *Consistency of parameter estimates for 2D and 3D model* of the main manuscript. The reference parameters used for the generation of artificial data and the lower and upper bounds are used for statistical inference are:

NAME	SYMBOL	UNIT	REFERENCE VALUE	LOWER BOUND	LOWER BOUND
Division rate	$k_{\max}^{\text{div}}$	1/h	$4.17 \times 10^{-2}$	$10^{-3}$	$10^{-1}$
Division depth	$L^{\text{div}}$	$\mu\text{m}$	$10^2$	$10^1$	$10^3$
Initial spheroid radius	$L^{\text{init}}$	$\mu\text{m}$	$1.2 \times 10^1$	$10^0$	$1.59 \times 10^1$
Initial quiescent cell fraction	$q^{\text{init}}$	-	$7.5 \times 10^{-1}$	$10^{-5}$	$10^0$
ECM production rate	$k_e^{\text{pro}}$	au/h	$5.0 \times 10^{-3}$	$10^{-5}$	$10^0$
ECM degradation rate	$k_e^{\text{deg}}$	1/h	$8.0 \times 10^{-4}$	$10^{-5}$	$10^0$
ECM division threshold	$e^{\text{div}}$	au	$10^{-2}$	$10^{-5}$	$10^0$

**Scenario II – nutrition limitation:** Glucose and oxygen are potentially limiting and all afore-described variables are simulated. Due to more possible reasons for cells to end up in  $G_0$ , we in addition allow them to reenter the cell cycle with rate  $k_{re}$ . We considered four experimental conditions with different glucose and oxygen concentrations.

Scenario II is studied in the section *Multi-experiment Data Integration* and section *Uncertainty-aware Prediction of Tumor Spheroid Growth* of the main manuscript. The lower and upper bounds for statistical inference are derived from the reference value  $\theta^{\text{ref}}$  provided by (Jagiella, 2012),  $\theta_{i,\min} = 10^{-2} \times \theta_i^{\text{ref}}$  and  $\theta_{i,\max} = 10^2 \times \theta_i^{\text{ref}}$  and are:

NAME	SYMBOL	UNIT	LOWER BOUND	LOWER BOUND
Division rate	$k_{\max}^{\text{div}}$	1/h	$3.2 \times 10^{-4}$	$3.2 \times 10^0$
Division depth	$L^{\text{div}}$	$\mu\text{m}$	$1.3 \times 10^0$	$1.3 \times 10^4$
Initial spheroid radius	$L^{\text{init}}$	$\mu\text{m}$	$1.2 \times 10^{-1}$	$1.2 \times 10^3$
Initial quiescent cell fraction	$q^{\text{init}}$	-	$7.5 \times 10^{-3}$	$7.5 \times 10^1$
ECM production rate	$k_e^{\text{pro}}$	au/h	$5.0 \times 10^{-6}$	$5.0 \times 10^{-2}$
ECM degradation rate	$k_e^{\text{deg}}$	1/h	$3.3 \times 10^{-5}$	$3.3 \times 10^{-1}$
ECM division threshold	$e^{\text{div}}$	au	$3.0 \times 10^{-5}$	$3.0 \times 10^{-1}$
Cell cycle reentrance rate	$k_{\max}^{\text{re}}$	1/h	$10^{-5}$	$10^{-1}$
Necrosis rate	$k^{\text{nec}}$	1/h	$10^{-4}$	$10^0$
Lysis rate	$k^{\text{lys}}$	1/h	$10^{-4}$	$10^0$
ATP synthesis division threshold	$k_a^{\text{deg}}$	mM/h	$9.0 \times 10^0$	$9.0 \times 10^4$
ATP necrosis division threshold	$k_a^{\text{nec}}$	mM/h	$6.0 \times 10^0$	$6.0 \times 10^4$
Lactate division threshold	$l^{\text{div}}$	mM	$2.0 \times 10^{-1}$	$2.0 \times 10^3$
Lactate necrosis threshold	$l^{\text{nec}}$	mM	$2.0 \times 10^{-1}$	$2.0 \times 10^3$
Waste diffusion coefficient	$D_w$	$\mu\text{m}^2/\text{h}$	$10^3$	$10^7$
Waste degradation rate	$k_w^{\text{upt}}$	1/h	$10^{-8}$	$10^{-4}$
Waste division threshold	$w^{\text{div}}$	mM	$8.0 \times 10^{-5}$	$8.0 \times 10^{-1}$
Maximum number of cell cycles under waste exposure / oxygen deprivation	$n_{o,w,\max}$	-	$8.0 \times 10^{-2}$	$8.0 \times 10^2$

