# CHOTA: A Higher Order Accuracy Metric for Cell Tracking

Timo Kaiser[1], Vladimír Ulman[2], and Bodo Rosenhahn[1]

[1] Institute for Information Processing – Leibniz University Hannover, Hannover, Germany
{kaiser,rosenhahn}@tnt.uni-hannover.de
[2] IT4Innovations, VSB – Technical University of Ostrava, Ostrava, Czech Republic
vladimir.ulman@vsb.cz

**Abstract.** The evaluation of cell tracking results steers the development of tracking methods, significantly impacting biomedical research. This is quantitatively achieved by means of evaluation metrics. Unfortunately, current metrics favor local correctness and weakly reward global coherence, impeding high-level biological analysis. To also foster global coherence, we propose the CHOTA metric (*Cell-specific Higher Order Tracking Accuracy*) which unifies the evaluation of all relevant aspects of cell tracking: cell detections and local associations, global coherence, and lineage tracking. We achieve this by introducing a new definition of the term 'trajectory' that includes the entire cell lineage and by including this into the well-established HOTA metric from general multiple object tracking. Furthermore, we provide a detailed survey of contemporary cell tracking metrics to compare our novel CHOTA metric and to show its advantages. All metrics are extensively evaluated on state-of-the-art real-data cell tracking results and synthetic results that simulate specific tracking errors. We show that CHOTA is sensitive to all tracking errors and gives a good indication of the biologically relevant capability of a method to reconstruct the full lineage of cells. It introduces a robust and comprehensive alternative to the currently used metrics in cell tracking. Python code is available at https://github.com/CellTrackingChallenge/py-ctcmetrics.

## 1 Introduction

Cell tracking aims to find and cluster cell instances in time-lapse videos to model spatio-temporal cell trajectories and their lineage relations. This automates laborious work in biomedical research and allows extensive research in complex domains, *e.g.* nervous system analysis [22] or spinal cord development [30].

Unfortunately, algorithms from general multiple object tracking (MOT) are usually not easily directly applicable because cells differ in their properties. Different cell instances are similar in their visual appearance, have no smooth movements due to large time steps in the time-lapse, and in particular cells divide during proliferation introducing parental relations between cell instances in the lineage. In contrast to MOT, which has a variety of publicly available datasets

(*e.g.* [1, 11, 13, 31, 38, 41]), there is less publicly available data for cell tracking since imaging and labeling require expensive hardware and biological experts with domain knowledge. To leverage the process of algorithmic development in cell tracking, the ongoing *Cell Tracking Challenge* (CTC) [28] was founded and provides a variety of public training data and annotations as well as undisclosed test annotations that allow a fair comparison on an evaluation server. While the CTC remains the most impacting cell tracking challenge and constantly evaluates the benchmark [27, 40], other public challenges [2] and datasets [19, 25, 33] followed and enrich the diversity and modalities. Also, synthetic data generation pipelines [36, 37, 39] were introduced to generate data.

While the accuracy of methods increases over time, the main quality measures and metrics have not been changed. As discussed in [12], this phenomenon can lead to ongoing improvements in specific aspects of tracking although dismissing other aspects that, in the worst case, are relevant for biomedical practitioners. Current tracking metrics [5, 29] focus on local correctness like detection accuracy and correct frame-to-frame associations. These measures rarely reflect the global perspective on tracking that questions the ability to correctly identify an instance over long temporal periods or lineage relations between cell trajectories. As shown later, there are secondary metrics that evaluate the latter but do not adequately address the local errors. Following [9], an ideal metric should be *sensitive* to all types of errors, *equally* treat errors of the same kind, and *continuously* exploit the metric's outcome interval fully to enable to rank competing methods. Observing that state-of-the-art methods on CTC accomplish metrics close to 100% on mostly all datasets even without correct lineage reconstruction, the current main metrics seem to miss either equality, sensitivity, or continuity.

The current cell tracking evaluation prefers local-centric methods [10, 21, 24] and does not reward conceptual improvements of global-centric approaches [3, 4, 17, 34]. To solve this issue in general MOT, Luiten *et al.* [23] introduced the HOTA metric (*Higher Order Tracking Accuracy*) that holistically unifies local and global tracking correctness by measuring the impact of local decisions on global coherence. HOTA puts global optimal approaches in perspective and promotes the development of global approaches like [8, 14, 42]. Unfortunately, HOTA does not address the biologically relevant proliferation of cells or lineage tracking.

To tackle these challenges, this paper proposes a metric called CHOTA (*Cell-specific Higher Order Tracking Accuracy*) that brings together local correctness, global coherence, and lineage tracking, aiming to present a holistic view of the cell tracking problem. CHOTA builds upon HOTA but follows a novel definition of the term 'trajectory' that includes lineage and thus covers biologically relevant aspects. We conduct extensive experiments on state-of-the-art results on nine (9) CTC datasets and specific error scenarios to reveal insights about metric behaviors. We analyze and compare CHOTA to all relevant metrics in the field and show that CHOTA is the only metric today that *equally* treats relevant errors, is *sensitive* to all aspects of cell tracking, and presents the evaluation in a *continuous* domain. The experimental results are quantitatively and qualitatively summarized in Figure 1 and Table 1, respectively.

**Table 1:** Cell tracking metrics and their *sensitivity* to specific error cases, the *equality* to similar errors, and the *continuity* between 0 and 1. Furthermore, local, global, and lineage evaluation capabilities are rated. Random influences are indicated with (✓).

|  |  | CTC | | | | | | | CTMC | | | | | | | |
|  |  | TRA | DET | LNK | CT | TF | BC($i$) | CCA | MOTA | IDF1 | Prec. | Rec. | MT | ML | HOTA | **CHOTA** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Sensitivity* | False Negatives | ✓ | ✓ | (✓) | ✓ | ✓ | (✓) | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ |
|  | False Positives | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✓ | ✓ |
|  | ID Switches | ✓ | ✗ | ✓ | ✓ | ✓ | ✗ | (✓) | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ |
|  | Missing Matches | ✓ | ✗ | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
|  | Missed Mitosis | ✓ | ✗ | ✓ | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
|  | Equality | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ |
|  | Continuity | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
|  | Local Correctness | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | (✓) | (✓) | ✓ | ✓ |
|  | Global Coherence | ✗ | ✗ | ✗ | ✓ | ✓ | ✗ | (✓) | ✗ | ✓ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ |
|  | Lineage Tracking | ✓ | ✗ | ✓ | ✗ | ✗ | ✓ | (✓) | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |

To compare CHOTA, this paper first offers an in-depth survey of all relevant cell tracking metrics in Section 2. Then, we introduce CHOTA in Section 3 and show experiments to benchmark CHOTA and all relevant metrics in Section 4.

## 2 Quality Measures in Cell Tracking

This section surveys evaluation protocols and quality measures of the most popular cell tracking benchmarks that facilitate fair evaluation conditions due to inaccessible test annotations. We discuss the *'Cell Tracking Challenge'*[3] (CTC) [27] and the *'Cell Tracking with Mitosis Detection Dataset Challenge'*[4] (CTMC) [2] to consolidate formal details of state-of-the-art cell tracking evaluation in one paper. Before presenting the details, preliminaries, and notations are discussed.

All quality measures in cell tracking require a reference set of *cell annotations* (*a.k.a. ground truth*) that correctly describe the real spatial and temporal position of unique cell instances in image sequences and their lineage relations. Usually, the ground truth is organized in spatial instance annotations labeled with IDs. The IDs are consistent over the image sequence and cluster annotations to trajectories (*a.k.a. tracks*). Given only raw images, the task of cell tracking is to create a set of *cell predictions* (also termed as *predicted detections* or just *detections*) with IDs that make up tracks and establish parent-daughter relations, all of which should be equal to the ground truth. The quality of predictions is evaluated using metrics that overall describe the amount disparities to the ground truth, typically with a number between 0 (worst) and 1 (perfect).

From the evaluation point of view, a distinction is made between *tracking*, *segmentation*, and *detection*. While *detection* is based on binary decisions if a cell annotation is sufficiently addressed by a corresponding cell prediction or not, *segmentation* incorporates the prediction quality (typically some form of the accuracy of cell outline). *Tracking* focuses on associations of cell instances to form trajectories.

