Tracking One-in-a-Million: Large-Scale Benchmark for Microbial Single-Cell Tracking with Experiment-Aware Robustness Metrics

Johannes Seiffarth[1,2(✉)], Luisa Blöbaum[3], Richard D. Paul[4], Nils Friederich[5,6], Angelo Jovin Yamachui Sitcheu[5], Ralf Mikut[5], Hanno Scharr[4], Alexander Grünberger[3,7], and Katharina Nöh[1]

[1] Institute of Bio- and Geosciences, IBG-1: Biotechnology, Forschungszentrum Jülich GmbH, Jülich, Germany
j.seiffarth@fz-juelich.de
[2] Computational Systems Biology (AVT-CSB), RWTH Aachen University, Aachen, Germany
[3] Multiscale Bioengineering, Bielefeld University, Bielefeld, Germany
[4] Institute for Advanced Simulation, IAS-8: Data Analytics and Machine Learning, Forschungszentrum Jülich GmbH, Jülich, Germany
[5] Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany
[6] Institute of Biological and Chemical Systems, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany
[7] Microsystems in Bioprocess Engineering, Karlsruhe Institute of Technology, Karlsruhe, Germany

Abstract. Tracking the development of living cells in live-cell time-lapses reveals crucial insights into single-cell behavior and presents tremendous potential for biomedical and biotechnological applications. In microbial live-cell imaging (MLCI), a few to thousands of cells have to be detected and tracked within dozens of growing cell colonies. The challenge of tracking cells is heavily influenced by the experiment parameters, namely the imaging interval and maximal cell number. For now, tracking benchmarks are not widely available in MLCI and the effect of these parameters on the tracking performance are not yet known. Therefore, we present the largest publicly available and annotated dataset for MLCI, containing more than 1.4 million cell instances, 29k cell tracks, and 14k cell divisions. With this dataset at hand, we generalize existing tracking metrics to incorporate relevant imaging and experiment parameters into experiment-aware metrics. These metrics reveal that current cell tracking methods crucially depend on the choice of the experiment parameters, where their performance deteriorates at high imaging intervals and large cell colonies. Thus, our new benchmark quantifies the

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-91721-9_20.

A. Del Bue et al. (Eds.): ECCV 2024, LNCS 15638, pp. 318–334, 2025.
https://doi.org/10.1007/978-3-031-91721-9_20

influence of experiment parameters on the tracking quality, and gives the opportunity to develop new data-driven methods that generalize across imaging and experiment parameters. The benchmark dataset is publicly available at https://zenodo.org/doi/10.5281/zenodo.7260136.

Keywords: Microbial Cell Tracking · Benchmark · Robustness Metrics · Live-Cell Imaging

1 Introduction

Detecting objects, segmenting their visual appearance into pixel-precise masks and tracking their movement through time is a fundamental challenge of computer vision providing crucial scene understanding necessary for autonomous driving [11], pedestrian management [27], sports or robotics [11]. Especially, in biomedical imaging, tracking the development of individual living cells allows gaining insights into the basic principles of life and diseases. For instance, single-cell tracking allows studying virus infections [34], pathogenic bacteria [15], cell aging [26], and cell interactions [14,41] at the single-cell level. In particular, microbial live-cell imaging (MLCI) is a technology that performs high-throughput screening of the temporal developments of individual cells (see Fig. 1). Herein, living microbial cells are introduced into microfluidic chip devices and trapped within thousands of micrometer-sized microfluidic structures called cultivation chambers. Within these structures, the cells grow in monolayers while their temporal development is recorded using automated microscopy. The microscope scans the cultivation chambers one by one, takes an image and repeatedly performs this within a loop, recording a time-lapse that captures the temporal development of the independent cell colonies. As a result, a single MLCI experiment records dozens of time-lapses and produces hundreds of gigabytes of raw imaging data. Clearly, automated segmentation and tracking methods are essential to extract information from the time-lapse images and to gain insights into microbial colony development and single-cell behavior.

However, tracking living microbial cells presents unique challenges distinct from those usually encountered in general object tracking. First, living microbial cells divide frequently, with division times ranging from a few minutes to hours. In MLCI experiments, few cells grow exponentially into dense and large colonies with up to thousands of cells captured in a single microscopy image (Fig. 1E) while their total number is limited by the size of the cultivation chamber. Second, microbial cells in phase-contrast microscopy are visually hard to distinguish, making it hard to track them by their appearance (see Fig. 1D). Third, the time-lapse recordings are affected by parameters such as the choice of the imaging interval between two consecutive phase contrast images and the number of concurrently monitored cultivation chambers. Moreover, both parameters are interdependent: lower imaging intervals usually simplify the tracking challenge but enforce shorter movement cycles of the microscope and, consequently, limit the number of concurrently monitored cultivation chambers. Notably, the