For Scenario I (no nutrient limitation) and Scenario II (nutrient limitation), some model parameters are fixed to previously published values:

NAME	SYMBOL	UNIT	VALUE	REFERENCE
Oxygen diffusion coefficient	$D_o$	$\mu\text{m}^2/\text{h}$	$6.3 \times 10^6$	(Schaller and Meyer-Hermann, 2005)
Glucose diffusion coefficient	$D_g$	$\mu\text{m}^2/\text{h}$	$3.78 \times 10^5$	(Schaller and Meyer-Hermann, 2005)
Lactate diffusion coefficient	$D_l$	$\mu\text{m}^2/\text{h}$	$7.56 \times 10^6$	(Rong et al., 2008)
Glucose uptake	$k_{m,g}$	$\text{mM}$	$6.8 \times 10^{-2}$	(Jagiella et al., 2016)
	$k_o$	$\text{mM}$	$3.1 \times 10^{-2}$	
	$q_g^{\text{min}}$	$\text{mM}/\text{h}$	$1.87 \times 10^2$	
	$q_g^{\text{max}}$	$\text{mM}/\text{h}$	$7.07 \times 10^2$	
Oxygen uptake	$k_{m,o}$	$\text{mM}$	$3.1 \times 10^{-2}$	(Jagiella et al., 2016)
	$k_g$	$\text{mM}$	$1.0 \times 10^{-1}$	
	$q_o^{\text{min}}$	$\text{mM}/\text{h}$	$1.2 \times 10^2$	
	$q_o^{\text{max}}$	$\text{mM}/\text{h}$	$3.07 \times 10^2$	
Cycle steps until growth	$m_g$	-	2	(Jagiella et al., 2016)
Cycle steps until division	$m_d$	-	10	(Jagiella et al., 2016)
Oxygen division threshold	$o^{\text{div}}$	$\text{mM}$	$7.0 \times 10^{-2}$	(Jagiella et al., 2016)

The boundary conditions for molecular species are.

MOLECULE	SCENARIO I	SCENARIO II			
		CONDITION I	CONDITION II	CONDITION III	CONDITION IV
Glucose, $g$	*	25 mM	5 mM	1 mM	25 mM
Oxygen, $o$	*	0.28 mM	0.28 mM	0.28 mM	0.07 mM
Lactate, $l$	*	0 mM	0 mM	0 mM	0 mM
ECM, $e$	0 mM	0 mM	0 mM	0 mM	0 mM
Waste, $w$	0 mM	0 mM	0 mM	0 mM	0 mM

In Scenario I, glucose and oxygen are available in excess and there is no lactate. This is indicated by an asterisk, \*  
The hybrid discrete-continuum model for tumor spheroid growth is implemented in C++.

## QUANTIFICATION AND STATISTICAL ANALYSIS

### Parallel Approximate Bayesian Computing Sequential Monte Carlo method

For this study we developed a simple parallelised version of the ABC SMC method introduced by (Toni et al., 2009). The master node runs the main routine which iteratively samples  $T$  generations,  $t = 0$  to  $t = T - 1$ , with decreasing thresholds,  $\epsilon_0 > \dots > \epsilon_{T-1}$ . To exploit multiple cores, candidate parameters are evaluated in parallel. To ensure convergence of the sampling to the true posterior, the main routine keeps track of the order of the candidate parameters. Only if the evaluation for the candidate parameters  $j = 1, \dots, J$  is finished and if these candidate parameters resulted in  $N$  accepted points, the algorithm continues with the next generation (Figure 2). The pseudocode of the main routine is:

#### Main routine: pABC SMC

In: Number of generations  $T$ , number of samples per generation  $N$  and number of available computing cores  $C$ .