---

[3] http://celltrackingchallenge.net/
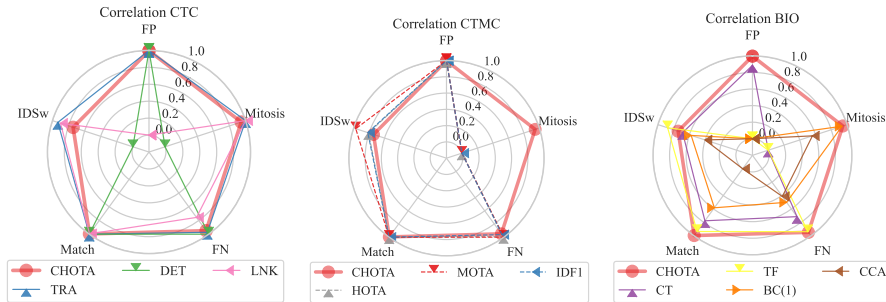
[4] https://motchallenge.net/data/CTMC-v1/

**Fig. 1:** Sensitivity of selected cell tracking metrics to specific tracking errors, measured by the correlation in our experiments. Error types are identity switches (IDSw), false (FP) or missing (FN) detections, unmatched detection-annotation pairs (Match), and missed mitosis detections (Mitosis). From left to right, we present the CTC [27], CTMC [2] with HOTA [23], and biological metrics (BIO) together with CHOTA. Some metrics are less sensitive to a particular type of error (correlation is $\ll 1.0$), or even not sensitive at all (= 0.0; *e.g.* CTMC metrics to mitosis detection errors). CHOTA and TRA are *sensitive* enough to all tracking issues, but CHOTA is the only metric that also rates individual errors by their global impact. Thus, the same amount of IDSw (or Mitosis) errors may influence variously (depending on how much the randomly inserted errors cancel out themselves) the global coherence, which acts as noise to the correlation and is actually visible in the correlation $< 1.0$.

To compare prediction and ground truth, a matching needs to be performed that assigns predicted detections to annotations and defines a set of *True Positive* (TP) pairs $c \in$ TP. For such a pair $c$, the ID of the corresponding detection and annotation are denoted as $\mathrm{prID}(c)$ and $\mathrm{gtID}(c)$, respectively. Unmatched detections are called *False Positives* (FP) and have a corresponding $\mathrm{gtID}(c) = 0$. Similarly, unmatched annotations are called *False Negatives* (FN) for which $\mathrm{prID}(c) = 0$. The parent-daughter relations in the ground truth or predictions are expressed by $p(a) = b$, indicating that a cell with ID $b$ is the parent of $a$.

It is worth noting that we use the term *match* to describe the correspondence between detection and annotation and *association* to describe that multiple objects are assigned to the same ID, as done in [23].

### 2.1 The *'Cell Tracking Challenge'* Measures

The CTC [27] provides multiple datasets to evaluate tracking, detection, and segmentation. For each dataset, two training and two test sequences exist. The training data contains only tens of pixel-wise accurate ground truth annotations for segmentation (*i.e.* not every instance is annotated) and all but simplified annotations for tracking and detection (*i.e.* every instance is annotated with a small circle in every frame).

According to the CTC, a predicted detection, represented as a set of pixels $S_c$, is matched to a ground truth annotation, pixel set $R_c$, and denoted as a $c \in$ TP if the size of the two sets intersection is greater than half the annotation

size. That way, all True Positive matches are as follows

$$\text{TP} = \big\{\, c \;\big|\; |S_c \cap R_c| > 0.5\,|R_c| \big\}. \tag{1}$$

This formulation restricts an annotation to be matched to at most one detection but allows multiple matches for a detection. This captures situations where a mitosis event is detected too late and both daughter cells are matched to one detection. Furthermore, the CTC does not allow temporal gaps in trajectories, leading to fragmentation (change of IDs) if a single instance is not detected.

Following these conventions, the CTC evaluates and ranks submitted approaches in a monthly cycle based on their detection, segmentation, and tracking capabilities. The ranking is done with the following measures.

*Segmentation Accuracy (SEG).* Being the only segmentation metric in this paper, the SEG score evaluates predictions with the Jaccard score $J(R, S)$ [16] (*a.k.a.* IoU). To achieve the best results, a pixel-accurate detection must be matched to each annotation. The SEG metric is defined by an averaged Jaccard score:

$$\text{SEG} = \sum_{c \in \text{TP}} \frac{J(R_c, S_c)}{|\text{TP}| + |\text{FN}|}. \tag{2}$$

*Complete Tracks (CT).* The CT [32, 40] metric describes the fraction of ground truth trajectories that are completely tracked without errors. A complete track has all its annotations assigned to detections with the same ID. Formally, CT is defined as the $F_1$ score over the ground truth and predicted tracks with valid IDs $i$ and $j$:

$$\text{CT} = \frac{2 \cdot \Big|\big\{(i,j) \,\big|\, \{c \in \text{TP} \cup \text{FN} \,|\, \text{gtID}(c) = i\} \subseteq \{c \in \text{TP} \,|\, \text{prID}(c) = j\}\big\}\Big|}{|\text{gtID}| + |\text{prID}|} \tag{3}$$

*Largest Track Fraction (TF).* The TF [40] metric is a relaxation of CT and averages the largest continuously tracked fraction of a ground truth track. It is important to note that TF does not mimic the $F_1$ score as CT and is therefore not sensitive to FP. The metric is defined as

$$\text{TF} = \frac{1}{|\text{gtID}|} \sum_{i \in \text{gtID}} \frac{\max_{j \in \text{prID}} \big|\{c \in \text{TP} \,|\, \text{gtID}(c) = i \;\wedge\; \text{prID}(c) = j\}\big|}{\big|\{c \in \text{TP} \cup \text{FN} \,|\, \text{gtID}(c) = i\}\big|}. \tag{4}$$

*Cell Cycle Accuracy (CCA).* To reflect the ability of an algorithm to discover the true distribution of cell cycle lengths in an image sequence, the CCA [40] metric was introduced. A cell cycle is the time from birth to death/branch and is only defined for tracks that are initiated and terminated by branching events (*i.e.* mitosis) visible in the data. The CCA measure indicates the maximal distance between the normalized cumulative distribution functions $\text{CDF}(t)$ of the predicted and ground truth cell life cycle lengths $t$. The metric is formulated as

$$\text{CCA} = 1 - \max_t \big(|\text{CDF}_{\text{pr}}(t) - \text{CDF}_{\text{gt}}(t)|\big). \tag{5}$$

*Branching Correctness (BC).* To measure the quality of mitosis detection, $\mathrm{BC}(i)$ [6, 26, 40] is the $F_1$ score of the detected branching events. A branching event for a match $c$ is detected if the corresponding tracks $\mathrm{prID}(c)$ and $\mathrm{gtID}(c)$ are both branching in the same frame or with a temporal distance up to $i$ frames. With correctly detected branching events $(\mathrm{BTP}(i))$, undetected events $(\mathrm{BFN}(i))$, and false detections $(\mathrm{BFP}(i))$, the score is defined as

$$\mathrm{BC}(i) = \frac{2\mathrm{BTP}(i)}{2\mathrm{BTP}(i) + \mathrm{BFP}(i) + 2\mathrm{BFN}(i)}. \tag{6}$$

*Acyclic oriented Graph Measure (AOGM).* The AOGM [29] metric counts all tracking and detection errors by spanning a predicted and ground truth graph and aggregating the operations that are necessary to transform the first into the latter. A graph is constructed by adding vertices for every detection/annotation and link-edges to connect vertices that are associated with the same track and are subsequent in time. Furthermore, parent-edges are added between the end vertex of a parent and the start vertex of the daughters. The following operations to transform the predicted graph into the ground truth graph are counted: Remove (FP) or add (FN) a vertex, remove (ED) or add (EA) an edge, split a vertex if a detection is falsely matched to multiple annotations (NS), and alter the semantic from link-edge to parent-edge or vice versa (EC). The measure is defined as the weighted sum of all errors, in which the weights reflect the effort to manually curate the tracking result and correct the respective errors:

$$\mathrm{AOGM} = w_{\mathrm{NS}} \cdot \mathrm{NS} + w_{\mathrm{FN}} \cdot \mathrm{FN} + w_{\mathrm{FP}} \cdot \mathrm{FP} + w_{\mathrm{ED}} \cdot \mathrm{ED} + w_{\mathrm{EA}} \cdot \mathrm{EA} + w_{\mathrm{EC}} \cdot \mathrm{EC}. \tag{7}$$

*Tracking- (TRA), Detection- (DET), and Linking Accuracy (LNK).* To transform AOGM into a metric, the $\mathrm{AOGM}_0$ score is introduced that counts the operations to create the ground truth graph from scratch. The main tracking metric of the CTC is TRA and, to give values between 0 and 1, is defined by the ratio of AOGM and $\mathrm{AOGM}_0$:

$$\mathrm{TRA} = 1 - \frac{\min\{\mathrm{AOGM}, \mathrm{AOGM}_0\}}{\mathrm{AOGM}_0}. \tag{8}$$

To measure the detection quality, DET does not incorporate linking related errors (*i.e.* $w_{\mathrm{ED}} = w_{\mathrm{EA}} = w_{\mathrm{EC}} = 0$). Similar, LNK only measures the linking quality by removing the vertex-related errors (*i.e.* $w_{\mathrm{NS}} = w_{\mathrm{FN}} = w_{\mathrm{FP}} = 0$).