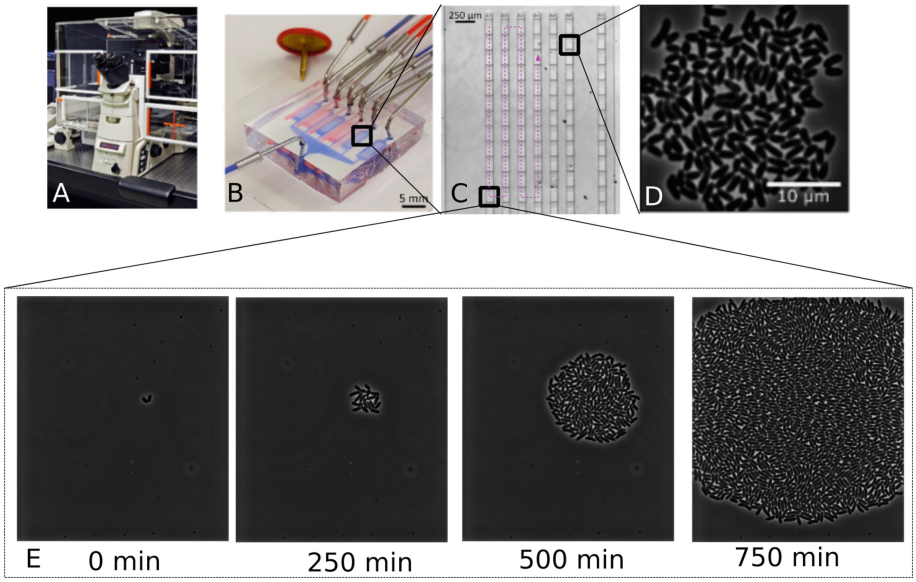

Fig. 1. Data acquisition in microbial live-cell imaging. A microfluidic cultivation device is mounted to an automated microscope (A, B). The device contains hundreds to thousands of rectangular cultivation chambers (C) that are imaged one-by-one, moving the microscope stage and capturing images at a series of time-points (D). This imaging-and-movement loop, indicated by the purple line (C), is repeated with a pre-set imaging interval and leads to time-lapse recordings capturing the temporal development of the cell colony from a few cells up to several thousands (E). The images in (E) depict a $80 \times 90 \mu m$ region. The figure is adapted from Blöbaum *et al.* [4].

imaging procedure itself may influence the growth behavior of the cells. While phase contrast imaging is not considered to impact microbial growth, fluorescence imaging may lead to phototoxicity and -bleaching effects [16]. Therefore, MLCI experiments are usually conducted with relatively low imaging rates. This leads to frequent cell divisions and larger cell displacement between consecutive frames. Consequently, we argue that MLCI not only needs robust and highly automated cell tracking, but also informed choices of experiment parameters, namely the imaging interval and the upper cell count limit.

In recent years, the rapid development in deep learning (DL) methods has driven segmentation and tracking methods [9,13,19,20,28,42–44]. Moreover, the increased availability of large-scale datasets such as ImageNet [29], KITTI [11], CityScapes [6], SA-1B [19], and LAION-5B [31] has shown to be a crucial driver for method development. Within the life sciences, we have seen a similar rapid development of methods. DL segmentation approaches have been developed [7,18,30,35,40], driven by annotated datasets [5,8,12,32,35,39]. For cell tracking, special methods have been developed that incorporate cell divisions [10,17,22,23,25,32,36]. However, public datasets for cell tracking [1,32,39] pri-

marily focus on microscopy images with eukaryotic cells. In contrast to microbial cells, eukaryotes usually show more distinctive visual features, less frequent cell division and the images contain fewer cell instances. For microbial cells, only few tracking datasets [26] are available, making it difficult to train data-driven tracking methods and benchmarking suitable experiment parameters. Due to the lack of datasets, the importance of the experiment parameters for high-quality automated tracking has not been investigated before.

Therefore, we establish a novel benchmark for cell tracking in MLCI and extend existing tracking metrics with experiment parameters to quantify their effect on the tracking performance. Our contributions are threefold: (1) we introduce a new annotated time-lapse dataset, recorded with low imaging interval for the segmentation and tracking of *Corynebacterium glutamicum* cells containing roughly 1.4 million cell masks and about 14k cell divisions. (2) We introduce experiment-aware metrics that extend existing metrics and incorporate the choice of the imaging interval and the maximum number of cells into the evaluation. (3) We evaluate state-of-the-art (SOTA) tracking methods using our devised metrics across a broad range of experiment parameters. We thereby show that the performance of SOTA tracking algorithms deteriorates, especially at lower imaging rates and higher cell counts. Notably, this fact that has not yet been quantified by the CTC community. Therefore, our benchmark represents a step forward to towards fully automated and robust data-driven microbial single-cell tracking and raises awareness about the importance of experiment parameters for cell tracking in MLCI experiments.

2 Related Work

Benchmark Datasets. Benchmark dataset for cell segmentation cover different cell tissues, morphologies and imaging modalities [5,8,32,35]. The availability of additional tracking information for full time-lapse recordings is much less common. The cell tracking challenge combines datasets of various cell types and provides partial dense segmentation and full tracking information [25,39]. Schwartz *et al.* introduced the DynamicNuclearNet dataset containing roughly 600K segmented nuclei instances with roughly 2k cell divisions [32]. Van Vliet *et al.* provide a dataset of six genetic variants of *Escherichia coli* in 39 time-lapse videos containing roughly 100k cell instances and 9k cell divisions. Anjum *et al.* introduce the CTMC challenge dataset contains roughly 2 million cell detections within 2.9k cell tracks and 457 cell division events [1]. Their segmentation annotation is restricted to bounding boxes and on average 13 cells are visible within a microscope image.

Tracking Methods. In the predominant tracking-by-detection scheme, cells are first detected in the microscopy time-lapse and then linked across frames to build biologically valid cell tracks. Thus, tracking-by-detection is usually formulated as a graph problem, where nodes represent the cell detections at specific points in time and edges link cell detections through time. Therefore, Jaqaman *et al.*

[17] formulated tracking as a linear assignment problem (LAP) where cell detections are linked into segments, which are then linked to incorporate cell division events. Theorell *et al.* [36] extended this approach into multi hypothesis tracking (MHT), where cell linking costs are derived from biological models and tracking predictions are sampled using a particle filter approach. In contrast, Löffler *et al.* [23] formulated a coupled minimum cost flow and correct segmentation errors during the tracking. All these methods require hand-tuned parameters to compute the costs of linking cells.

