Accurate Segmentation of Bacterial Cells using Synthetic Training Data

Georgeos Hardo[1], Maximilian Noka[1], and Somenath Bakshi[1,*]

[1]Department of Engineering, University of Cambridge, CB2 1PZ, UK
[*]Corresponding author: Somenath Bakshi, sb2330@cam.ac.uk

July 2021

Abstract

We present a novel method of bacterial image segmentation using machine learning based on Synthetic Micrographs of Bacteria (SyMBac). SyMBac allows for rapid, automatic creation of arbitrary amounts of training data that combines detailed models of cell growth, physical interactions, and microscope optics to create synthetic images which closely resemble real micrographs, with access to the ground truth positions of cells. We also demonstrate that models trained on SyMBac data generate more accurate and precise cell masks than those trained on human annotated data, because the model learns the true position of the cell irrespective of imaging artefacts.

Main

The performance of any deep-learning algorithm is fundamentally dependent on the quality and volume of training data. However, the task of generating training data with accurate ground truth is not only slow and difficult but a near-impossible one for microscopic images of micron-sized bacterial cells. The point-spread function (PSF) of the microscope corrupts the images of tiny bacterial cells, blurring them to the extent that neither expert users nor traditional computational segmentation programs can accurately annotate the correct pixels. Similarly, geometric impacts on the 2D projection of 3D objects lead to inaccuracies in contrast based segmentation by both humans and computational programs. Inaccuracies in the ground-truth of training data cause deep learning models to misinterpret the relations between the object and its image, resulting in systematically artifactual masks from test data. This limits one's ability to infer a cell's true size and shape and thus biases.

To address this fundamental limitation, we have developed a pipeline to generate realistic micrographs of bacteria growing in a microfluidic device called the mother machine. This enables a neural network trained on this data to precisely learn about the corruption-modes introduced during image formation, and as a result output accurate cell masks. We combine detailed understanding of cell growth and morphology, its position and interaction with the imaging platform, and the microscope's optics to render synthetic micrographs capable of training highly accurate segmentation CNNs without human annotated data. This not only greatly speeds up the process of image segmentation (because no human annotation is necessary), but most importantly generates accurate masks of cells, enabling precise analysis of size regulation and dynamics. Using our method, given any change in experimental settings, it becomes trivial to generate high volumes of synthetic images and ground truths, and rapidly analyse new data. This also solves the robustness problem of image processing using deep-learning because models can now be easily adapted and benchmarked, since synthetic data can be generated at any desired spatial and temporal resolution.

In this paper, we focus our attention to the microfluidic device called the mother machine (Figure 1a), a device that enables very high-throughput time-lapse imaging of bacterial cells to measure their physiology and gene-expression under precisely controlled conditions [1, 2]. Bacterial single cell measurements of size and gene expression in the mother machine have revived the field of quantitative bacterial physiology and revolutionised the field of systems and synthetic microbiology [3, 4, 5, 6, 7]. We tackle the challenging properties of this imaging platform: 1) Phase contrast mother machine data is difficult to segment due to diffractive interference between the device and the cells. 2) The

data is generally very high throughput ($> 10^7$ cells per experiment). Moreover the higher the throughput the lower the magnification and resolution of the images, with cells being often less than 100 pixels in size. Therefore it has previously been challenging to achieve high segmentation accuracy. However, since the mother machine is often used in high throughput studies for detecting rare events [2, 8] or screening libraries of circuits [7], inaccuracies in segmentation are very costly. These challenges can, however, be tackled for 3 main reasons: 1) Owing to its design, the architecture of the mother machine is regular. 2) Cell growth and division is generally predictable. 3) The microscope's optics can be modelled using only a few parameters. This makes the whole system very amenable to simulations, as we illustrate in Figure 1a.

Much work has been done to develop image processing software for the segmentation and tracking of cells, including software specifically for mother machine data. A comparison of the tracking and segmentation algorithms of different state of the art tools are given in [9]. Such a survey reveals that the tools can be categorised into "Machine Learning" (ML) and "traditional"/"Non-ML" approaches. Overall, the ML tools are the most effective pipelines in terms of accuracy and processing speed (excluding time to annotate data). These toolsets, however, do not generalise well to new experimental set-ups, which require new training data to be generated for each application. If an experiment involves multiple strains or changing growth-conditions, which all require their own training data, this can be prohibitive. Further, the final trained algorithm can be biased by the qualities of the original human annotation or segmentation algorithms used to generate the training-set, as is demonstrated in Figure 2f.

The solution to the robustness problem that ML techniques encounter is the need to adapt the models to new experimental setups and datasets. In order to train an ML algorithm for the first time for a new set-up, usually a non-ML technique is used to generate a first order set of training data. This data is then manually curated to fix segmentation errors. This approach has three disadvantages: 1) For good performance of the final model, the training-set must be error free and therefore manual curation is always required, which is time consuming and cumbersome. The authors of the DeLTA pipeline report that curating a dataset of sufficient quality and size (8300 frames) takes two days [10]. 2) The way the training data was segmented will involve bias (either human or machine) or be of low quality, which in turn biases the trained model. 3) If the training data only covers a narrow domain of growth conditions, the accuracy of ML methods can suffer when applied to situations where the statistics behind cell growth and division change, even if the imaging and microfluidic set-up is otherwise the same.