*Biologically relevant (BIO) and Overall Performance (OP).* To comprehensively rank competition submissions, the CTC introduces metrics that harmonize multiple measures. The biologically relevant metrics CT, BC, TF, and CCA are averaged to

$$\mathrm{BIO} = 0.25 \cdot (\mathrm{CT} + \mathrm{BC}(i) + \mathrm{TF} + \mathrm{CCA}). \tag{9}$$

To focus solely on the segmentation or tracking, the DET and SEG metrics, and SEG and TRA are averaged, respectively:

$$\mathrm{OP}_{\mathrm{CSB}} = 0.5 \cdot (\mathrm{DET} + \mathrm{SEG}), \qquad \mathrm{OP}_{\mathrm{CTB}} = 0.5 \cdot (\mathrm{DET} + \mathrm{TRA}). \tag{10}$$

Recently, the $\mathrm{OP_{CLB}}$ metric was added to measure the overall performance of association strategies. To remove the influence of the detection quality, $\mathrm{OP_{CLB}}$ is evaluated on a prepared set of detections and is defined as

$$\mathrm{OP_{CLB}} = 0.5 \cdot (\mathrm{BIO} + \mathrm{LNK}). \tag{11}$$

## 2.2   The *'CTMC-MOTChallenge'* Measures

The CTMC [2] provides a large-scale cell tracking dataset covering 14 cell types. Compared to the CTC, CTMC includes fewer tracks (2900 vs. 11318) but has more frames (152584 vs. 5927). Annotations for cell instances are given as bounding boxes that cause ambiguous associations of pixels to cell instances because a pixel can be included in multiple bounding boxes. The matching procedure between detections and annotations depends on the metric. The CTC matching is used to calculate the TRA metric. On the other hand, Hungarian matching [20] creates a bijective mapping between detections and annotations. The costs for the matching are created using the Jaccard score:

$$\arg\max_{\mathrm{TP}} \sum_{c \in \mathrm{TP}} J(R_c, S_c) \quad \text{with} \quad \forall c \in \mathrm{TP} : \ J(R_c, S_c) > 0.5 \tag{12}$$

A match is typically only allowed, iff the Jaccard score is larger than a threshold, usually 0.5. This strategy does not allow multiple matches to a detection.

Since there are no pixel-wise segmentation masks, the CTMC does not provide a segmentation benchmark. Submitted methods are evaluated with metrics that are sensitive to tracking and detection errors, jointly. The CTMC evaluates submitted results ad-hoc, but evaluations per month is limited to avoid test-data fitting. The next section presents metrics used by the CTMC. Furthermore, FP, FN, and TRA scores introduced in Section 2.1 are also reported by the CTMC.

*Multiple Object Tracking Accuracy (MOTA) and Identity Switches (IDSw).* The main CTMC tracking metric to rank approaches is the MOTA metric [5]. Using IDSw that denote association errors, in which two subsequent detections $S_{c_1}$ and $S_{c_2}$ of a predicted trajectory (*i.e.* $\mathrm{prID}(c_1) = \mathrm{prID}(c_2)$) are matched to different ground truth trajectories (*i.e.* $\mathrm{gtID}(c_1) \neq \mathrm{gtID}(c_2)$), MOTA is defined as:

$$\mathrm{MOTA} = 1 - \frac{|\mathrm{FN}| + |\mathrm{FP}| + |\mathrm{IDSw}|}{|\mathrm{TP}| + |\mathrm{FN}|}. \tag{13}$$

*Identification $F_1$ Score (IDF$_1$).* The IDF$_1$ metric [35] measures the $F_1$ score with TPs, FPs, and FNs but only considers TPs that belong to matched trajectory pairs. The metric assumes that a ground truth trajectory $i$ can only be assigned to a single predicted trajectory $j$, and vice-versa. Matching trajectory pairs are defined by an optimized trajectory assignment $(i, j) \in M$, such that the IDF$_1$ score is maximized. The overall metric is defined as

$$\mathrm{IDF_1} = \max_{M} \left( \frac{\sum_{(i,j) \in M} 2 \cdot \left| \left\{ c \in \mathrm{TP} \mid \mathrm{gtID}(c) = i \ \wedge \ \mathrm{prID}(c) = j \right\} \right|}{2|\mathrm{TP}| + |\mathrm{FN}| + |\mathrm{FP}|} \right). \tag{14}$$

*Recall, Precision, and False Alarms per Frame (FAF).* The detection quality is also evaluated using recall, precision, and the FAF quantity. The FAF quantity is the number of FP normalized over the total number of frames in the image sequence. With an image sequence $\mathcal{I}$, the measures are defined as

$$\text{Precision} = \frac{|\text{TP}|}{|\text{TP}| + |\text{FP}|}, \quad \text{Recall} = \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}|}, \quad \text{and} \quad \text{FAF} = \frac{|\text{FP}|}{|\mathcal{I}|}. \quad (15)$$

*Mostly Tracked (MT) and Mostly Lost (ML).* The CTMC provides the additional criteria MT and ML to evaluate the tracking performance. The measure MT is the ratio of ground-truth trajectories that are covered by a track hypothesis for at least 80% of their respective life span. ML indicates the ratio with at most 20% coverage, respectively.

## 3     Unifying Local Correctness with Global Consistency and Lineage Tracking

Current cell tracking metrics do not follow ideal metric criteria [9]: No metric is *sensitive* to all error types, *equal* to similar errors, and *continuous*. The main metrics analyze local detection quality and the association between subsequent frames (TRA, DET, LNK, MOTA), while ignoring global coherence. Secondary metrics evaluate biological characteristics and global coherence, *i.e.* large trajectory fractions are predicted correctly (CT, TF, CCA, BC, $IDF_1$, MT, ML), but do not adequately address local correctness.

The main focus on locality led to outstanding results in metrics like TRA while simultaneously being bad in biologically relevant metrics. This disparity between local correctness and global coherence was dissolved in general MOT by introducing the HOTA metric [23] that unifies MOTA and $IDF_1$:

$$\text{HOTA} = \sqrt{\frac{\sum_{c \in \text{TP}} \mathcal{A}(c)}{|\text{TP}| + |\text{FN}| + |\text{FP}|}} \quad (16)$$

HOTA treats local errors similar to MOTA but weights each TP individually by its global coherence using an association score $\mathcal{A}(c) \in [0, 1]$ that is defined later.

However, HOTA does not address biologically relevant aspects of cell tracking, namely the cell divisions. To close this gap, we propose CHOTA which unifies HOTA and lineage tracking. In the next sections, we reinvent the term 'trajectory' to include lineage information. Then, we redefine the score $\mathcal{A}(c)$ utilizing our trajectory definition. Both combined result in CHOTA.

### 3.1     Lineage-oriented Trajectories

Tracking metrics evaluate the correctness of predicted trajectories. Across all (cell) tracking literature, a trajectory is defined as a set of objects that share the same ID. Especially in general MOT, this is reasonable and implies that
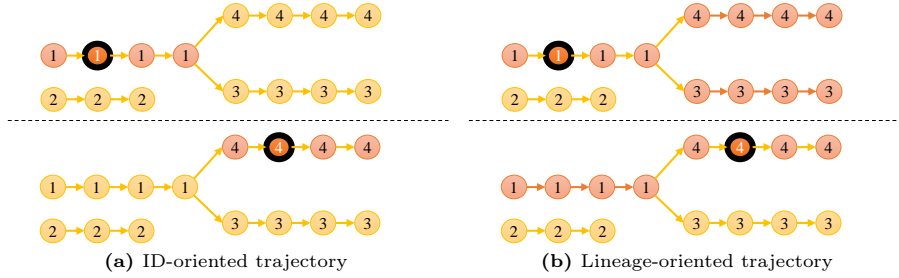
**(a)** ID-oriented trajectory            **(b)** Lineage-oriented trajectory

**Fig. 2:** The common ID-oriented and our lineage-oriented trajectory. A node represents a cell in a specific frame with its respective ID and an edge emphasizes the evolution over time. Each line shows the same lineage tree but highlights a different trajectory (red) for a specific cell detection (black). Our definition accounts for ID-oriented information and also biologically relevant relations to ancestors and descendants.

objects have no relation to objects with different IDs, as indicated with the red trajectories in Figure 2a. However, this does not hold for cell tracking where cells can proliferate and split into multiple daughter cells. Research like embryonic studies [7, 15, 18] not only require the information if two cells have the same ID, they also need to know if they have lineage relations. Thus, an ideal cell tracking metric should evaluate trajectories that include both aspects.