Data-driven approaches promise to derive these linking costs purely from training data. Ben-Haim *et al.* [2] use a graph neural network (GNN) to predict cell linking costs and utilize features computed from contrastive visual embeddings. However, cell divisions are performed heuristically among the cells' neighborhood. Similarly, Schwartz *et al.* [32] utilize a graph attention network to predict linking costs used for building a LAP. Gallusser *et al.* [10] extract cell detection features and use a transformer network to predict the linking costs. The tracking graph is constructed using a greedy scheme or optimizing an integer linear program (ILP) for minimizing linking costs. O'Connor *et al.* [26] predict the dense evolution of cell mask to the next time point for every cell instance using a U-Net architecture. Cells are then linked by their mask overlap. While Gallusser *et al.* and O'Connor *et al.* have been using microbial datasets, the other data-driven approaches have focused on tracking eukaryotic cells.

Tracking Metrics. General object tracking metrics such as MOTA [3] and HOTA [21] have been established for tracking a wide range of objects, but lack the consideration of object division crucial in cell tracking. Thus, specialized tracking metrics have been developed and established in cell tracking, especially by the cell tracking challenge (CTC) [39]. Herein, we distinguish technical and biological tracking metrics. Technical tracking metrics such as the *TRA* and *LNK* are based on the Acyclic Oriented Graph Matching (AOGM) [24]. The AOGM is a weighted sum of costs for the minimal set of operations to transform the predicted segmentation and tracking into the ground truth segmentation and tracking. While the *TRA* metric scores both segmentation and tracking errors, the *LNK* metric solely rates errors in cell linking. Biological metrics are motivated by biological events that are of special interest. For instance, the complete tracks score (CT) measures the number of completely correctly reconstructed cell tracks from the first detection of a cell to its division or disappearance. Moreover, the mitotic branch correctness (MC) rates the quality of reconstructing cell division events.

3 Benchmark Dataset

In this work, we present a new MLCI benchmark dataset for microbial single-cell tracking, briefly termed 'Tracking one-in-a-million' (TOIAM). The dataset consists of microscopy time-lapses of growing *C. glutamicum* that show a characteristic 'snapping' division behavior, adding another challenge to the cell tracking. The images are recorded using phase contrast imaging at low imaging intervals of

one image per minute. The recorded microscopy frames were annotated with cell segmentation masks and tracking information using a semi-automated workflow. We highlight the special characteristic of microbial datasets that are crucial to consider for robust cell tracking.

3.1 Data Acquisition

MLCI experiments are usually carried out in three steps [38]: First, the microbial organism is cultivated in a so-called preculture until reaching a certain biomass measured by the optical density (OD). Second, the cell suspension is introduced into a microfluidic cultivation chip, trapping individual cells within the cultivation chambers. Third, medium supply is connected to the microfluidic chip and the imaging routine is started recording hundreds of cultivation chambers at a specific imaging interval.

In our case, we cultivated *C. glutamicum* (ATCC 13032) in BHI-medium at 30 °C. From an overnight preculture, the main culture has been inoculated the next day starting with an OD600 of 0.05 and grown at 120 rpm to an OD600 of 0.25. A microfluidic chip has been fabricated according to Täuber *et al.* [37], and fixed to the microscope's stage. The main culture cells were transferred to monolayer cultivation chambers (height of 720 nm) on the microfluidic chip within the inoculation procedure. Constant medium flow through the microfluidic device has been provided by pressure driven pumps with a pressure of 100 mbar on the medium reservoir.

The time-lapse phase contrast images of five cultivation chambers have been recorded every minute using an inverted microscope (Nikon Eclipse Ti2) with a 100x oil immersion objective and a DS-QI2 camera (Nikon) at 15 % relative DIA-illumination intensity and 100 ms exposure time. The recording procedure has been performed for 800 minutes, leading to a total of 4, 000 recorded microscopy images. The recorded images provide a spatial resolution of 0.072 micrometers per image pixel in both spatial dimensions.

3.2 Semi-automated Segmentation and Tracking Annotation

To provide high quality annotation for the large number of recorded images with a limited amount of manual annotation workload, we decided to first perform an initial segmentation and tracking using Omnipose [7] and UAT [36], respectively. The result was subsequently corrected by an expert in the annotation tool ObiWan-Microbi [33]. For cell segmentation, we focused on providing annotations for every single-cell. Therefore, over- and under-segmentation, false positive and false negative segmentations were corrected. Based on the corrected segmentation, the tracking edges were manually checked and corrected. Table 1 shows the number of manual corrections carried out for the different time-lapse recordings. Only few manual segmentation corrections were performed. For cell tracking, the majority of corrections had to be performed towards the end of the time-lapse sequences, where the cell count and divisions events increased substantially and cells leave the field of view at the left and right image borders.

Table 1. Amount of manual tracking correction actions to correct errors in the semi-automated annotation workflow. A correction action is the addition and deletion of a tracking link or adding, deleting and editing of segmentation masks.