S1 Set the generation indicator  $t = 0$ .

Set the initial threshold  $\epsilon_0 = \infty$ .

S2 Set the candidate number  $j = 1$ .

S3 If number of jobs on queue is below or drops below  $C$ , determine from stored files the smallest candidate number  $J + 1$  for which no results are available.

- If number of accepted candidates in the set  $j = 1, \dots, J$  is  $N$ , load parameters  $\{\theta_{t-1}^{(i)}\}_{i=1}^N$  and unnormalized weights  $\{w_{t-1}^{(i)}\}_{i=1}^N$  and normalize the weights.
- Else, start new job on computing cluster by executing the subroutine **getSample**( $t, \epsilon_t, j, \{\theta_{t-1}^{(i)}\}_{i=1}^N, \{w_{t-1}^{(i)}\}_{i=1}^N$ ), set  $j = j + 1$  and go to S3.

S4 If  $t < T$  set  $t = t + 1$ ,  $\varepsilon_t = f(\{d_{t-1}^{(i)}\}_{i=1}^N)$  and go to S2.

Else, stop algorithm and output results.

Out: Samples  $\{\theta_t^{(i)}\}_{i=1}^N$  and weights  $\{w_t^{(i)}\}_{i=1}^N$ .

The main routine calls a subroutine which runs on a slave and initiates an individual sample. In generation  $t$  a sample from generation  $t - 1$  is selected and perturbed using the perturbation kernel  $K_t(\theta | \theta')$  to obtain a new candidate parameter. If a stochastic simulation using these candidate parameters yields a distance  $d$  between simulation and data below the threshold  $\varepsilon_t$ , this candidate is accepted. Otherwise, it is rejected. The pseudocode for this subroutine is:

**Subroutine: Sampling and evaluation of candidate parameter, getSample (·)**

In: Generation number  $t$ , threshold  $\varepsilon_t$ , candidate number  $j$ , sample  $\{\theta_{t-1}^{(i)}\}_{i=1}^N$  and weights  $\{w_{t-1}^{(i)}\}_{i=1}^N$  from previous generation.

S1 If  $t = 0$ , sample candidate parameter  $\theta^*$  independently from the prior,  $\theta^* \sim p(\theta)$ .

Else, sample  $\theta'$  from the previous generation  $\{\theta_{t-1}^{(i)}\}_{i=1}^N$  with probabilities  $\{w_{t-1}^{(i)}\}_{i=1}^N$  and perturb it to obtain the candidate parameter  $\theta^* \sim K_t(\theta | \theta')$ . If prior probability of  $\theta^*$  is zero,  $p(\theta^*) = 0$ , return to S1.

S2 Sample candidate dataset  $\mathcal{D}^*$  by simulating the model,  $\mathcal{D}^* \sim p(\mathcal{D} | \theta^*)$ .

S3 Create a file indicating the generation and candidate number,  $t$  and  $j$ , and write the parameter candidate  $\theta^*$ , distance  $d(\mathcal{D}^*, \mathcal{D}, \infty)$  and weight

$$w_{t-1}^{(j)} = \begin{cases} 1, & \text{if } t = 0, \\ \frac{p(\theta^*)}{\sum_{i=1}^N w_{t-1}^{(i)} K_t(\theta_{t-1}^{(i)} | \theta^*)}, & \text{otherwise,} \end{cases}$$

in the file created.