To facilitate lineage information, we redefine the notion of the term 'trajectory': Two cells with IDs $i$ and $j$ belong to the same trajectory **if** $i$ is equal to $j$ **or if** $i$ is an ancestor of $j$ or vice versa. The implications are visualized in Figure 2b where the parent cell trajectory includes both daughters but the daughter trajectory has no relation to its sibling. This reflects that a mutation in ID 1 affects IDs 3 and 4, but ID 4 cannot be affected by ID 3. The trajectory can be formalized using an indicator function $\sigma(i, j)$ and a parent-daughter relation $p^n(a) = b$ which states that cell $b$ is the $n$-th degree ancestor of cell $a$:

$$\sigma(i, j) = \begin{cases} 1, & \text{if } i = j \\ 1, & \text{if } \exists n : \ i = p^n(j) \ \text{or} \ j = p^n(i) \\ 0, & \text{else.} \end{cases} \tag{17}$$

Considering two True Positive detections $c_1 \in \mathrm{TP}$ and $c_2 \in \mathrm{TP}$, the lineage-oriented trajectory indicator can be used to evaluate all relevant cell tracking aspects: By comparing the prediction $\sigma(\mathrm{prID}(c_1), \mathrm{prID}(c_2))$ and ground truth $\sigma(\mathrm{gtID}(c_1), \mathrm{gtID}(c_2))$, local association correctness can be evaluated if $c_1$ and $c_2$ are in subsequent frames, global coherence if they have a temporal gap, and lineage relations if they do not share the same ID.

### 3.2   CHOTA

The score $\mathcal{A}(c)$ from Equation (16) evaluates the global trajectory coherence of a True Positive $c \in \mathrm{TP}$ individually. This is done by comparing the corresponding
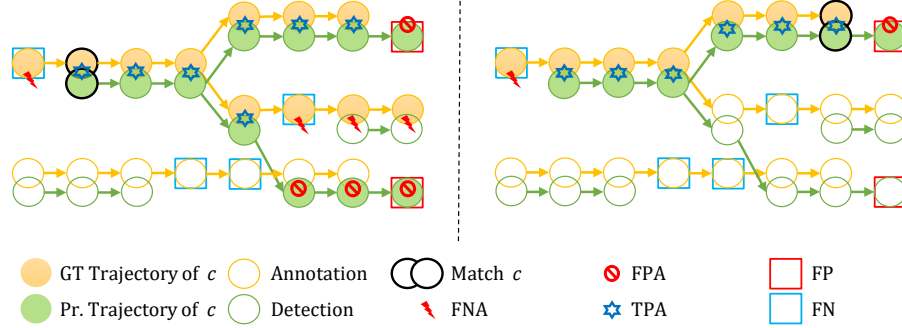
**Fig. 3:** An example prediction (green) to annotation (yellow) mapping presenting the relations to calculate $\mathcal{A}^{\sigma}(c)$ from Equation (18) for two matched pairs $c$ (black). Using our lineage-oriented trajectory, TPA, FPA, and FNA reflect inter- and intra-ID relations. On the left, the ground truth annotation of $c$ relates to two daughter cells of which only one is represented in the prediction, leading to a low $\mathcal{A}^{\sigma}(c) = \frac{7}{15}$. On the right, the lineage of $c$ is almost correct, leading to a higher $\mathcal{A}^{\sigma}(c) = \frac{6}{8}$.

predicted and ground truth trajectories with the Jaccard score [16]. Qualitatively speaking, the score is high if both trajectories are similar over time. To open $\mathcal{A}(c)$ to our trajectory definition, we utilize our formerly introduced indicator $\sigma(i, j)$ and extend the association measures from [23] named *True Positive Associations* (TPA), *False Negative Associations* (FNA), and *False Positive Associations* (FPA), to calculate a lineage-oriented score

$$\mathcal{A}^{\sigma}(c) = \frac{|\text{TPA}^{\sigma}(c)|}{|\text{TPA}^{\sigma}(c)| + |\text{FNA}^{\sigma}(c)| + |\text{FPA}^{\sigma}(c)|}. \tag{18}$$

Using the indicator, the set $\text{TPA}^{\sigma}(c)$ consists of TPs that are associated with the same ground truth and predicted trajectory:

$$\text{TPA}^{\sigma}(c) = \big\{ k \in \text{TP} \,\big|\, \sigma\big(\text{prID}(k), \text{prID}(c)\big) \wedge \sigma\big(\text{gtID}(k), \text{gtID}(c)\big) \big\}. \tag{19}$$

They are indicated with blue stars in Figure 3 and reflect the overlap of the two trajectories, including the entire reconstructed lineage tree. The set $\text{FNA}^{\sigma}(c)$ contains all annotations that are associated with the same ground truth trajectory but matched with a different or no predicted trajectory:

$$\text{FNA}^{\sigma}(c) = \begin{aligned} &\big\{ k \in \text{TP} \,\big|\, \overline{\sigma\big(\text{prID}(k), \text{prID}(c)\big)} \wedge \sigma\big(\text{gtID}(k), \text{gtID}(c)\big) \big\} \\ &\cup \big\{ k \in \text{FN} \,\big|\, \sigma\big(\text{gtID}(k), \text{gtID}(c)\big) \big\}. \end{aligned} \tag{20}$$

Similarly, $\text{FPA}^{\sigma}(c)$ is the union of detections associated with the same predicted trajectory but matched with a different or no ground truth trajectory:

$$\text{FPA}^{\sigma}(c) = \begin{aligned} &\big\{ k \in \text{TP} \,\big|\, \sigma\big(\text{prID}(k), \text{prID}(c)\big) \wedge \overline{\sigma\big(\text{gtID}(k), \text{gtID}(c)\big)} \big\} \\ &\cup \big\{ k \in \text{FP} \,\big|\, \sigma\big(\text{prID}(k), \text{prID}(c)\big) \big\}. \end{aligned} \tag{21}$$

Indicated with red markers in Figure 3, FNAs and FPAs count the missing and incorrect parts of the predicted trajectory, respectively.

By applying the new trajectory definition to $\mathcal{A}^\sigma(c)$, CHOTA is defined as

$$\text{CHOTA} = \sqrt{\frac{\sum_{c \in \text{TP}} \mathcal{A}^\sigma(c)}{|\text{TP}| + |\text{FN}| + |\text{FP}|}} \tag{22}$$

The new lineage-oriented trajectory allows $\mathcal{A}^\sigma(c)$ to reward or penalize all relevant cell tracking events. Similar to HOTA and MOTA, CHOTA penalizes local errors like FPs and FNs in its denominator. Extending TPA, FPA, and FNA *equally* to all detections that have lineage relations to $c$ makes the reward $\mathcal{A}^\sigma(c)$ also *sensitive* to lineage errors. The TPs are individually rewarded relative to their global trajectory coherence and thus implicitly incorporate ID switches and mitosis detections. The individual $\mathcal{A}^\sigma(c)$ is conditioned on the broader impact as shown in Figure 3: While not traversing the coherence in the right trajectory, the wrong association in one daughter cell gets penalized in the left trajectory because it mistakenly creates a lineage relation to an independent cell instance.

It is worth noting that CHOTA can be calculated efficiently in a quasi-linear dependency on the number of IDs as shown in the Appendix. Equivalent to HOTA [23], CHOTA can be decomposed into the detection (DetA) and association (AssA) sub-metrics which are not further elaborated in this paper. Furthermore, we use the CTC matching procedure from Section 2.1. Other strategies like bijective matching from Section 2.2 or HOTAs $\alpha$-integral optimization [23] can be applied, too, but require accurate segmentation annotations that are not always available, *e.g.* in CTC.

## 4    Experiments and Discussion

This section presents insights about CHOTA and the metrics introduced in Section 2. First, a benchmark of all metrics is created on real-world tracking results showing that CHOTA unifies the outcome of other metrics and is the most holistic measure. Then, specific error types are induced synthetically into perfect tracking results to analyze the influence of typical tracking failure cases on the metrics. The experimental results are qualitatively summarized in Table 1.

### 4.1    Real-World Tracking

To evaluate metric behaviors in real tracking scenarios, we re-evaluated all metrics on the state-of-the-art method *EmbedTrack* [24]. To this end, we applied and evaluated the publicly available framework on the training data of 9 diverse datasets from the CTC. The selection of datasets contains various scenarios with long, short, dense, and sparsely populated image sequences with weak to heavy proliferation. The results and relevant dataset properties are given in Table 2.