Manual Correction	Time-lapses					
	0	1	2	3	4	Total
Segmentation	143	319	202	1087	25	1,776
Tracking	773	1,463	4,057	1,823	1,111	9,227

3.3 Dataset Statistics

The annotated dataset contains more than 1.4 million densely annotated cell instances, 29k cell tracks, and 14k cell divisions (see Table 2). Thus, our data set is large enough to be split into meaningful train, validation, test splits. Moreover, Fig. 2 shows that all five time-lapses show similar temporal developments of cell numbers, divisions and disappearances. Nevertheless, cell count and, thereby, the number of cell divisions are exponentially increasing throughout the time-lapses (Fig. 1E) leading to a temporal imbalance. For instance, the large number of cells at the end of the time-lapses leads to 50% of the overall division events occurring in the last 100 minutes of the time-lapse recordings. Moreover, the number of cell disappearances is much higher towards the end of the time-lapse recordings as the colony exceeds the size of the cultivation chamber and cells leave the field of view (Fig. 2B, C).

Table 2. Statistics for five time-lapse sequences in the benchmark dataset and its split into train, validation and test sets. The table shows the number of densely annotated segmentation masks (cell instances), cell tracks and cell divisions.

Split	Time-lapse Index	# Images	# Cell Instances	# Cell Tracks	# Cell Divisions
Train	0	800	238,364	4,918	2,448
Train	1	800	292,070	6,137	3,053
Train	2	800	327,832	6,884	3,428
Val	3	800	292,995	6,184	3,064
Test	4	800	264,011	5,740	2,844
Total	**0,1,2,3,4**	**4,000**	**1,415,272**	**29,863**	**14,837**
Train	0,1,2	2400	858,266	17,939	8,929

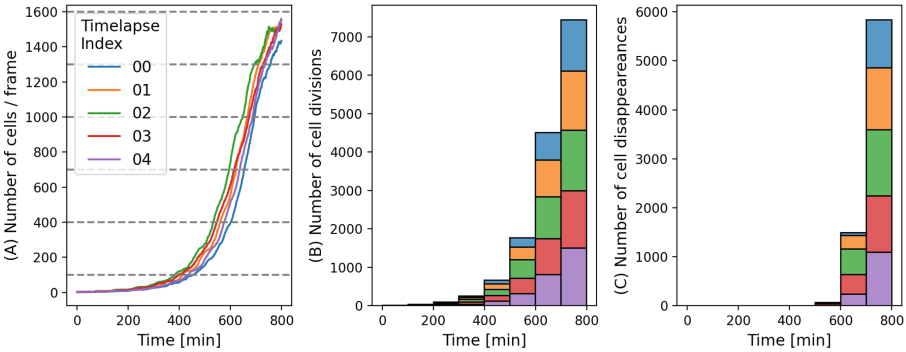

Fig. 2. Temporal imbalance of the five microbial time-lapse recordings of the benchmark dataset. We measured the temporal development of cell count (A), cell divisions (B), and cell disappearance events (C) for each time-lapse. Cell division and disappearance events are grouped into bins of 100 min. The dashed lines in (A) indicate cell count limits (100, 400, 700, 1,000, 1,300, 1,600).

3.4 Implications of Imaging Interval Subsampling

The temporal imbalance of microbial time-lapse datasets in Fig. 2 is amplified by the choice of the imaging interval. Recording the TOIAM dataset with a low imaging interval allows us to simulate higher intervals by considering only every kth recorded image. We call this subsampling with a factor $k \in \mathbb{N}$ in the following. Due to the imaging interval of one minute, a subsampling of k leads to an imaging interval of k minutes. Using the subsampling procedure, we created subsampled datasets with higher imaging intervals.

Figure 3 shows that changing the imaging interval has a tremendous impact on the structure of the dataset and, therefore, on the challenge to track the living cells. Time-lapses with simulated higher imaging intervals contain fewer images, but the number of cell divisions and cell appearances stays constant. Thus, the number of cell divisions and disappearances between consecutive frames increases strongly with higher imaging intervals (Fig. 3C, D). More cell divisions between consecutive frames lead to much larger cell displacements due to the 'snapping' cell division of *C. glutamicum* (see Fig. 3E). However, also the frequency of cell divisions increases. While at an imaging interval of one minute roughly 1 % of cell links are cell divisions, higher imaging intervals lead to a strong increase, with up to 34 % of the cell links being divisions at an imaging interval of 40 minutes (Fig. 3F). Thus, having a good estimate on the frequency of cell divisions, for example based on previous experiments, is crucial.