In response to all these problems, we have devloped SyMBac: Synthetic Micrographs of Bacteria. SyMBac is a tool which rapidly generates synthetic training data of bacteria growing in microfluidic devices. The most obvious advantage of synthetic training data is that it can be generated virtually instantly, and on demand (10,000 images in less than 10 minutes fig 1d and f). Another advantage of synthetic data is that it is accompanied by perfect ground truth information. This means large sets of training data can be generated without the need for curation. Additionally, biological ground truth and imaging effects can be tuned independently. Synthetic data can therefore be adapted to different biological conditions, imaging platforms, and imaging modalities (Figure 1e). With real data, many experiments would need to be conducted to gain the same kind of coverage in training data.

In Figure 1b, we show the steps involved in generating synthetic phase contrast data. This involves the generation of a scene using a rigid body physics simulation of cell growth (details in Figure S1). Properties from the simulation are extracted and used to render an optical path length image that signifies relative phase shift in the image—observed as as a change in intensity—due to local refractive index (Figure S2). This image is then convolved with the microscope's point spread function which can be generated using known parameters of the objective lens and phase ring/annulus. These properties are easy to find and thus any change in optics can be readily simulated. The intensity distribution and noise spectrum of the convolved image is then optimised in order to maximise its similarity to the real image. Put together, this enables us to produce realistic phase contrast images of bacterial cells in the mother machine, and even kymographs (Figure 1c). SyMBac takes advantage of multiprocessing, and where possible offloads array operations onto the GPU, which allows it to be approximately 10,000x faster than a human at generating training data (and if deployed on computing clusters, this could be extended by a further order of magnitude). It is important to mention here that SyMBac does not involve any *ad hoc* image augmentation to increase the training data coverage, and therefore maintains perfect relation between the ground truth and the corresponding images.

Most importantly, synthetic images in SyMBac are accompanied by perfect ground truth masks. This plays a key