In order to increase computational efficiency, we stop the model simulation in step S2 as soon as the threshold  $\varepsilon_t$  is reached. In this case,  $d(\mathcal{D}^*, \mathcal{D}, \infty) > \varepsilon_t$  is returned. This is possible as we use a distance  $d(\mathcal{D}^*, \mathcal{D}, t_{\text{sim}}) > \varepsilon_t$  which monotonically increases in the simulation time  $t_{\text{sim}}$ .

**Distance measure**

In this study, we consider artificial and measured data for

- the time-dependent spheroid radius,  $r^m(t_{g,k})$ ,
- the time-dependent fraction of proliferation cells at different distances from the spheroid rim,  $\rho^m(t_{h,k}, d_l)$ ,
- the time-dependent fraction of necrotic cells at different distances from the spheroid rim,  $n^m(t_{h,k}, d_l)$ , and
- the time-dependent ECM intensity at different distances from the spheroid rim,  $e^m(t_{h,k}, d_l)$ .

The artificial and measured data are the averages over all available replicates.

We use as distance measure the sum of weighted least-squares,

$$\begin{aligned} d(\mathcal{D}^*, \mathcal{D}, t_{\text{sim}}) = & \frac{1}{n_g} \sum_{k=1}^{n_g} H[t_{\text{sim}} - t_{g,k}] w_{k,l}^r (r^m(t_{g,k}) - r(t_{g,k}, \theta^*))^2 \\ & + \frac{1}{n_h n_d} \sum_{k=1}^{n_h} H[t_{\text{sim}} - t_{h,k}] \sum_{l=1}^{n_d} w_{k,l}^{\rho} (\rho^m(t_{h,k}, d_l) - \rho(t_{h,k}, d_l, \theta^*))^2 \\ & + \frac{1}{n_h n_d} \sum_{k=1}^{n_h} H[t_{\text{sim}} - t_{h,k}] \sum_{l=1}^{n_d} w_{k,l}^n (n^m(t_{h,k}, d_l) - n(t_{h,k}, d_l, \theta^*))^2 \\ & + \frac{1}{n_h n_d} \sum_{k=1}^{n_h} H[t_{\text{sim}} - t_{h,k}] \sum_{l=1}^{n_d} w_{k,l}^e (e^m(t_{h,k}, d_l) - e(t_{h,k}, d_l, \theta^*))^2 \end{aligned}$$

in which the simulation results for a proposed parameter  $\theta^*$  are denoted by  $r(t_{g,k}, \theta^*)$ ,  $\rho(t_{h,k}, d_l, \theta^*)$ ,  $n(t_{h,k}, d_l, \theta^*)$  and  $e(t_{h,k}, d_l, \theta^*)$  and the weights are denoted by  $w_{k,l}^r$ ,  $w_{k,l}^{\rho}$ ,  $w_{k,l}^n$ , and  $w_{k,l}^e$ . The sums in the individual lines penalize the error in the spheroid radius, the error in the fraction of proliferation cells, the error in the fraction of necrotic cells and the error in the ECM intensity, respectively. All contributions are normalized with the corresponding number of measurements to facilitate an equal weighting of different datasets. As the

simulation is run till time point  $t_{\text{sim}}$ , merely measurements with  $t_k > t_{\text{sim}}$  are considered. For  $t_{\text{sim}} > \max\{t_{n_g}, t_{n_n}\}$ , all measurement data are considered. The final distance is denoted by  $d(\mathcal{D}^*, \mathcal{D}, \infty)$ .