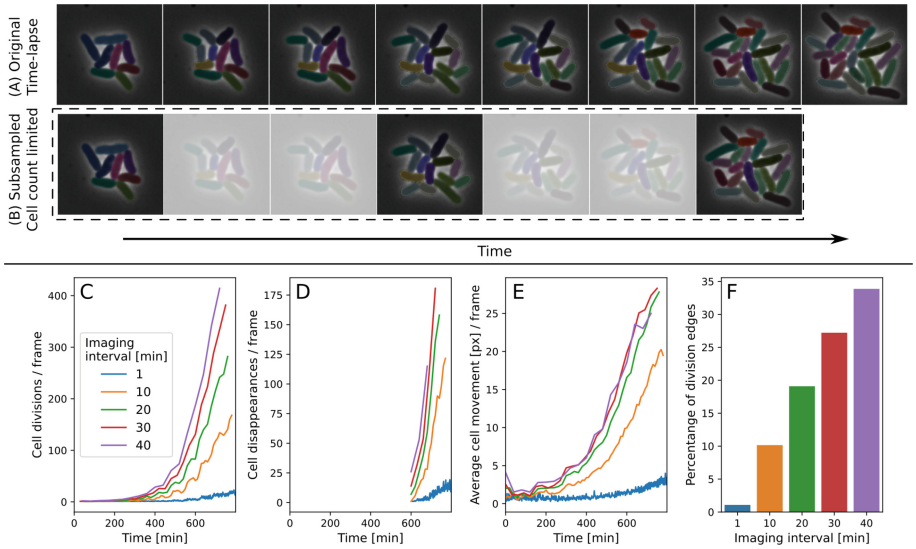

Fig. 3. Subsampling and cell count limiting of time-lapse sequences. (A) shows an excerpt of an MLCI time-lapse. (B) shows an exemplary subsampling with a factor of 3 and truncation at a cell count limit of 21 leading to 3 frames in total. Grayed out images denote frames removed due to subsampling, the dashed box denotes the cell count limit. (C-E) shows the temporal changes in the number of cell division, disappearance and movement per microscopy frame when subsampling to different imaging intervals. The curves have been exponentially smoothed. (F) shows the percentage of division events in contrast to non-division links. (C-F) show data from the TOIAM test split.

4 Experiment-Aware Microbial Live-Cell Tracking Metrics

For a benchmark, suitable metrics are essential that rate the quality of the method performance and serve as an objective comparison tool. However, we have shown that MLCI time-lapses have unique characteristics that make the application of existing metrics difficult. Therefore, we present two new metrics that build on top of the well-established CTC metrics but introduce experiment awareness: First we incorporate the influences of experiment parameter choices, i.e. the imaging interval and maximum number of cells. We decompose the existing metrics along these experiment parameters and term these experiment-aware tracking metrics ($EATM$). Second, we summarize the performance of tracking metrics across a wide range of these parameters within a single robustness metric (RM).

4.1 Cell Tracking Metrics

The CTC introduces several metrics for measuring the quality of a tracking prediction in comparison with ground truth information [39]. The *TRA* and *LNK* metrics are based on the Acyclic Oriented Graph Matching ($AOGM$) score that determines the minimum number of operations needed to convert a predicted tracking result into its corresponding ground truth [24]. These operations include corrections of over- and under-segmentation and the addition or removal of tracking links. Each of these operations is awarded a constant penalty cost, and the AOGM gives their weighted sum.

The *TRA* metric is the $AOGM$, normalized to $[0, 1]$:

$$TRA = 1 - \frac{\min\left(AOGM, AOGM_0\right)}{AOGM_0} \tag{1}$$

where $AOGM_0$ is the $AOGM$ of an empty tracking graph (no nodes, no edges). The *LNK* metric scores the quality of the tracking:

$$LNK = 1 - \frac{\min\left(AOGMA, AOGMA_0\right)}{AOGMA_0} \tag{2}$$

where $AOGMA$ denotes the $AOGM$ where only edge operation costs are considered and the $AOGMA_0$ denotes the $AOGMA$ of the ground truth graph without edges, respectively.

Moreover, the *DIV* and *CT* metrics focus on correct reconstruction of 'biological events' that are cell divisions (DIV) and complete cell tracks (CT). Therefore, for both types of events true positives (TP), false positives (FP), and false negatives (FN) are computed, while precision and recall are summarized in the F1-score

$$F_1 = \frac{2}{1/precision + 1/recall} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}. \tag{3}$$

4.2 *EATM* Cell Tracking Metric for MLCI

Existing tracking metrics, such as those introduced by the CTC, have proven to be useful to rate the tracking quality. However, they do not consider the specific characteristics of microbial time-lapses that are crucially influenced by experiment parameter choices. Thus, we extend the existing metrics and decompose them in to the devised *EATM* tracking metric, that considers the imaging interval and the maximal number of living cells.

Let $S = \{S_1, \ldots, S_L\}$ be the given segmentation ground truth of a time-lapse with $L \in \mathbb{N}$ frames. Then, some tracking-by-detection method T predicts a tracking graph $G = (V, E)$ containing the segmentation detections as nodes, $V = \bigcup_{l \in \{1,\ldots,L\}} S_l$, and links between cell detections as tracking edges, $E \subseteq V \times V$. We define a tracking metric to be a function $m(\cdot, \cdot)$ that compares a predicted tracking graph $\hat{G} = T(S)$ to the ground truth tracking graph $G^\star$ using a metric m with normalized score:

$$0 \leq m(\hat{G}, G^\star) \leq 1. \tag{4}$$

The *TRA*, *LNK* and *DIV* metrics defined before are such metrics.

First, we make such a metric sensitive to experiment parameters. Therefore, we evaluate the metric on temporally subsampled and cell count limited versions of the time-lapses (Fig. 4). Let $k \in \mathbb{N}$ be the subsampling parameter for reducing the temporal resolution. Let $N_{max} \in \mathbb{N}$ be the cell count limit. Then a time-lapse is truncated to the last frame where the cell count does not exceed the limit N_{max} (Fig. 3B). We denote the subsampled and truncated segmentation information with $S|^k_{N_{max}}$ and the ground truth tracking graph with $G^\star|^k_{N_{max}}$, respectively. Then we define the experiment-aware tracking metric (EATM) version $\tilde{m}$ of m

$$\tilde{m}^k_{N_{max}}(T, S, G^\star) := m\left(T(S|^k_{N_{max}}), G^\star|^k_{N_{max}}\right) \tag{5}$$

that evaluates the metric m on tracking prediction for the subsampled and truncated segmentation $T(S|^k_{N_{max}})$ with the subsampled and truncated ground truth tracking $G^\star|^k_{N_{max}}$.