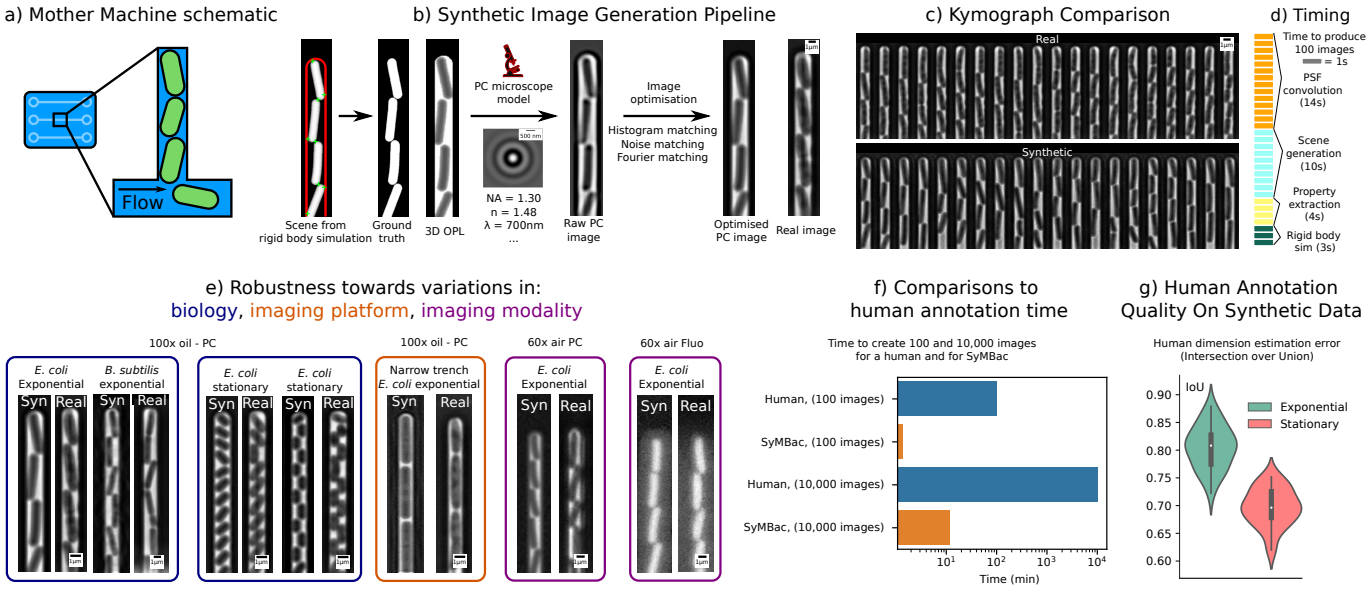


Figure 1: The synthetic image generation process: **a)** Schematic of the mother machine. **b)** Steps involved in generating a synthetic image. The outputs of a rigid body simulation of cellular growth are used to draw the "Scene". This contains information about the position of each cell, the walls of the microfluidic device (red) and any contact and collision points between cells (green crosses). From the scene, an accompanying ground truth mask is generated. This represents the true pixel regions where the cell is. Three-dimensional Optical Path Length (OPL) images of the scene is rendered from this scene (Figure S2). A phase contrast (PC) microscope model, in the form of the microscope's point spread function (PSF) in then used to generate a first order (raw) synthetic phase-contrast image. This raw synthetic image is further optimised to maximise similarity to a real image. Optimisation methods include histogram matching, camera noise simulations, and matching of the rotational Fourier spectrum of the synthetic image to a real image. **c)** A real kymograph of phase contrast images of cells in a mother-machine trench is compared to a synthetic kymograph. **d)** Representative timeline of the steps to generate 100 synthetic images. **e)** Comparisons between synthetic and real images across changes in biology, imaging platform, and imaging modality shows the robustness of the pipeline. Blue boxes: We demonstrate the ability to produce images of *E. coli* and *B. subtilis* in their exponential growth phase, where cells are long and randomly stacked. We can also generate images of cells in stationary phase, with the ability to reproduce the two regular stacking modes. Orange box: Changes in the device geometry, such as narrower or wider trench can be accommodated. Purple boxes: Even changes in imaging device, e.g. a low magnification low NA objective, or different imaging modality, such as fluorescence imaging can be simulated. **f)** SyMBac can generate training data 3 orders of magnitude faster than a human. 10,000 synthetic images can be produced in 10 minutes with a high core count CPU and fast enough GPU. **g)** Researchers were asked to annotate synthetic images of bacteria. The pixel-wise error—computed as the Jaccard index (intersection over union)—reveals that humans are not good at perceiving the true boundary of a cell. The results vary vastly across users, however they consistently annotate smaller masks on average.

role in determining the performance of machine learning algorithms trained with it. Diffraction limited images of small bacterial cells, where the cell size is of a similar scale as the size of the point-spread function of the microscope, pose a fundamental challenge to human annotation. Looking purely at an unprocessed image, it cannot be exactly known where the object ends and background starts. This is further complicated by geometric problems, as edges of cells have fewer volume elements and thereby appear lighter than the regions with more volume element near the centre. This causes human annotators to incorrectly label pixels as foreground and background (which itself is also variable across users). To illustrate this issue, we asked three researchers to annotate synthetic data (100x oil) generated by SyMBac and then compared their drawn masks with the ground truth. The intersection over union (IoU) between true mask and human annotated mask is given in Figure 1g. The results demonstrate not only that masks from different users have high variability, but on average are significantly different to the actual cell, even when cells are large (exponential phase). The errors are even higher for smaller cells in stationary phase. This suggests that even a curated set from real data relies on the subjectivity of the observer, which varies highly between experts and remains inaccurate. Further investigation showed that the way in which human annotators were getting masks wrong was by severe underestimation of cell lengths and widths, with errors higher than 10% and 20% respectively (Figure 2g). This has implications for models trained on human generated training data, as this bias is propagated to the neural network, causing the model to learn incorrect relations between the image and the mask. In the next section we illustrate how training models with SyMBac generates more precise and accurate masks of bacterial cells from low resolution images of diverse types and enables rigorous checks on size variability across the population and over time.