The biologically relevant metrics that reflect global tracking and lineage consistency are very small (*e.g.* < 5%) on long datasets with large proliferation

**Table 2:** Benchmark on all tracking measures used by CTC [27] and CTMC [2], as well as HOTA [23] and our CHOTA metric, calculated on the state-of-the-art results of *EmbedTrack* [24]. CHOTA unifies both, general and biological tracking aspects.

| | | BF-C2DL-HSC | BF-C2DL-MuSC | DIC-C2DH-HeLa | Fluo-C2DL-MSC | Fluo-N2DH-GOWT1 | Fluo-N2DL-HeLa | PhC-C2DH-U373 | PhC-C2DL-PSC | Fluo-N2DH-SIM+ |
|---|---|---|---|---|---|---|---|---|---|---|
| Properties | Frames | 3526 | 2750 | 166 | 94 | 182 | 182 | 228 | 598 | 213 |
| | Instances | 73178 | 12706 | 2150 | 616 | 4575 | 34059 | 1457 | 128207 | 5978 |
| | Branches | 168 | 44 | 13 | 0 | 6 | 303 | 0 | 992 | 72 |
| | Tracks | 340 | 119 | 70 | 25 | 86 | 939 | 20 | 2395 | 202 |
| | Cell Cycles | 165 | 39 | 0 | 0 | 0 | 151 | 0 | 783 | 36 |
| CTC-Metrics | SEG ↑ | 0.883 | 0.778 | 0.889 | 0.643 | 0.942 | 0.884 | 0.901 | 0.766 | 0.837 |
| | CT ↑ | 0.032 | 0.017 | 0.329 | 0.020 | 0.529 | 0.439 | 0.251 | 0.194 | 0.574 |
| | TF ↑ | 0.721 | 0.596 | 0.857 | 0.692 | 0.960 | 0.933 | 0.920 | 0.871 | 0.923 |
| | BC(1) ↑ | 0.451 | 0.163 | 0.493 | - | 0 | 0.625 | - | 0.513 | 0.827 |
| | CCA ↑ | 0.108 | 0.051 | - | - | - | 0.615 | - | 0.498 | 0.628 |
| | BIO(1) ↑ | 0.328 | 0.207 | 0.560 | 0.356 | 0.496 | 0.653 | 0.585 | 0.519 | 0.738 |
| | TRA ↑ | 0.976 | 0.957 | 0.952 | 0.908 | 0.977 | 0.982 | 0.960 | 0.962 | 0.973 |
| | DET ↑ | 0.974 | 0.962 | 0.953 | 0.914 | 0.978 | 0.983 | 0.963 | 0.963 | 0.975 |
| | LNK ↑ | 0.988 | 0.925 | 0.945 | 0.866 | 0.976 | 0.977 | 0.938 | 0.950 | 0.957 |
| | $OP_{CSB}$ ↑ | 0.929 | 0.870 | 0.921 | 0.778 | 0.960 | 0.934 | 0.932 | 0.865 | 0.906 |
| | $OP_{CTB}$ ↑ | 0.930 | 0.867 | 0.921 | 0.775 | 0.960 | 0.933 | 0.930 | 0.864 | 0.905 |
| | $OP_{CLB}$ ↑ | 0.658 | 0.566 | 0.753 | 0.611 | 0.736 | 0.815 | 0.762 | 0.735 | 0.848 |
| CTMC-Metrics | TP ↑ | 36583 | 6321 | 1039 | 299 | 2241 | 16976 | 727 | 63538 | 2925 |
| | FP ↓ | 3617 | 1705 | 137 | 163 | 81 | 2221 | 161 | 14562 | 46 |
| | FN ↓ | 6 | 32 | 36 | 9 | 46 | 53 | 1 | 565 | 63 |
| | IDSW ↓ | 421 | 335 | 10 | 25 | 5 | 199 | 1 | 1054 | 23 |
| | Precision ↑ | 0.819 | 0.783 | 0.887 | 0.632 | 0.964 | 0.890 | 0.821 | 0.814 | 0.985 |
| | Recall ↑ | 1.000 | 0.995 | 0.966 | 0.973 | 0.981 | 0.997 | 0.998 | 0.991 | 0.979 |
| | FAF ↓ | 2.1 | 1.3 | 1.6 | 3.4 | 0.9 | 24.7 | 1.6 | 50.8 | 0.6 |
| | MT ↑ | 0.428 | 0.289 | 0.810 | 0.383 | 0.934 | 0.897 | 0.929 | 0.793 | 0.879 |
| | ML ↓ | 0.002 | 0.011 | 0 | 0 | 0 | 0.001 | 0 | 0 | 0 |
| | MOTA ↑ | 0.736 | 0.649 | 0.830 | 0.284 | 0.942 | 0.860 | 0.751 | 0.738 | 0.952 |
| | IDF1 ↑ | 0.599 | 0.529 | 0.843 | 0.535 | 0.962 | 0.836 | 0.829 | 0.785 | 0.943 |
| | HOTA | 0.688 | 0.617 | 0.851 | 0.576 | 0.959 | 0.863 | 0.843 | 0.809 | 0.944 |
| | **CHOTA** | 0.542 | 0.451 | 0.849 | 0.630 | 0.947 | 0.865 | 0.859 | 0.745 | 0.959 |

(BF-C2DL-HSC/MuSC) and relatively high on small datasets (*e.g.* Fluo-N2DH-SIM+). Simultaneously, the more relevant and local-centric CTC main tracking metrics TRA, DET and LNK, are consistently close to 1 in mostly all results. This demonstrates the imbalance in cell tracking algorithm development and mistakenly could lead to the belief that the tracking task is close to being solved.

Taking a closer look at the CTMC metrics MOTA or IDF1 reveals that they utilize a larger value range. Low MOTA scores indicate a large number of local errors (FP, FN, and IDSw), and low IDF1 shows that these local errors are not recovered globally over time. As shown in general MOT, the HOTA metric harmonizes local and global error measures [23]. The range of values is from 58% to 96% and indicates that the general tracking task is between half and close to being solved, leaving much space for algorithmic improvements. While MOTA, IDF1, and HOTA are sensitive to local, global, or both tracking aspects, they are insensitive to lineage tracking by definition and discard valuable information.

Our proposed lineage-oriented CHOTA metric is also sensitive to biologically relevant lineage information. While CHOTA is close to HOTA on the small datasets where lineage information is weak, it decreases drastically on datasets with difficult trajectories where lineage construction fails according to the biological metrics (BF-C2DL-HSC/MuSC, and PhC-C2DL-PSC). Furthermore, CHOTA is slightly higher than HOTA on Fluo-C2DL-MSC. The cause is revealed when having a closer look into the results: The ground truth contains
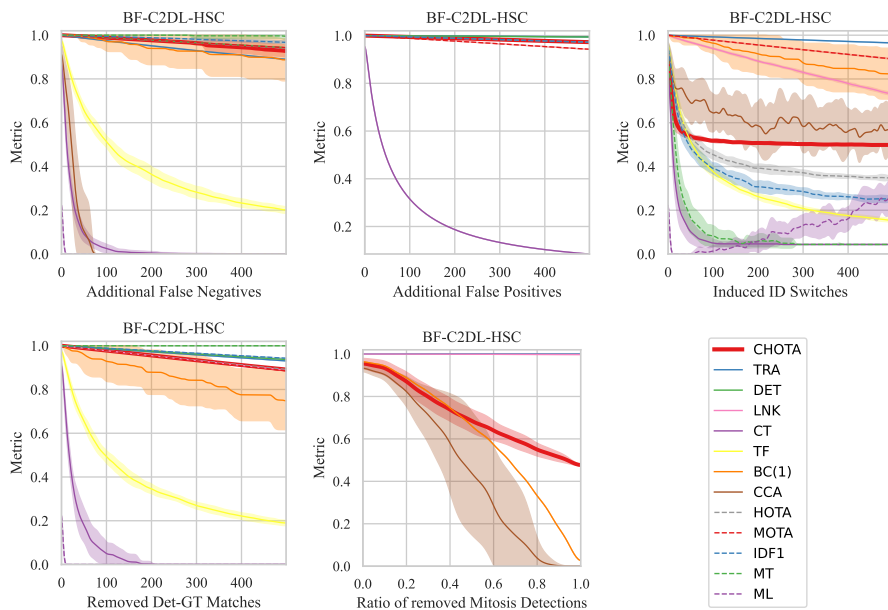
**Fig. 4:** The behavior of metrics if specific tracking errors are induced into a perfect tracking result (BF-C2DL-HSC, Seq. 01). The biological metrics (CT, TF, BC, CCA) are not *equally* treating similar errors, visible in the large variances. TRA is not adequately reflecting the main tracking issues of ID switches and missed mitosis detections, apparent in consistent values close to 1. CHOTA is the only metric that is both, *equal* and *continuously* exploiting a large value range when tracking errors are induced.

no mitotic events but the tracking results falsely detected over-segmentations as mitosis, *i.e.* two detections (TP and FP) are matched to a single annotation. While this destroys an entire trajectory in HOTA, biological relations still hold in the novel lineage-oriented trajectory of CHOTA leading to the better evaluation. This showcases the ability of CHOTA to be sensitive to all cell tracking aspects and to provide a holistic evaluation. Moreover, it continuously employs a larger value range that helps to interpret the current state of different datasets.