For the artificial data, the number of replicates is sufficiently high to obtain robust estimates of the standard deviations of individual observations. Accordingly, we set the weights to  $w_k^r$ ,  $w_{k,J}^p$ ,  $w_{k,J}^n$ , and  $w_{k,J}^e$  to inverses of the squared standard deviations. For the measured data the number of replicates is too small – for some settings only two – to compute robust estimates of the standard deviations. Therefore, we set the weights to inverse of the squared dynamic rang of the signal,

$$w_k^r = \frac{1}{R_r^2} \text{ with } R_r = \max_{k'} r^m(t_{g,k'}) - \min_{k'} r^m(t_{g,k'}),$$

$$w_{k,J}^p = \frac{1}{R_p^2} \text{ with } R_p = \max_{k',J'} p^m(t_{g,k'}, d_{J'}) - \min_{k',J'} p^m(t_{g,k'}, d_{J'}),$$

$$w_{k,J}^n = \frac{1}{R_n^2} \text{ with } R_n = \max_{k',J'} n^m(t_{g,k'}, d_{J'}) - \min_{k',J'} n^m(t_{g,k'}, d_{J'}),$$

$$w_{k,J}^e = \frac{1}{R_e^2} \text{ with } R_e = \max_{k',J'} e^m(t_{g,k'}, d_{J'}) - \min_{k',J'} e^m(t_{g,k'}, d_{J'}).$$

The use of these weights yields dimensionless residuals and should facilitate the comparability of residuals associate to different observables.

*Remark:* ABC methods are to a certain degree robust with respect to the choice of the distance measure. For a detailed discussion we refer to (Toni et al., 2009; Toni and Stumpf, 2010; Nunes and Balding, 2010).

In parts of the manuscript, several experimental conditions are considered simultaneously. In this case, the overall distance  $d$  is the sum of the distances for the individual conditions.

### Adaptation of perturbation kernel and threshold

The efficiency of ABC SMC methods depends critically on the perturbation kernels (Filippi et al., 2013) and the threshold sequences (Silk et al., 2013). To facilitate the applicability of the algorithms to a wide range of inference problems, we implemented adaptive methods. As perturbation kernel in generation  $t$  we use a multi-variate normal distribution,

$$K_t(\theta | \theta') = N(\theta | \theta', \Sigma_t),$$

with covariance matrix

$$\Sigma_t = N^{-\frac{2}{n_\theta+4}} C_{t-1}$$

Here  $n_\theta$  denotes the number of parameters and  $C_{t-1}$  denotes the sample covariance matrix of generation  $t - 1$ ,

$$C_{t-1} = \frac{1}{N-1} \sum_{i=1}^N (\theta_{t-1}^{(i)} - m_{t-1}) (\theta_{t-1}^{(i)} - m_{t-1})^T \text{ with } m_{t-1} = \sum_{i=1}^N \theta_{t-1}^{(i)}.$$

The choice of the proposal covariance matrix  $\Sigma_t$  is inspired by kernel density estimation, namely, Scott's rule (Scott, 1992). This perturbation kernel adapts to the correlation structure of the sample, thereby improving the representation of the distribution. The threshold for generation  $t$  is set to the median of the accepted distances in generation  $t - 1$ .

### Parameterization, prior distribution and parameter bounds

In this manuscript we sample the log-transformed parameter  $\xi_i = \log(\theta_i)$  instead of the parameter  $\theta_i$ . Previous studies revealed that this improves the computational efficiency (Raue et al., 2013; Hug et al., 2013). For the log-transformed parameters  $\xi_i$  we used lower and upper bounds which are consistent with previous publications. To account for the large uncertainty of the model parameters, we assumed uniform prior distributions for the log-transformed parameters  $\xi_i$  between lower and upper bounds.

### Population size and analysis of convergence

In the manuscript we employed population size of  $N = 100$  and  $N = 1000$ . These population sizes are rather low but proved to be appropriate for the respective problems in a series of test scenarios. The use of a large population size would increase the robustness and the accuracy of the method, however, the computation time increases proportionally with the population size.

The convergence of pABC SMC and the sufficiency of the population size  $N$  was monitored manually by assessing

- the inter-quantile ranges and the objective function values of subsequent populations and
- the effective sample size.