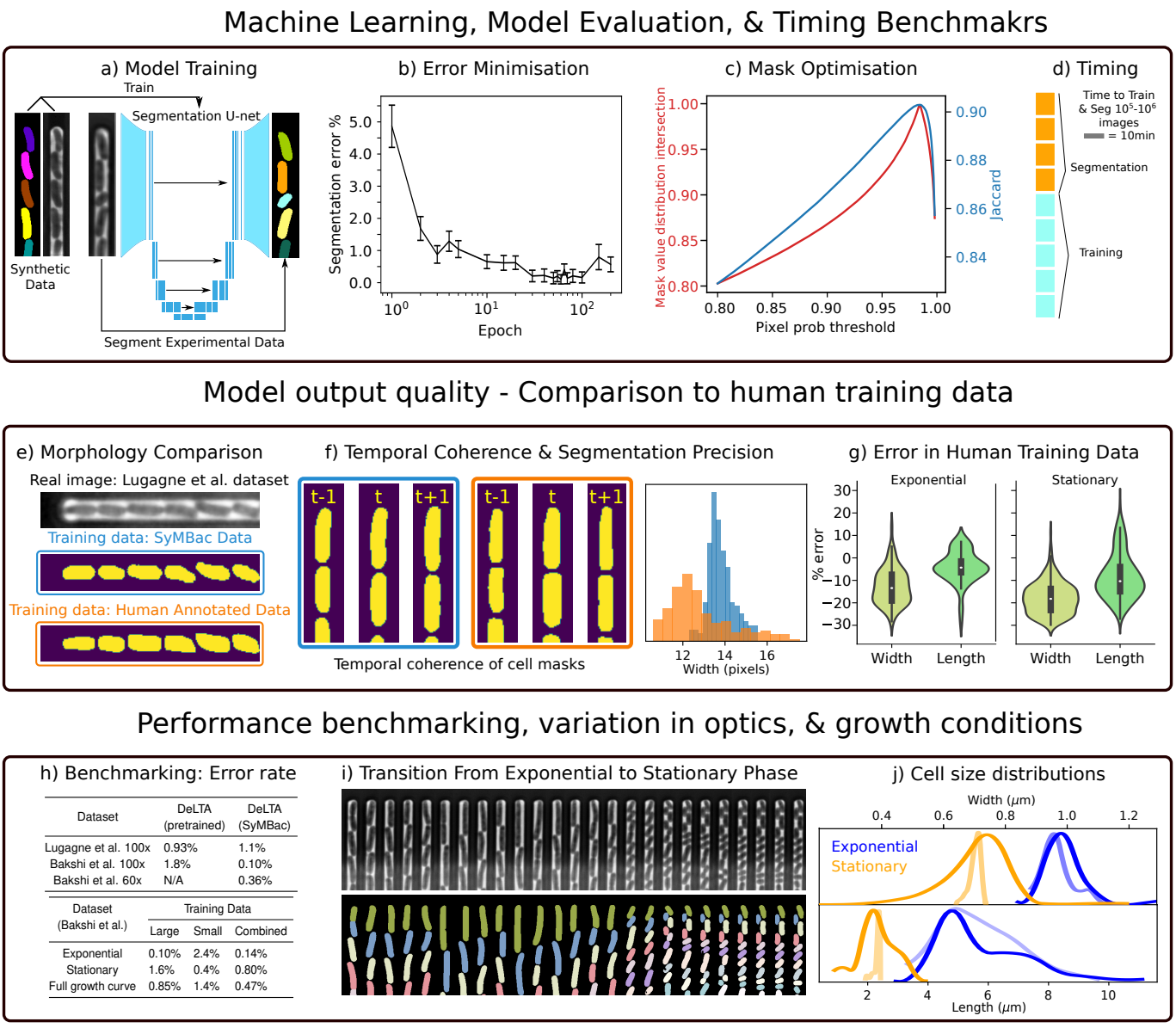
Once realistic synthetic images are generated, a segmentation deep neural network (U-net [11]) can be trained.

This network will learn the relation between the synthetic phase contrast image and ground truth pair to infer image to object relationships. The network can then reverse-engineer this relation to learn about positions of objects from their image. We evaluated the usefulness of the synthetic data for this purpose by applying it to the DELTA implementation of U-net [10]), which is a well used and very robust pipeline for analysing mother machine images. The performance of neural networks for mother machine image segmentation can be considered through two aspects: identification of cells and accurate segmentation of their boundaries. Since synthetic data is being used for training, validation data is not available in the traditional sense. Instead, the model is evaluated automatically epoch-by-epoch in terms of identification errors, and the model with the lowest error is kept (Figure 2b). Automatic error rate calculation is possible without validation data due to the predictable nature of bacterial growth in the mother machine (explained in Figure S8). In general, 100 epochs were enough to bring the identification error below 0.5% (before over-fitting began to take place).

Because the output from the U-net is a probabilistic mask, a correct probability threshold must be determined to compute the binary masks for cells. We developed a novel validation mechanism which utilises synthetic data to estimate the optimal mask pixel probability threshold by passing an independent set of synthetic validation data through the model and thresholding the output masks until they are maximally similar to the ground truth masks (Figure 2c). To maximise the mask quality, the threshold value which maximises the Jaccard index between the predicted and ground truth masks was chosen. Alternatively, one can compare the distributions of ground truth length and width to those predicted by the model for different pixel probability thresholds. The maximal distribution intersection then corresponds to the optimal mask threshold. As can be seen, these two optimisation methods yield the same optimal mask threshold (Figure 2c). We emphasise that models trained with SyMBac data require no pre-processing of training data, and no post-processing of output masks, whereas in [10] data used for training the model was pre-processed, and output masks were post-processed using morphological operators, which by their very nature will deform the cell masks.

To illustrate the superior ground truth quality of synthetic images generated with SyMBac, we first compared the output masks from a DeLTA model trained with SyMBac, and a DeLTA model trained with human human curated training data (dataset from Lugagne et al. [10]). Alongside their algorithm, the authors of DeLTA provide curated segmentation training and evaluation datasets. In the case of DeLTA, no good segmentation algorithm existed for the data, so a semi-automatic tool was used. As a result, the training data becomes artifactual. Generating training data using such tools requires the use of morphological operators, various thresholding techniques, and other traditional segmentation methods. The parameters for these methods, and the order in which they are applied is subjective. This results in training data with oddly shaped masks, and these odd shapes are learned by the network during training (Figure 2e). These shapes are also variable in time and non-representative of cells seen in the images. Moreover, in creating the training data and in thresholding the final probabilistic masks, the authors of DeLTA used morphological operators (such as binary opening) to separate cells. This causes cell masks to have flat ended poles. In our comparisons—to maintain consistency—we did not apply any post-processing to output masks beyond simple thresholding. To perform a rigorous comparison, we generated a synthetic training dataset based on reference images from the original DeLTA paper. Since the model has access to perfect ground truth for synthetic data, it produces less flawed cell masks near cell-to-cell borders and visually more "natural" cell shapes (Figure 2e).