## 4.2   Random Tracking Error Induction

We evaluate the influence of specific error sources to perfect tracking results. To do so, we use the ground truth as tracking result and randomly induce errors. We randomly add False Positive detections (FP), remove detections (FN), remove correct detection-to-ground truth matchings, remove parent-daughter relations (mitosis detections), and induce ID switches. Only errors of one type are added to the result at the same time and every experiment is repeated at least 10 times with different seeds. We use the same datasets as in Section 4.1 and visualize the

smoothed mean and variance of the metrics on BF-C2DL-HSC in Figure 4. The results on all other datasets are comparable and can be found in the Appendix.

In Figure 4, variance reflects random influence, meaning that similar errors are not treated *equally*. It shows that MT, ML, CT, TF, CCA, and BC($i$) are prone to local randomness, while local metrics like TRA, DET, LNK, and MOTA have no variance by their definition. IDF1, HOTA, and CHOTA only incorporate small randomness on ID switches and mitosis. The randomness is caused by errors that dissolve former errors and lead to more global coherence.

Our CHOTA metric is the only metric that is adequately *sensitive* and *continuous* to general tracking and mitosis detection errors. While TRA and LNK remain close to 100% even if all (!) mitosis detections are removed, CHOTA decreases to 50%, showing the harmonization of general and lineage tracking. Although the BC($i$) metric is *sensitive* and *continuous* to mitosis detection by its definition, it is just randomly influenced by all other tracking errors.

To summarize the effect of specific error types on the metrics, we visualize the linear correlation of the number of induced errors to the metrics in Figure 1. For visualization purposes, we present the magnitude of the correlation averaged over all datasets. Our CHOTA metric covers all cell tracking relevant aspects with a relatively large correlation. The only metric that also correlates to all error sources is the TRA metric. The correlation to ID switches is larger than by CHOTA which reflects CHOTAs ability to reward coherence-promoting ID switches. With few exceptions like CT to FP and BC($i$) to Mitosis, there is a low correlation between biologically relevant metrics and cell tracking errors, underlining the random influence in the evaluation. Overall, CHOTA is the only metric that addresses all cell tracking failure cases and global tracking coherence.

## 5    Conclusion

This work addresses the problem of unbalanced evaluation metrics in cell tracking that mostly focus on local correctness and do not appropriately reflect global coherence and lineage tracking. To tackle this problem, we introduce the novel metric CHOTA - the *Cell-specific Higher Order Tracking Accuracy* - that unifies local and global tracking aspects as well as biologically relevant lineage consistency. Furthermore, we provide an in-depth survey and analysis of all relevant cell tracking metrics that help researchers in their choice of metrics. Our analysis shows that CHOTA is the only metric that is *sensitive* and *equal* to all relevant cell tracking error types while *continuously* utilizing the full value range. We hope that CHOTA steers algorithmic development more sustainably by incorporating local correctness, global coherence, and lineage information. Addressing all of them in one metric will hopefully lead to advancements that also address all facets of cell tracking.

## 6    Acknowledgements

## References

1. Abeysinghe, C., Reid, C., Rezatofighi, H., Meyer, B.: Tracking different ant species: an unsupervised domain adaptation framework and a dataset for multi-object tracking. arXiv preprint arXiv:2301.10559 (2023)
2. Anjum, S., Gurari, D.: CTMC: Cell tracking with mitosis detection dataset challenge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2020)
3. Bao, R., Al-Shakarji, N.M., Bunyak, F., Palaniappan, K.: Dmnet: Dual-stream marker guided deep network for dense cell segmentation and lineage tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2021)
4. Ben-Haim, T., Raviv, T.R.: Graph neural network for cell tracking in microscopy videos. In: Proceedings of the IEEE/CVF European Conference on Computer Vision. Springer (2022)
5. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: The clear mot metrics. EURASIP journal on image and video processing (2008)
6. Bise, R., Yin, Z., Kanade, T.: Reliable cell tracking by global data association. In: Proceedings of IEEE International Symposium on Biomedical Imaging (2011)
7. Bondarenko, V., Nikolaev, M., Kromm, D., Belousov, R., Wolny, A., Blotenburg, M., Zeller, P., Rezakhani, S., Hugger, J., Uhlmann, V., Hufnagel, L., Kreshuk, A., Ellenberg, J., van Oudenaarden, A., Erzberger, A., Lutolf, M.P., Hiiragi, T.: Embryo-uterine interaction coordinates mouse embryogenesis during implantation. The EMBO Journal (2023)
8. Cetintas, O., Brasó, G., Leal-Taixé, L.: Unifying short and long-term tracking with graph hierarchies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
9. Chen, L., Wu, Y., Stegmaier, J., Merhof, D.: Sortedap: Rethinking evaluation metrics for instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2023)
10. Chen, Y., Song, Y., Zhang, C., Zhang, F., O'Donnell, L., Chrzanowski, W., Cai, W.: Celltrack r-cnn: A novel end-to-end deep neural network for cell segmentation and tracking in microscopy images. In: Proceedings of the IEEE International Symposium on Biomedical Imaging. IEEE (2021)
11. Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., Leal-Taixé, L.: Mot20: A benchmark for multi object tracking in crowded scenes. arXiv preprint arXiv:2003.09003 (2020)

12. Eisenmann, M., Reinke, A., Weru, V., Tizabi, M.D., Isensee, F., Adler, T.J., Ali, S., Andrearczyk, V., Aubreville, M., Baid, U., Bakas, S., Balu, N., Bano, S., Bernal, J., Bodenstedt, S., Casella, A., Cheplygina, V., Daum, M., de Bruijne, M., Depeursinge, A., Dorent, R., Egger, J., Ellis, D.G., Engelhardt, S., Ganz, M., Ghatwary, N., Girard, G., Godau, P., Gupta, A., Hansen, L., Harada, K., Heinrich, M.P., Heller, N., Hering, A., Huaulmé, A., Jannin, P., Kavur, A.E., Kodym, O., Kozubek, M., Li, J., Li, H., Ma, J., Martín-Isla, C., Menze, B., Noble, A., Oreiller, V., Padoy, N., Pati, S., Payette, K., Rädsch, T., Rafael-Patiño, J., Bawa, V.S., Speidel, S., Sudre, C.H., van Wijnen, K., Wagner, M., Wei, D., Yamlahi, A., Yap, M.H., Yuan, C., Zenk, M., Zia, A., Zimmerer, D., Aydogan, D.B., Bhattarai, B., Bloch, L., Brüngel, R., Cho, J., Choi, C., Dou, Q., Ezhov, I., Friedrich, C.M., Fuller, C.D., Gaire, R.R., Galdran, A., Faura, A.G., Grammatikopoulou, M., Hong, S., Jahanifar, M., Jang, I., Kadkhodamohammadi, A., Kang, I., Kofler, F., Kondo, S., Kuijf, H., Li, M., Luu, M., Martinčič, T., Morais, P., Naser, M.A., Oliveira, B., Owen, D., Pang, S., Park, J., Park, S.H., Plotka, S., Puybareau, E., Rajpoot, N., Ryu, K., Saeed, N., Shephard, A., Shi, P., Štepec, D., Subedi, R., Tochon, G., Torres, H.R., Urien, H., Vilaça, J.a.L., Wahid, K.A., Wang, H., Wang, J., Wang, L., Wang, X., Wiestler, B., Wodzinski, M., Xia, F., Xie, J., Xiong, Z., Yang, S., Yang, Y., Zhao, Z., Maier-Hein, K., Jäger, P.F., Kopp-Schneider, A., Maier-Hein, L.: Why is the winner the best? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
13. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2012)
14. Hornakova*, A., Kaiser*, T., Rolinek, M., Rosenhahn, B., Swoboda, P., Henschel, R., equal contribution), .: Making higher order mot scalable: An efficient approximate solver for lifted disjoint paths. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
15. Ichikawa, T., Zhang, H.T., Panavaite, L., Erzberger, A., Fabrèges, D., Snajder, R., Wolny, A., Korotkevich, E., Tsuchida-Straeten, N., Hufnagel, L., et al.: An ex vivo system to study cellular dynamics underlying mouse peri-implantation development. Developmental cell (2022)
16. Jaccard, P.: Étude comparative de la distribution florale dans une portion des alpes et des jura. Bull Soc Vaudoise Sci Nat (1901)
17. Kaiser, T., Schier, M., Rosenhahn, B.: Cell tracking according to biological needs–strong mitosis-aware random-finite sets tracker with aleatoric uncertainty. arXiv preprint arXiv:2403.15011 (2024)
18. Kaufmann, A., Mickoleit, M., Weber, M., Huisken, J.: Multilayer mounting enables long-term imaging of zebrafish development in a light sheet microscope. Development (2012)
19. Ker, D.F.E., Eom, S., Sanami, S., Bise, R., Pascale, C., Yin, Z., Huh, S.i., Osuna-Highley, E., Junkers, S.N., Helfrich, C.J., et al.: Phase contrast time-lapse microscopy datasets with automated and manual cell tracking annotations. Scientific Data (2018)
20. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Research Logistics Quarterly (1955)
21. Löffler, K., Scherr, T., Mikut, R.: A graph-based cell tracking algorithm with few manually tunable parameters and automated segmentation error correction. PloS one (2021)