The effective sample size assessed using the approximation by Kong et al. (1994) and Liu (1996),

$$\text{EES}_t = \frac{N}{1 + \text{var}(w_t^{(j)})} \quad \text{with} \quad \text{var}(w_t^{(j)}) = \frac{1}{N-1} \sum_{j=1}^N (w_t^{(j)} - 1)^2$$

with normalized weights  $w_t^{(j)}$ . Complementary to the online evaluation, we performed for the 2D model test runs with altered population size  $N$  to ensure that  $N$  was sufficiently high.

### Analysis of parameter and prediction uncertainties

The pABC SMC algorithm provided a parameter sample  $\{\theta^{(i)}\}_{i=1}^N$  and a corresponding sample of simulation results. The uncertainty of parameters and simulation results was assessed by evaluating the (Bayesian) confidence intervals, more precisely, the percentile intervals of the samples. The confidence regions for parameter pairs were computed using kernel density estimation and subsequent thresholding. To assess the parameter uncertainty analysis in the high-dimensional space, we carried out a Principal Component Analysis (PCA) and the MATLAB Statistics Toolbox.

### Prediction of spheroid growth characteristics

We used the 2D model to predict the depth of the proliferating zone, the depth of the viable rim zone and the initial growth rate. We computed the depth of the proliferating zone by evaluating the percentage of proliferating cells at different distances from the spheroid rim and subsequent integration over the distance. The calculation of the depth of the viable zones is performed accordingly by considering the percentage of all cells which are viable (not necrotic). To calculate the initial growth rate, the trajectory of the spheroid radius and the linear growth regime is detected. The observed spheroid radii in the linear regime are fitted with a regression model, providing the initial growth rate.

### Assessment of the importance of individual data types

Beyond the studies discussed in the main manuscript, we employed the pABC SMC algorithm to study for Scenario I the necessity of the different datasets for reliable prediction of tumor spheroid growth. We considered the following datasets:

- Dataset 1: Spheroid radius.
- Dataset 2: Spheroid radius and fraction of proliferating cells.
- Dataset 3: Spheroid radius and ECM abundance.
- Dataset 4: Spheroid radius, fraction of proliferating cells and ECM abundance.

Datasets 1-3 provided reduced sets of information compared to Dataset 4, which has been used in the previous section. The spheroid radius is included in all datasets as it is easy to assess compared to the histological information. We used pABC SMC to estimate the model parameters from the Dataset 1-4. The algorithm was terminated as soon as the acceptance rate dropped substantially.

We found that only a subset of the system properties were predicted correctly if reduced datasets are used for inference (Figure S2). In particular, the model predictions for the fraction of proliferating cells and the ECM abundance were only consistent with the experimental data if the respective datasets were used in the fitting. This indicated that the histological information was essential and that a further reduction of the dataset was not possible. Accordingly, Dataset 4 already provided a minimal dataset for the development of predictive models of tumor spheroid growth.

### Implementation of the statistical analysis

The pABC SMC algorithm and the scripts used for the evaluation were implemented in MATLAB.

*Random number generation:* Sampling methods like the pABC SMC algorithm rely on (pseudo) random number generators. In this study, we employed the default random number generator implemented in MATLAB ('mt1993ar'), which is a Mersenne Twister algorithm (Matsumoto and Nishimura, 1998). This random number generator is a 32-bit multiplicative congruential generator with an approximate period in full precision of  $2^{19937} - 1$ . Mersenne Twister algorithms are widely used in practice. As they do however not pass the CRUSH test in the TestU01 software suite of random number tests (L'Ecuyer and Simard, 2007) and are computationally expensive, we also implemented the Random123 (Salmon et al., 2011) in our routines. Random123 produces better streams of random numbers is easy to parallelise.

## DATA AND SOFTWARE AVAILABILITY

### Data resources

The growth curves and the radial profiles of the histological stainings have been deposited in GitHub (<https://github.com/ICB-DCM/pABC-SMC>).

### Software resources

The code for the simulation and inference has been deposited in GitHub (<https://github.com/ICB-DCM/pABC-SMC>). The implementation of the sampling is tailored to our local grid infrastructure.