Furthermore, we observed that the DeLTA model trained on human annotated data demonstrated abnormal frame-to-frame temporal decoherence in cell width (Figure 2f), while masks from SyMBac trained models showed high temporal coherence. To illustrate this further, we compare the distribution of cell widths from a single trace over time for masks produced using models trained with the different types of training data. The masks from SyMBac trained models maintain a narrow distribution of widths, while the masks from DeLTA display wide variation with the peak shifted to lower values (Figure 2f). The observed distribution of cell mask widths is reminiscent of our experiments with human annotation errors. When we compared the widths and lengths of human annotated synthetic data with the ground truth widths and lengths of ground truth masks (Figure 2g), we found that human annotated masks are narrower than the true masks ($\sim 15\%$ for large cells in exponential phase and $\sim 20\%$ for smaller cells in stationary phase cells), which provides a likely explanation for the observed performance bias and variability in models trained with this type of training data. This analysis establishes the power of U-net models trained with synthetic data for



Dataset	DeLTA (pretrained)	DeLTA (SyMBac)
Lugagne et al. 100x	0.93%	1.1%
Bakshi et al. 100x	1.8%	0.10%
Bakshi et al. 60x	N/A	0.36%

Dataset (Bakshi et al.)	Training Data		
	Large	Small	Combined
Exponential	0.10%	2.4%	0.14%
Stationary	1.6%	0.4%	0.80%
Full growth curve	0.85%	1.4%	0.47%

Figure 2: Image segmentation with SyMBac: **a)** The DeLTA [10] model is trained on synthetic data and corresponding ground truth masks, and real data can then be segmented using the model. **b)** Segmentation error of trained models are evaluated automatically epoch-by-epoch by analysing the sawtooth curves of cell length changes over time (Figure S8), and the model with the lowest error is kept. **c)** The optimal probability threshold is identified by passing synthetic validation data through the model, and the threshold is adjusted to maximise the Jaccard index between the predicted masks and the ground truth masks, or the maximal distribution intersection between size distributions. In this case, these two optimisation parameters yield the same optimal mask threshold. **d)** A representative timeline for model training and prediction of a large dataset is shown. **e)** Comparison of morphology between the output masks from a model trained on synthetic data vs a model trained on the DeLTA human annotated data. **f)** Masks from models trained on synthetic data demonstrate better temporal coherence between frames. Comparison of the cell width distributions for a single cell over time shows that models trained on human training data underestimate the cell width, but also have a large tail towards overestimating cell width. The model trained on SyMBac data has a tight width distribution demonstrating good temporal coherency. **g)** Human annotated training data consistently reports lower cell widths than the model trained on synthetic data. **h)** Table (top): Performance benchmarking showing the error rate of mis-segmentation of the mother cell over all three datasets. The error of the DeLTA pretrained model on 60x data was too high to accurately measure. (**Note**: the error rate for Lugagne et al. [10] differs largely from that reported in their original paper, because we use automated sawtooth analysis to measure error, and we also do not post-process output masks.) Table (bottom): performance benchmarking of models trained with synthetic data of large cells, small cells, and combined sizes, on images from growth-curve experiment [2]. **i)** A kymograph of cells transitioning from exponential to stationary phase, and accompanying masks generated from a model trained on synthetic data. This demonstrates the model's robustness across multiple cell sizes. **j)** Width and length distributions of single cells (light orange and blue) in exponential growth and stationary phase compared to overall width and length distributions of 300 single cells (dark orange and blue).

segmentation of bacterial cells. The superior quality of cell masks will be beneficial in the study of size and width regulation analysis of bacterial cells. As we show below, DeLTA models trained with SyMBac can generate very accurate masks and enable estimation of both cell length and cell width with a precision of ~ 10 nm (Figure 2j and

Figure S9). This is several times more precise than existing cell segmentation algorithms (Figure S10), more accurate, and is also robust to large variations in cell size (Supplementary Notes 8). Additionally, we also investigated and established that common methods of estimating cell size—such as Otsu thresholding and using membrane dyes to highlight the cell perimeter—are also not good estimators of the cell's true shape (Supplementary Notes 7) and suffer from variable underestimation of cell dimensions.