Second, we define a new robustness metric (*RM*) to summarize the robustness of the tracking algorithm across a wide range of imaging intervals and cell count limits. We define a set of subsamplings $SF \subset \mathbb{N}$ and maximum cell counts $MC \subset \mathbb{N}$. The robustness metric *RM* of a metric m measures the normalized frequency that the *EATM* of m surpasses a given threshold $\vartheta \in [0, 1]$:

$$RM(m, \vartheta, SF, MC) := \frac{1}{|SF| \cdot |MC|} \sum_{k \in SF} \sum_{mc \in MC} \mathbb{1}\left[\tilde{m}^k_{mc}(T, S, G^\star) \geq \vartheta\right], \tag{6}$$

where $\mathbb{1}[\,\cdot\,]$ is the indicator function.

5 Tracking Evaluation

We evaluated the performance of tracking methods on our new dataset using the *EATM* and *RM* metrics. For the comparison, we selected three representative tracking methods: The `Distance` method provides a baseline using pure distance information for linking costs and predicting links using a greedy scheme. The `LAP` method uses mask overlap for linking costs and optimizes the linking cost between consecutive frames in an LAP [17]. The `Trackastra` [10] method is the best performing tracker according to the current CTC leaderboard. In our evaluation, `Trackastra` represents the data-driven methods for predicting linking costs.

We evaluated all three tracking methods with various imaging intervals and cell count limits on the TOIAM test split. Figure 4 shows resulting heatmaps of the *EATM* based on the *DIV* metric. Across all three tracking methods, the *DIV* metric decreases for larger imaging intervals as well as higher cell limits, indicating deteriorating performance. The evaluation shows that higher cell numbers, more frequent cell divisions, and lower temporal resolution make the cell tracking task notably more difficult.

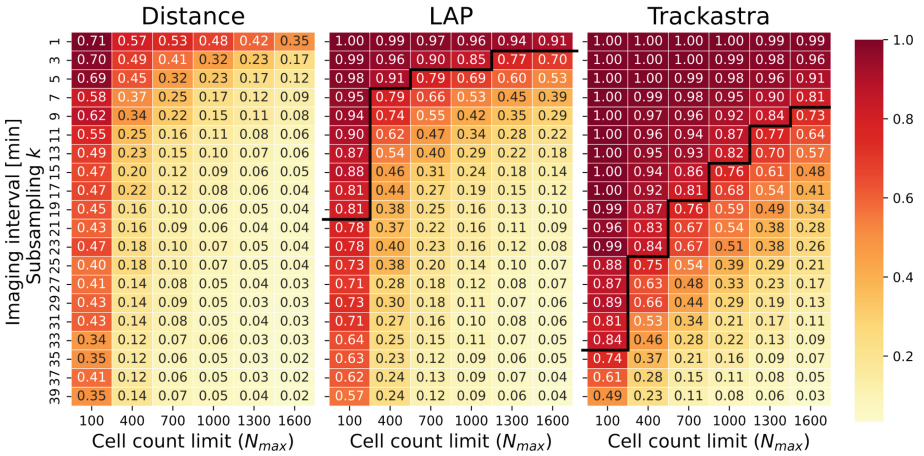

Fig. 4. The *EATM* based on the *DIV* metric measured across different imaging intervals and cell count limits. The black line marks the region surpassing the 80 % threshold (*RM*@0.8). The higher the value of the *DIV* metric, the better is the reconstruction of the cell divisions, with a value of 1 meaning perfect cell division reconstruction. Evaluations have been performed on the test split.

Among all three tracking methods, the `Distance` method performs worst across all experiment parameters. Therefore, making cell linking inferences based purely on distances, is not suitable for MLCI, especially when using a greedy scheme. The `LAP` method performs slightly better in the *DIV* metric at lower imaging intervals, benefiting from the more informative overlap costs and the non-greedy `LAP`. However, its performance drops rapidly when many cells are present and the imaging interval is increased leading to larger cell movement and, therefore, limiting the usefulness of overlap costs. In contrast, `Trackastra` shows much stronger *DIV* scores across the various experiment parameters. The transformer network trained on the training split seems to learn patterns that generalize to higher imaging intervals. Thus, `Trackastra` robustly performs division reconstruction at various experiment parameters, opening the opportunity to monitor more cultivation chambers concurrently, which is important in the context of high-throughput MLCI-based screening applications.

While the *EATM* heatmaps give detailed insights into the methods' performance, they only visualize a single tracking metric. To summarize and compare the methods across different metrics, we summarize the robustness of the tracking method to the experiment parameters using the *RM* metric. Table 3 shows the *RM* score for the *TRA*, *LNK*, and *DIV* metrics using a threshold of 80% and 90%. For the *DIV* metric, the *RM* score is also visualized by the black line in Fig. 4.

We observe that the *TRA* metric is not sensitive enough to give robustness insights when ground truth segmentation masks are provided. The correct segmentation will always lead to scores above 0.8 in all experiment parame-

Table 3. Robustness metric version of the *TRA*, *LNK*, and *DIV* metrics evaluated on the test split using *RM* thresholds of 80% and 90%. The *RM* metric has been computed over the subsampling factors (*SF*) and cell count limits (*MC*) used in Fig. 4.