22. Lovas, J.R., Yuste, R.: Ensemble synchronization in the reassembly of hydra's nervous system. Current Biology (2021)
23. Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B.: Hota: A higher order metric for evaluating multi-object tracking. International Journal of Computer Vision (2020)
24. Löffler, K., Mikut, R.: Embedtrack—simultaneous cell segmentation and tracking through learning offsets and clustering bandwidths. IEEE Access (2022)
25. Ma, J., Xie, R., Ayyadhury, S., Ge, C., Gupta, A., Gupta, R., Gu, S., Zhang, Y., Lee, G., Kim, J., et al.: The multimodality cell segmentation challenge: toward universal solutions. Nature methods (2024)
26. Magnusson, K.E.G., Jaldén, J., Gilbert, P.M., Blau, H.M.: Global linking of cell tracks using the viterbi algorithm. IEEE Transactions on Medical Imaging (2015)
27. Maška, M., Ulman, V., Delgado-Rodriguez, P., Gómez de Mariscal, E., Necasova, T., Guerrero Peña, F.A., Ing Ren, T., Meyerowitz, E., Scherr, T., Löffler, K., Mikut, R., Guo, T., Wang, Y., Allebach, J., Bao, R., Al-Shakarji, N., Rahmon, G., Toubal, I.E., Palaniappan, K., Ortiz-de Solorzano, C.: The cell tracking challenge: 10 years of objective benchmarking. Nature Methods (2023)
28. Maška, M., Ulman, V., Svoboda, D., Matula, P., Matula, P., Ederra, C., Urbiola, A., España, T., Venkatesan, S., Balak, D.M., et al.: A benchmark for comparison of cell tracking algorithms. Bioinformatics (2014)
29. Matula, P., Maška, M., Sorokin, D.V., Matula, P., Ortiz-de Solorzano, C., Kozubek, M.: Cell tracking accuracy measurement based on comparison of acyclic oriented graphs. PLOS ONE (2015)
30. May, M., Denecke, B., Schroeder, T., Götz, M., Faissner, A.: Cell tracking in vitro reveals that the extracellular matrix glycoprotein tenascin-c modulates cell cycle length and differentiation in neural stem/progenitor cells of the developing mouse spinal cord. Biology open (2018)
31. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016)
32. Miller, E., Chen, M., Kanade, T., Weiss, L., Campbell, P.: Cell population tracking and lineage construction with spatiotemporal context. Medical Image Analysis (2008)
33. Moen, E., Borba, E., Miller, G., Schwartz, M., Bannon, D., Koe, N., Camplisson, I., Kyme, D., Pavelchek, C., Price, T., et al.: Accurate cell tracking and lineage construction in live-cell imaging experiments with deep learning. Biorxiv (2019)
34. Nguyen, T.T.D., Vo, B.N., Vo, B.T., Kim, D.Y., Choi, Y.S.: Tracking cells and their lineages via labeled random finite sets. IEEE Transactions on Signal Processing (2021)
35. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: Proceedings of the IEEE/CVF European Conference on Computer Vision. Springer (2016)
36. Sorokin, D.V., Peterlík, I., Ulman, V., Svoboda, D., Nečasová, T., Morgaenko, K., Eiselleová, L., Tesařová, L., Maška, M.: Filogen: a model-based generator of synthetic 3-d time-lapse sequences of single motile cells with growing and branching filopodia. IEEE Transactions on Medical Imaging (2018)
37. Sturm, M., Cerrone, L., Hamprecht, F.A.: Syncellfactory: Generative data augmentation for cell tracking. arXiv preprint arXiv:2404.16421 (2024)
38. Sun, P., Cao, J., Jiang, Y., Yuan, Z., Bai, S., Kitani, K., Luo, P.: Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)

39. Svoboda, D., Ulman, V.: Mitogen: a framework for generating 3d synthetic time-lapse sequences of cell populations in fluorescence microscopy. IEEE Transactions on Medical Imaging (2016)
40. Ulman, V., Maška, M., Magnusson, K.E., Ronneberger, O., Haubold, C., Harder, N., Matula, P., Matula, P., Svoboda, D., Radojevic, M., et al.: An objective comparison of cell-tracking algorithms. Nature methods (2017)
41. Xu, N., Yang, L., Fan, Y., Yang, J., Yue, D., Liang, Y., Price, B., Cohen, S., Huang, T.: Youtube-vos: Sequence-to-sequence video object segmentation. In: Proceedings of the IEEE/CVF European conference on computer vision (2018)
42. You, S., Yao, H., Bao, B.k., Xu, C.: Utm: A unified multiple object tracking model with identity-aware feature enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)

# A  Appendix

## A.1  CHOTA Implementation

The CHOTA metric can be implemented efficiently. It is separated into two steps: First, object detections are matched to ground truth annotations. Second, the TPA, FPA and FNA for every True Positive are calculated and summed up. The next section describes both steps in detail and shows that the second step is very efficient in practice. We use the notation of Section 2 and 3.

*Matching.* As discussed in Section 2, every tracking metric requires a precedent matching of predicted detections to ground truth annotations. Depending on the protocol, the matching can directly be deduced from the intersection of the predicted and ground truth masks (see Section 2.1) or with a subsequent bilinear matching (see Section 2.2). Practically, the matching procedure is computationally intensive: Every ground truth and predicted mask needs to be loaded from disk and compared pixel-by-pixel. When neglecting the bilinear mapping, the matching process is basically the calculation of the intersection of instance masks and therefore linear depending on the dataset characteristics, *i.e.* the number of frames and size of the images. However, since the matching needs to be performed for every metric, it is not specific for CHOTA.

*Association Scores.* After the matching procedure, We have TP, FP, and FN. Furthermore, we have lineage relations $L^{gt}$ and $L^{pr}$ between IDs according to our trajectory definition in Equation (17). The lineage relations can be expressed as a mapping $l(i) = \{i, j_1, \ldots, j_n\}$ that maps a set of $n+1$ IDs to an ID $i$. Every ID $i$ has at least a relation to itself and to $n$ other IDs depending on the lineage tree. For example, the trajectories in Figure 2b have mappings $l(1) = \{1, 3, 4\}$, $l(2) = \{2\}$, $l(3) = \{1, 3\}$, and $l(4) = \{1, 4\}$.

These sets are used to create a prediction-to-ground-truth mapping accumulator $M \in \mathcal{R}^{(I+1)\times(J+1)}$ with the number of predicted and ground truth IDs $I = |\text{PrID}|$ and $J = |\text{GtID}|$. An element $M_{ij}$ reflects the number of TPs matched to the $i$-th predicted ID and $j$-th ground truth ID. Furthermore, row 0 is used to count FNs, such that $M_{0j}$ contains the number of $|\{c \in \text{FN} \,|\, \text{gtID}(c) = j\}|$ of ID $j$. Similarly, column 0 aggregates FPs, such that $M_{i0}$ contains the number of $|\{c \in \text{FP} \,|\, \text{prID}(c) = i\}|$ of the predicted ID $i$. The number of operations to fill the accumulator is linear depending on the number of detections and annotations, *i.e.* there are $|\text{TP}| + |\text{FP}| + |\text{FN}|$ additions required. The resulting accumulator $M$ is usually extremely sparse because most predicted trajectories only overlap with a single or up to a few ground truth trajectories.

Now, the matching score $\mathcal{A}^\sigma(c)$ from Equation (18) for a True Positive $c \in \text{TP}$ can be calculated using $M$, $L^{gt}$ and $L^{pr}$. For an arbitrary $c$ with $\text{PrID}(c) = i$ and $\text{GtID}(c) = j$, the association measures FPA, FNA and TPA can be calculated as follows:

$$\text{TPA}^\sigma(c) = \sum_{i' \in l(i)} \sum_{j' \in l(j)} M_{i'j'} \tag{23}$$