The cell identification accuracy in the test dataset provided by the DeLTA authors (estimated from segmentation error as detailed in Figure S8) was comparable between models trained trained with the training data provided with the paper, and models trained with the SyMBAC pipeline (table in Figure 2h). This is a remarkable achievement, since it shows that SyMBAC is able to generate realistic data from just the reference images. However, the true advantage of SyMBac is in its ability to generalise to variations in imaging setting and growth conditions (as shown in Figure 1e), which makes the performance of DeLTA robust towards these effects. To illustrate this, we first analysed the performance of the pre-trained DeLTA model on a separate 100x dataset with different imaging conditions (detailed in Bakshi et al. [2]). The error-rate of DeLTA increased 2x for this dataset. We must emphasise that the pre-trained model from Lugagne et al. performs remarkably well considering it has not seen our experimental data before. However, SyMBac allows us to rapidly generate new training data which matches images from Bakshi et al. [2], and effortlessly brings the error rate down to 0.1%, improving performance by 18-fold. The pretrained DeLTA model could not produce any useful masks from a data set collected with a 60x air objective and therefore we could not quantify the error. On the other hand, SyMBac can generate similar images in a matter of minutes, yielding highly accurate segmentation with an error rate of 0.36% (table (top) in Figure 2h). This analysis highlights two key powerful attributes of SyMBac. Firstly, it is very easy for a user to generate large volumes of high quality training data for any new experiment experiment where optics, illumination, platform, or the biology might have changed. Secondly, SyMBac enables high-accuracy segmentation from very low resolution low signal-to-noise data, collected with air objectives. This will have important implications in using the mother machine for high-throughput screening (for persisters [2], or identifying optimum designs of synthetic circuits [7]). Low resolution air objectives can provide the necessary travel distance and speed, as well as the large field of view that is required for such experiments. We hope that the ability of SyMBac to provide an accurate segmentation platform for such datasets will help to advance this field, furthermore we are currently producing an equivalent pipeline which will simulate the growth of whole colonies, and produce synthetic data for their segmentation as well.

Segmentation algorithms (both ML and non-ML types) face a severe challenge when the data type changes during the experiment, for example cells changing size along a growth-curve experiment as they transition from exponential to stationary phase [2]. Since cells in the stationary phase are very small, the annotation errors are very high and compromise the performance of the ML algorithm (Figure 1g). On the other hand, non ML models suffer from changes in object size, as no common filter (or structural element) size can be used for objects which span many size scales. SyMBac can address this issue by generating high-quality training datasets for stationary phase cells (Figure 1e). However, while models trained with small cell training data perform well in stationary phase, this comes at the cost of reduced accuracy in exponential phase. Conversely, models trained with large cells perform well in the exponential growth phase but generalise poorly to stationary phase (bottom table in Figure 2h). To optimise the performance throughput the entire dataset, we verified the performance of models trained with SyMBac data that has a combination of large and small cells. This resulted in the best overall performance across the entire growth curve (error rate 0.47%), but did slightly worse than the specialised models. We show an example kymograph of segmented masks next to the real kymograph to illustrate the performance quality (Figure 2i).

The precise and accurate cell masks enabled us to compare width and length regulation along the growth-curve, for the first time. The orange distributions in Figure 2j are from the stationary phase. From single cell (lighter curves) timeseries analysis, we could estimate the precision of our masks to be as good as 10 nm (Figure 2j). This is possible since the cell dimensions are essentially constant in the stationary phase (figure S15). The distribution of widths and lengths from all the cells in the dataset show that cells have largely variable widths and lengths in the stationary phase. However, when we compare the distribution of lengths and widths of single cells over time vs many cells at any given point in exponential phase, we find that those are comparable. The width and length of cells vary in exponential phase due to growth and division, but the variability across the population is low, and is indicative of a homeostasis

mechanism. On the other hand, the large variability across the population in stationary phase suggests cells enter stationary phase at various phases of their growth-cycle and are "locked in" there. Thus SyMBac enables the detailed analysis of the size regulation of minuscule microbes.

In summary, we have designed a pipeline to rapidly generate realistic synthetic images of bacterial growth in microfluidic devices in order to train deep-learning networks. SyMBac's synthetic data can be generated in a fraction of the time it takes for a human to annotate the same number of images, is highly similar to real data, and is capable of training highly accurate deep neural networks for image segmentation. Not only does the performance of models trained on SyMBac data approach or match models trained on human annotated data, but the quality of the output masks far exceeds the quality of masks generated by human trained models. This is validated by noting that masks demonstrate high temporal coherence and realistic cell morphology.

Methods

Image Generation

Our method employs several steps which make up a pipeline:

- Simulation of growth and division of cells and their interactions with each other and the microfluidic device.
- Extracting geometric information and rendering images of the bacteria and device in the scene.
- Convolution of these images with phase contrast or fluorescence point-spread functions.
- Optional image optimisation to further improve similarity to experimental images.
- Generation of synthetic data.
- training ML segmentation algorithms with training data.
- Performance analysis.

  We give brief descriptions for each of these steps below, with additional information presented in the supplementary material.

Step 1: Simulation of growth and division of cells and their interactions with each other and the microfluidic device

We built an agent based model (ABM) of cell growth and division, taking into account the three size regulation modes of bacteria (adder, sizer, timer), while also adding variability to the key parameters governing cell division. Cells are agents with properties such as length, width, position, age, and can be customised to include new properties such as time-dependent fluorescence intensity. These cell agents exist as dynamic objects in a rigid body physics world called a "scene", modelled using Python bindings of the popular physics engine Chipmunk (pymunk) [12]. We add static objects to these scenes which are shaped into a mother machine trench geometry, with configurable length and width. A global event loop keeps track of cell growth, divisions, inter-cellular collisions and cell-trench collisions at each time-step. A simple API was created to allow simulations to be run by entering the desired distributions of cell lengths and widths by defining the mean, variance and size-regulation mode, and then selecting the desired trench geometry.