Method	RM@0.8			RM@0.9		
	TRA ↑	*LNK* ↑	*DIV* ↑	*TRA* ↑	*LNK* ↑	*DIV* ↑
`Distance` (greedy)	**1.00**	0.16	0.00	0.46	0.12	0.00
`LAP`	**1.00**	0.23	0.16	0.47	0.16	0.11
`Trackastra` (greedy)	**1.00**	**0.50**	**0.45**	**0.81**	**0.42**	**0.32**

ters, yielding a misleading robustness score of 1. The *RM*s of *LNK* and *DIV* are more sensitive and provide insights for both robustness thresholds. `LAP` shows low results in *LNK* and *DIV* metrics, highlighting that it is only suitable for tracking cells at low imaging intervals and small cell colony sizes. The `Trackastra` method shows robustness up to 50% at an 80% threshold and, therefore, allows performing automated tracking at various experiment settings. Across all evaluations, the *DIV* metric is consistently lower than the *LNK* metric, indicating that cell division reconstruction is more difficult than linking non-dividing cells. This, underlines the larger focus on predicting cell divisions in MLCI datasets.

6 Conclusions

In this work, we have presented a new benchmark for cell tracking in MLCI with increased experiment awareness in metric ratings. The presented TOIAM dataset is the largest publicly available for MLCI in terms of annotated cell masks, cell tracks and cell divisions. Moreover, we have highlighted that MLCI data comes with unique challenges due to exponentially growing cell colonies and frequent cell divisions. These challenges are strongly influenced by experiment parameters such as the imaging interval and the maximum number of cells per frame. To capture these influences in appropriate metrics, we extended existing metrics towards experiment-awareness (*EATM*) and summarize them in a robustness metric (*RM*). We have shown that these *EATM*s and *RM*s give crucial insights into the practical suitability of tracking methods across a wide range of experiment parameters. Thus, our efforts aim to closely integrate method development and experiment design, and to open a stringent approach for experimenters to make informed decisions about their experiment parameter choices.

For now, our TOIAM dataset is limited to a single type of microbe cultivated within a single experiment. Therefore, we are looking forward to extending the dataset to other cell types and cultivation conditions and also apply the introduced metrics to other existing datasets in the future. Moreover, the evaluation of the tracking methods with erroneous segmentation and imperfect image data as well as the use of other biologically motivated metrics, such as the CT or BIO metrics from the CTC, is crucial for further evaluating their practical applicability. Adapting tracking methods to the challenging imaging conditions in MLCI,

for example, by tuning hyper-parameters or establishing temporal subsampling during the training procedure might lead to more robust cell tracking methods.

Summarizing, our large-scale benchmark represents a step forward towards robust data-driven microbial single-cell tracking and facilitates tight integration of experiment parameters and tracking method development using experiment-aware metrics.

Acknowledgments. This work was supported by the President's Initiative and Networking Funds of the Helmholtz Association of German Research Centres [EMSIG ZT-I-PF-04-044]. JS and KN acknowledge the inspiring scientific environment provided by the Helmholtz School for Data Science in Life, Earth and Energy (HDS-LEE), and thank Wolfgang Wiechert for continuous support. NF, AJYS and RM were supported by the Helmholtz Program NACIP and the Helmholtz Information and Data Science School for Health (HIDDS4Health).

References

1. Anjum, S., Gurari, D.: CTMC: cell tracking with mitosis detection dataset challenge. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 4228–4237. IEEE (2020)
2. Ben-Haim, T., Raviv, T.R.: Graph Neural Network for Cell Tracking in Microscopy Videos (2022). arXiv:2202.04731 [cs]
3. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the CLEAR MOT Metrics. EURASIP J. Image Video Process. **2008**(1), 1–10 (2008). https://doi.org/10.1155/2008/246309
4. Blöbaum, L., Torello Pianale, L., Olsson, L., Grünberger, A.: Quantifying microbial robustness in dynamic environments using microfluidic single-cell cultivation. Microbial Cell Factories, p. 44 (2024)
5. Caicedo, J.C., et al.: Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. Nature Methods, pp. 1247–1253 (2019)
6. Cordts, M., et al.: The Cityscapes Dataset for Semantic Urban Scene Understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
7. Cutler, K.J., et al.: Omnipose: a high-precision morphology-independent solution for bacterial cell segmentation. Nature Methods, pp. 1438—1448 (2022)
8. Edlund, C., et al.: LIVECell-A large-scale dataset for label-free live cell segmentation. Nature Methods, pp. 1038–1045 (2021)
9. Fang, Y., et al.: EVA: exploring the limits of masked visual representation learning at scale (2022). arXiv:2211.07636 [cs]
10. Gallusser, B., Weigert, M.: Trackastra: Transformer-based cell tracking for live-cell microscopy (2024). arXiv:2405.15700 [cs]
11. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The KITTI dataset. The International Journal of Robotics Research, pp. 1231–1237 (2013)
12. Greenwald, N.F., et al.: Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. Nature Biotechnology, pp. 555–565 (2022)
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. arXiv:1703.06870 [cs] (2018). arXiv: 1703.06870

