# Supplementary Material

## A. Experimental Design

The main dataset in this paper, BBBC039, was split in three subsets, using 50% of the images for training, 25% for validation, and 25% for testing (holdout set). All optimization steps of the deep learning strategies were tuned using the training and validation sets, and the reported results are obtained with final evaluations using the holdout test set (unless otherwise indicated).

We evaluate five segmentation strategies, two based on deep learning, one based on classical machine learning, and two baseline pipelines based on the CellProfiler software. The deep learning strategies are U-Net [(16)](https://paperpile.com/c/ZdopOx/dtm6o), representing a family of neural networks originally designed for segmentation; and DeepCell [(17)](https://paperpile.com/c/ZdopOx/H8DJe), representing another family of neural networks for patch-based, pixel-wise classification. Random Forest was selected as the classical machine learning algorithm using the implementation provided by Ilastik [(10)](https://paperpile.com/c/ZdopOx/gWwBy), which is specialized for segmentation. The baseline segmentation pipelines include an advanced settings mode and a basic settings mode.

With trained deep learning models and calibrated CellProfiler pipelines, we proceeded to segment all images in the test set. To evaluate and compare the estimated masks we use the average F1-score and modes of error.

## B. Post-processing operations

The following are the post-processing steps used to recover single object instances from pixel classification probability maps: 1) max-binarization: the prediction for each pixel is a 3-way probability distribution. These predictions are binarized by assigning 1 to the category with maximum probability and 0 to the rest. 2) Isolation of the interior channel. The other channels only serve the purpose of indicating where the boundaries and background are located, thus they are ignored after determining binary values for each pixel. 3) Use the connected components algorithm [(50)](https://paperpile.com/c/ZdopOx/J43L2) to label single objects in the interior channel. After removing the boundaries and background, single nuclei are separated in a binary map. Each of these regions is assigned a different number. 4) Dilation of single objects. The dilation morphological operation [(51)](https://paperpile.com/c/ZdopOx/YprFm) is applied to all objects using a square structuring element of 3x3 pixels to compensate for the loss of boundaries. This procedure was used for the two evaluated deep learning models: DeepCell and U-Net. No additional boundary refinement algorithms --such as watershed or similar-- were applied.

****

**Supplementary Figure S1.** F1-score has better resolution for detecting performance differences than the Jaccard index. In addition, the F1-score is penalized by biologically relevant errors that can be explained in terms of missing, merged and split objects. a) F1-score and Jaccard index are highly correlated measures of segmentation performance, but F1-score displays more resolution to distinguish biologically meaningful errors. Each dot in the plot is an image in the test set segmented with the CellProfiler advanced pipeline. b) Example images that get similar Jaccard indexes but have very different segmentation errors as shown by the yellow arrows. F1-score captures these differences more naturally, which is explained by the number of false negatives resulting from the merges. c) Performance of the evaluated methods on the entire test set. The Jaccard index reports little difference between the two CellProfiler baselines, and also between Random Forest and the two deep learning methods. With average F1-score, the biologically relevant errors are measured more explicitly, and performance differences can be observed more clearly.



**Supplementary Figure S2**. Total number of missed (false negative) and extra (false positive) objects generated by each segmentation model in the test set. The total number of objects in the test set is 5,720. Deep learning algorithms reduce both missed and extra objects with respect to the baseline. U-Net misses 47% less objects than the CellProfiler advanced pipeline, while Random Forest misses 6% more. Similarly, DeepCell introduces 77% less extra objects, providing cleaner segmentations. Deep learning algorithms display a tradeoff between accurately recovering all relevant objects and not introducing segmentation artifacts (by splits or debris).



**Supplementary Figure S3.** Z’-factor values for a set of 10 compounds selected for evaluation. a) The Z’-factor values indicate whether a compound has an observable cell cycle disruption in a population of cells with respect to negative controls. To compute the Z’-factor, images are first segmented using CellProfiler or a U-Net model. Then, the DNA content is measured within each nucleus using the integrated intensity of pixels. The distribution of DNA content reflects the expected 2N/4N cell cycle. b) Z’-factor values obtained for the two evaluated methods. Higher values are better. c) Scatter plot of the values reported in b. The Y axis presents the result obtained with deep learning and the X axis presents the result with CellProfiler segmentations.