Extracting geometric information and rendering images of the bacteria and device in the scene

After simulations are complete, geometric information about the cell the trench position are extracted. This includes each cell's vertices in every time-point. The cell's 3D structure is a spherocylinder, which affects how light interacts with it, yet the simulation engine runs only in 2D. Therefore to render 3D cells, the cell's vertices are revolved along

their major axis to generate each cell's spherocylindrical hull. All of these 3D cells are then morphed to simulate cell bending and then re-rendered and in a 3D scene, computationally represented by a 3D NumPy array. A 2D projection of this 3D array is then taken, such that an image is generated where the intensity in each pixel is proportional to the optical path length in that region (Figure S2). The details of this calculation is provided in Supplementary Notes 6. This simulated image corresponds to what would be seen if a microscope had no PSF, and thus induced no diffraction artefacts in an image. From the simulated images at this stage, we also save the ground truth images of cell the positions.

Microscope model for generating phase-contrast and fluoescence images

The microsope model generates images from the scenes by convolving them with PSFs of relevant optics. First, the raw image is rendered at a high resolution (typically 3x the real pixel size) and then convolved with a variety of fluorescence or phase contrast PSFs (also rendered at higher resolution) depending on the application. After PSF convolution, the image is sub-sampled back to the original resolution. Convolution at higher resolution is more accurate, because sub-resolution features of the PSF are maintained, and propagated to the image. Otherwise, the PSF convolution results would be corrupted by low resolution artefacts. The user can select an ideal fluorescence PSF or phase contrast PSF which are modelled as an Airy disk and obscured Airy disk respectively (Supplementary Notes 2). The PSFs needs to be parameterised by inputting the microscope objective's numerical aperture, refractive index of the medium, emission wavelength, camera pixel size, and phase ring and annulus dimensions (details provided in Supplementary Notes 2). Upon defining the PSF of choice, it is convolved over the raw image to produce a synthetic micrograph.

Image Optimisation

Next, the user can use the optional image optimisation to maximise the similarity between the synthetic image and an example real microscope image. Multiple image manipulations are possible, including intensity histogram matching, noise matching, rotational Fourier spectrum matching, camera noise simulation, stage drift simulation, and even manual intensity corrections. This is done using an interactive IPython Notebook interface, where the user is presented with a side-by-side comparison of a synthetic image and a real image from their experiment. Two plots showing the relative errors in intensity and variance between cells, the mother machine device, and the space between cells are given such that the user can manipulate their image to minimise this error (Figure S6). (Note: While examples of black box optimisers for image similarity matching are included, we avoid their use for this step due to the very noisy error-landscape between synthetic and real images. Moreover, we were unable to define a perfect objective function which guarantees perfect image similarity). Synthetic images are then generated according to the optimal parameters. These parameters can be varied according to a uniform or Gaussian distribution during the image generation process to simulate minor fluctuations in the experimental setup. This also acts as a form of data augmentation, but occurs during the image generation process, which preserves the mapping between object and image. For this reason we do not use data augmentation during the machine learning process.

Testing

Interfacing code for training machine learning models using DeLTA [10] and StarDist [13, 14] are provided. For most analysis, U-net models were trained with DeLTA, as this provided the most accurate masks. StarDist, while not pixel-perfect, can be usefol for especially low magnification data (Supplementary Notes 5) because of its robust ability to separate densly packed star-convex shapes.

We tested SyMBac against 4 datasets, which have their details given in Table 1. 4 sets of data were generated, each with approximately 2000 synthetic images. 3 of the datasets are images of *E. coli* growing in a mother machine, taken with a 100x oil objective, and one dataset was imaged with a 40x air objective, using a 1.5x post-magnification lens, giving an effective magnification of 60x. However synthetic images for this dataset were still generated according to the optics of the 40x objective, and then scaled to the appropriate pixel size.

Table 1: Datasets analysed using SyMBac

Dataset	Optics	Biology	Notable features
Lugagne et al. (exponential)	100x oil	*E. coli* in balanced growth	Poor contrast data
Bakshi et al. (exponential)	100x oil (NA = 1.30)	*E. coli* in balanced growth	High contrast data
Bakshi et al. (stationary)	100x oil (NA = 1.30)	*E. coli* in stationary phase	Small stacked cells
Bakshi et al. 60x (exponential)	40x air + 1.5x post mag (NA = 0.95)	E. coli in balanced growth	Low resolution + non-uniform illumination

Model Evaluation

For evaluation purposes, the DeLTA implementation of U-net was used to train models using synthetic data. Because the data is purely synthetic, validation data in the traditional sense does not exist. For this reason we saved the model every epoch and evaluated performance through the analysis of growth data. Bacterial growth in the mother machine traces out a predictable sawtooth wave (Figure S7i). These sawtooth waves and their derivatives were analysed for spurious peaks, which are the signature of over and under-segmentation errors (Figure S8).