14. Helfrich, S., et al.: Live cell imaging of SOS and prophage dynamics in isogenic bacterial populations. Molecular Microbiology, pp. 636–650 (2015)
15. Hockenberry, A.M., Micali, G., Takács, G., Weng, J., Hardt, W.D., Ackermann, M.: Microbiota-derived metabolites inhibit *Salmonella* virulent subpopulation development by acting on single-cell behaviors. Proceedings of the National Academy of Sciences, p. e2103027118 (2021)
16. Hoebe, R.A., Van Oven, C.H., Gadella, T.W.J., Dhonukshe, P.B., Van Noorden, C.J.F., Manders, E.M.M.: Controlled light-exposure microscopy reduces photobleaching and phototoxicity in fluorescence live-cell imaging. Nature Biotechnology, pp. 249–253 (2007)
17. Jaqaman, K., Loerke, D., Mettlen, M., Kuwata, H., Grinstein, S., Schmid, S.L., Danuser, G.: Robust single-particle tracking in live-cell time-lapse sequences. Nature Methods, pp. 695–702 (2008)
18. Jeckel, H., Drescher, K.: Advances and opportunities in image analysis of bacterial cells and communities. FEMS Microbiology Reviews (2020)
19. Kirillov, A., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4015–4026 (2023)
20. Liu, Y., Wang, Y., Wang, S., Liang, T., Zhao, Q., Tang, Z., Ling, H.: CBNet: A Novel Composite Backbone Network Architecture for Object Detection (2019). arXiv:1909.03625 [cs]
21. Luiten, J., Ošep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B.: HOTA: a higher order metric for evaluating multi-object tracking. Int. J. Comput. Vis., 548–578 (2021)
22. Löffler, K., Mikut, R.: EmbedTrack-simultaneous cell segmentation and tracking through learning offsets and clustering bandwidths. IEEE Access, pp. 77147–77157 (2022)
23. Löffler, K., Scherr, T., Mikut, R.: A graph-based cell tracking algorithm with few manually tunable parameters and automated segmentation error correction. PLOS ONE, p. e0249257 (2021)
24. Matula, P., Maška, M., Sorokin, D.V., Matula, P., Ortiz-de Solórzano, C., Kozubek, M.: Cell tracking accuracy measurement based on comparison of acyclic oriented graphs. PLOS ONE, p. e0144959 (2015)
25. Maška, M., et al.: The Cell Tracking Challenge: 10 years of objective benchmarking. Nature Methods, pp. 1010–1020 (2023)
26. O'Connor, O.M., Alnahhas, R.N., Lugagne, J.B., Dunlop, M.J.: DeLTA 2.0: A deep learning pipeline for quantifying single-cell spatial and temporal dynamics. PLOS Computational Biology, p. e1009797 (2022)
27. Pellegrini, S., Ess, A., Schindler, K., van Gool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 261–268 (2009)
28. Qiao, S., Chen, L.C., Yuille, A.: DetectoRS: Detecting Objects with Recursive Feature Pyramid and Switchable Atrous Convolution (2020). arXiv:2006.02334 [cs]
29. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vision **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
30. Scherr, T., Löffler, K., Böhland, M., Mikut, R.: Cell segmentation and tracking using CNN-based distance predictions and a graph-based matching strategy. PLOS ONE, p. e0243219 (2020)
31. Schuhmann, C., et al.: LAION-5B: an open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems, pp. 25278–25294 (2022)

32. Schwartz, M.S., et al.: Caliban: accurate cell tracking and lineage construction in live-cell imaging experiments with deep learning. bioRxiv (2023)
33. Seiffarth, J., et al.: ObiWan-Microbi: OMERO-based integrated workflow for annotating microbes in the cloud. SoftwareX (2024)
34. Sivaraman, D., Biswas, P., Cella, L.N., Yates, M.V., Chen, W.: Detecting RNA viruses in living mammalian cells by fluorescence microscopy. Trends in Biotechnology, pp. 307–313 (2011)
35. Stringer, C., Wang, T., Michaelos, M., Pachitariu, M.: Cellpose: a generalist algorithm for cellular segmentation. Nature Methods, pp. 100–106 (2021)
36. Theorell, A., Seiffarth, J., Grünberger, A., Nöh, K.: When a single lineage is not enough: Uncertainty-Aware Tracking for spatio-temporal live-cell image analysis. Bioinformatics, pp. 1221–1228 (2019)
37. Täuber, S., Golze, C., Ho, P., Von Lieres, E., Grünberger, A.: dMSCC: a microfluidic platform for microbial single-cell cultivation of *Corynebacterium glutamicum* under dynamic environmental medium conditions. Lab on a Chip, pp. 4442–4455 (2020)
38. Täuber, S., Schmitz, J., Blöbaum, L., Fante, N., Steinhoff, H., Grünberger, A.: How to perform a microfluidic cultivation experiment-a guideline to success. Biosensors, p. 485 (2021)
39. Ulman, V., et al.: An objective comparison of cell-tracking algorithms. Nature Methods, pp. 1141–1152 (2017)
40. Upschulte, E., Harmeling, S., Amunts, K., Dickscheid, T.: Uncertainty-aware contour proposal networks for cell segmentation in multi-modality high-resolution microscopy images. In: Proceedings of The Cell Segmentation Challenge in Multi-modality High-Resolution Microscopy Images, pp. 1–12. PMLR (2023)
41. van Vliet, S., Dal Co, A., Winkler, A.R., Spriewald, S., Stecher, B., Ackermann, M.: Spatially correlated gene expression in bacterial groups: the role of lineage history, spatial gradients, and cell-cell interactions. Cell systems, pp. 496–507.e6 (2018)
42. Wang, Y.H., Hsieh, J.W., Chen, P.Y., Chang, M.C., So, H.H., Li, X.: SMILEtrack: SiMIlarity LEarning for Occlusion-Aware Multiple Object Tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 5740–5748 (2024)
43. Zhang, Y., et al.: ByteTrack: Multi-Object Tracking by Associating Every Detection Box (2022). arXiv:2110.06864 [cs]
44. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: FairMOT: on the fairness of detection and re-identification in multiple object tracking. Int. J. Comput. Vision **129**(11), 3069–3087 (2021). https://doi.org/10.1007/s11263-021-01513-4