The output from a U-net is a probabilistic threshold image. This image needs to be converted into a binary mask by thresholding pixel values between 0 and 1. Thresholds close to 0 will generate larger, connected masks, and thresholds close to 1 will generate smaller less connected masks. In order to identify the optimal threshold value which generates the most accurate masks, an independent set of synthetic data is segmented using the neural network, and the probability thresholds are adjusted to maximise the Jaccard index between the ground truth and the model output (Figure 2c). Depending on the type of dataset this value can range between 0.6 and 0.99. If the thresholding value is low, then masks will be connected. To alleviate this we use seeded watershed segmentation to cut masks appropriately. Seeds are generated by first thresholding masks at very high probabilities ($P > 0.999$), and performing the watershed on the optimal probability binary mask. It must be noted that we do not perform any post-processing on the training data or the segmentation results, as this would affect the shape of the mask.

Comparison To Human Drawn Masks, Otsu's Method and Membrane Dye Drawn Masks

In order to test how accurate humans are at annotation of training data, we sent 3 researchers a set of 100 synthetic images. The researchers were asked to label the images and time themselves. We then compared the length, width and pixelwise (IoU) accuracy of the human generated masks to the ground truth masks of the synthetic data.

In order to evaluate the quality of masks generated by models trained on human generated data and computer generated data, we compared the masks outputted by the pretraiend model supplied with the DeLTA paper on its own test data, and a model we trained on synthetic data. We first made qualitative observations of mask quality, and then quantified these by assessing the temporal coherence of single cell width between frames.

To compare the accuracy of other segmentation methods, such as traditional Otsu thresholding and cell perimeter evaluation using membrane dyes, we generated synthetic images of single cells in phase contrast and with membrane dyes, along with accompanying ground truth. We then segmented the cells with Otsu's method and the membrane dye method, comparing the output masks to the ground truth.

References

[1] Wang, P., Robert, L., Pelletier, J., Dang, W. L., Taddei, F., Wright, A. *et al.* Robust Growth of Escherichia coli. *Current Biology* **20**, 1099–1103 (2010).

[2] Bakshi, S., Leoncini, E., Baker, C., Cañas-Duarte, S. J., Okumus, B. & Paulsson, J. Tracking bacterial lineages in complex and dynamic environments with applications for growth control and persistence. *Nature Microbiology 2021 6:6* **6**, 783–791 (2021).

[3] Lord, N. D., Norman, T. M., Yuan, R., Bakshi, S., Losick, R. & Paulsson, J. Stochastic antagonism between two proteins governs a bacterial cell fate switch. *Science* **366**, 116–120 (2019).

[4] Potvin-Trottier, L., Lord, N. D., Vinnicombe, G. & Paulsson, J. Synchronous long-term oscillations in a synthetic gene circuit. *Nature 2016 538:7626* **538**, 514–517 (2016).

[5] Taheri-Araghi, S., Bradde, S., Sauls, J. T., Hill, N. S., Levin, P. A., Paulsson, J. *et al.* Cell-Size Control and Homeostasis in Bacteria. *Current Biology* **25**, 385–391 (2015).

[6] Robert, L., Ollion, J., Robert, J., Song, X., Matic, I. & Elez, M. Mutation dynamics and fitness effects followed in single cells. *Science* **359**, 1283–1286 (2018).

[7] Luro, S., Potvin-Trottier, L., Okumus, B. & Paulsson, J. Isolating live cells after high-throughput, long-term, time-lapse microscopy. *Nature Methods 2019 17:1* **17**, 93–100 (2019).

[8] Manuse, S., Shan, Y., Canas-Duarte, S. J., Bakshi, S., Sun, W.-S., Mori, H. *et al.* Bacterial persisters are a stochastically formed subpopulation of low-energy cells. *PLOS Biology* **19**, e3001194 (2021).

[9] Hardo, G. & Bakshi, S. Challenges of analysing stochastic gene expression in bacteria using single-cell time-lapse experiments. *Essays in Biochemistry* **65**, 67–79 (2021).

[10] Lugagne, J. B., Lin, H. & Dunlop, M. J. Delta: Automated cell segmentation, tracking, and lineage reconstruction using deep learning. *PLoS Computational Biology* **16**, e1007673 (2020).

[11] Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marrakchi, Y. *et al.* U-Net: deep learning for cell counting, detection, and morphometry. *Nature Methods* **16**, 67–70 (2019).

[12] Blomqvist, V. Pymunk: A easy-to-use pythonic rigid body 2d physics library (version 6.0.0) (2007). URL `https://www.pymunk.org`.

[13] Weigert, M., Schmidt, U., Haase, R., Sugawara, K. & Myers, G. Star-convex Polyhedra for 3D Object Detection and Segmentation in Microscopy. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)* (2020).

[14] Schmidt, U., Weigert, M., Broaddus, C. & Myers, G. Cell Detection with Star-Convex Polygons. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II*, 265–273 (2018).

Code availability

Code is available at https://github.com/georgeoshardo/SyMBac

Acknowledgements

We thank William Earley for very detailed feedback on the manuscript. We thank the Dunlop Lab for producing the easy to use DeLTA, and for their permission to include the source code in SyMBac's codebase.

Author Contributions

G.H wrote the code, performed the simulations and analysis, and wrote the paper. M.N performed pilot and feasibility studies. S.B conceived of the idea, performed analysis, and wrote the paper.