---

**Algorithm 1** Calculate CHOTA

---

**Require:** TP, FP, FN, $L^{\mathrm{pr}}$ $L^{\mathrm{gt}}$
1:  $I, J \leftarrow |\mathrm{PrID}|, |\mathrm{GtID}|$
2:  $M \leftarrow \{0\}^{(I+1)\times(J+1)}$
3:  **for** $c \in \mathrm{TP}$ **do**                     ▷ Accumulate True Positives
4:      $i \leftarrow \mathrm{PrID}(c)$
5:      $j \leftarrow \mathrm{GtID}(c)$
6:      $M_{ij} \leftarrow M_{ij} + 1$
7:  **end for**
8:  **for** $c \in \mathrm{FP}$ **do**                     ▷ Accumulate False Positives
9:      $i \leftarrow \mathrm{PrID}(c)$
10:      $M_{i0} \leftarrow M_{i0} + 1$
11:  **end for**
12:  **for** $c \in \mathrm{FN}$ **do**                    ▷ Accumulate False Negatives
13:      $j \leftarrow \mathrm{GrID}(c)$
14:      $M_{0j} \leftarrow M_{0j} + 1$
15:  **end for**
16:  $\mathcal{A}^{\sigma} \leftarrow 0$                  ▷ Aggregate Matching Scores $\mathcal{A}^{\sigma}(c)$
17:  **for** $M_{ij} \in \{M_{ij} \mid M_{ij} > 0;\ i > 0;\ j > 0\}$ **do**
18:      $\mathrm{TPA}^{\sigma} \leftarrow \sum_{i' \in l(i)} \sum_{j' \in l(j)} M_{i'j'}$
19:      $\mathrm{FPA}^{\sigma} \leftarrow \sum_{i' \in l(i)} \sum_{j'=0}^{J+1} M_{i'j'} - \mathrm{TPA}^{\sigma}$
20:      $\mathrm{TPA}^{\sigma} \leftarrow \sum_{j' \in l(j)} \sum_{i'=0}^{J+1} M_{i'j'} - \mathrm{TPA}^{\sigma}$
21:      $\mathcal{A}^{\sigma} \leftarrow \mathcal{A}^{\sigma} + M_{ij} \cdot \frac{\mathrm{TPA}^{\sigma}}{\mathrm{TPA}^{\sigma}+\mathrm{FPA}^{\sigma}+\mathrm{FNA}^{\sigma}}$
22:  **end for**
23:  **return** CHOTA $\leftarrow \sqrt{\frac{\mathcal{A}^{\sigma}}{|\mathrm{TP}|+|\mathrm{FP}|+|\mathrm{FN}|}}$

---

$$\mathrm{FPA}^{\sigma}(c) = \sum_{i' \in l(i)} \sum_{j' \notin l(j)} M_{i'j'} = \sum_{i' \in l(i)} \sum_{j'=0}^{J+1} M_{i'j'} - |\mathrm{TPA}^{\sigma}(c)| \qquad (24)$$

$$\mathrm{FNA}^{\sigma}(c) = \sum_{j' \in l(j)} \sum_{i' \notin l(i)} M_{i'j'} = \sum_{j' \in l(j)} \sum_{i'=0}^{I+1} M_{i'j'} - |\mathrm{TPA}^{\sigma}(c)| \qquad (25)$$

The value $\mathcal{A}^{\sigma}(c)$ is equal for all $c \in \mathrm{TP}$ that share the same prediction and ground truth ID $\mathrm{PrID}(c) = i$ and $\mathrm{GtID}(c) = j$. Thus, $\mathcal{A}^{\sigma}(c)$ needs to be calculated only once for all TPs that are represented by an element $M_{ij}$. In other words, the number of to calculate values $\mathcal{A}^{\sigma}(c)$ equals the number of non-zero elements $M_{ij} > 0$ with $i > 0$ and $j > 0$. Assuming an extremely sparse $M$ with $n$ non-zero elements per ID on average, the entire calculation is quasi-linear depending on the number of trajectory IDs and $n$. However, since the number of trajectory IDs is usually much smaller than the number of detections, the CHOTA calculation complexity is practically negligible compared to the matching procedure. The complete algorithm is shown in Algorithm 1.
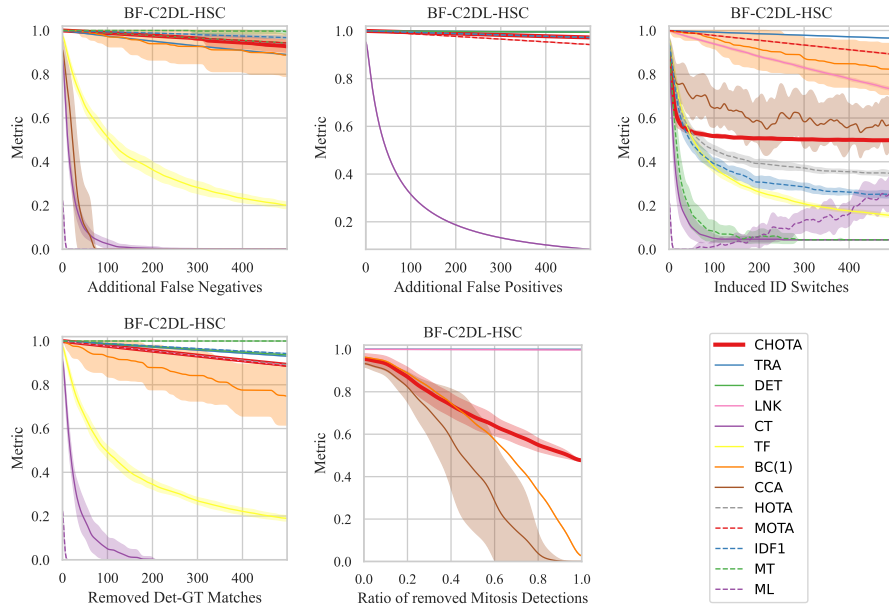
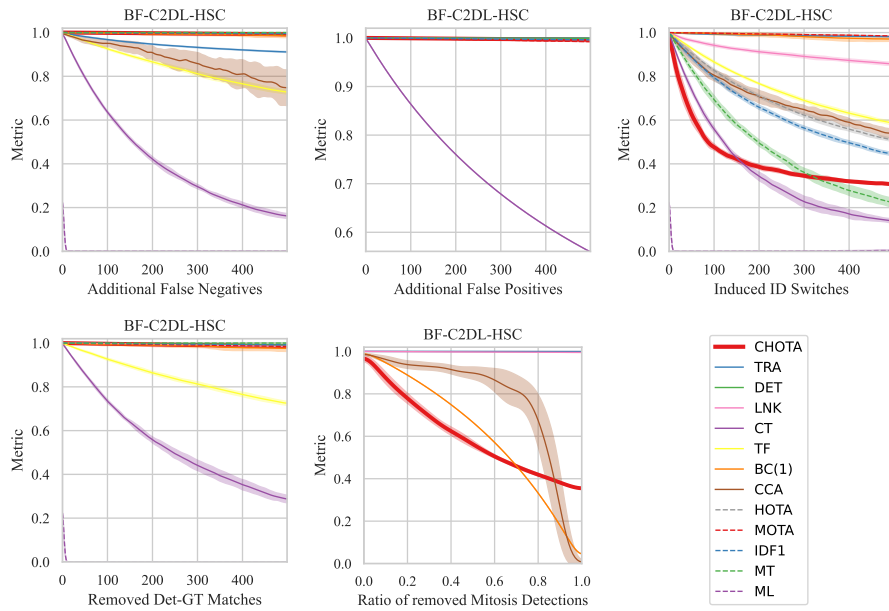**Fig. 5:** Synthetic error induction for Dataset BF-C2DL-HSC (Sequence 01)



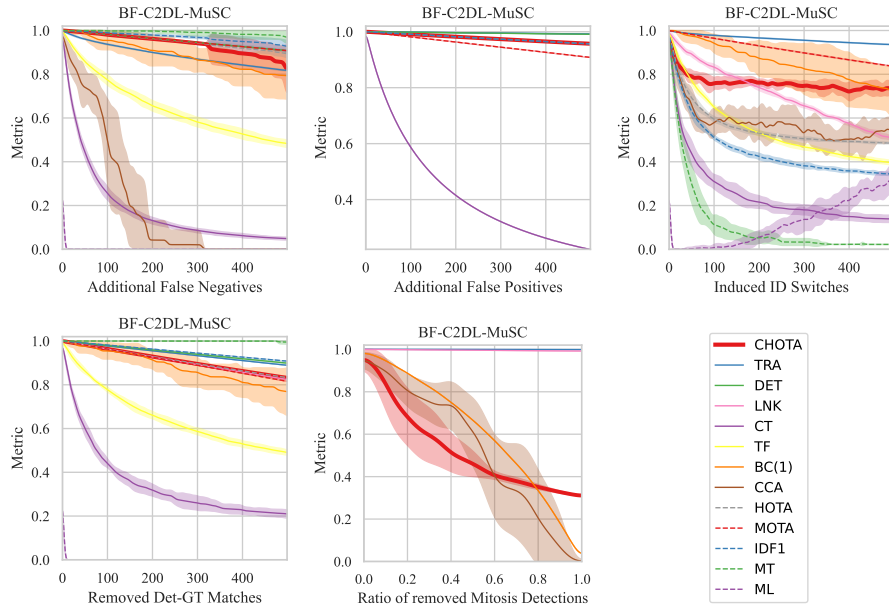**Fig. 6:** Synthetic error induction for Dataset BF-C2DL-HSC (Sequence 02)

**Fig. 7:** Synthetic error induction for Dataset BF-C2DL-MuSC (Sequence 01)
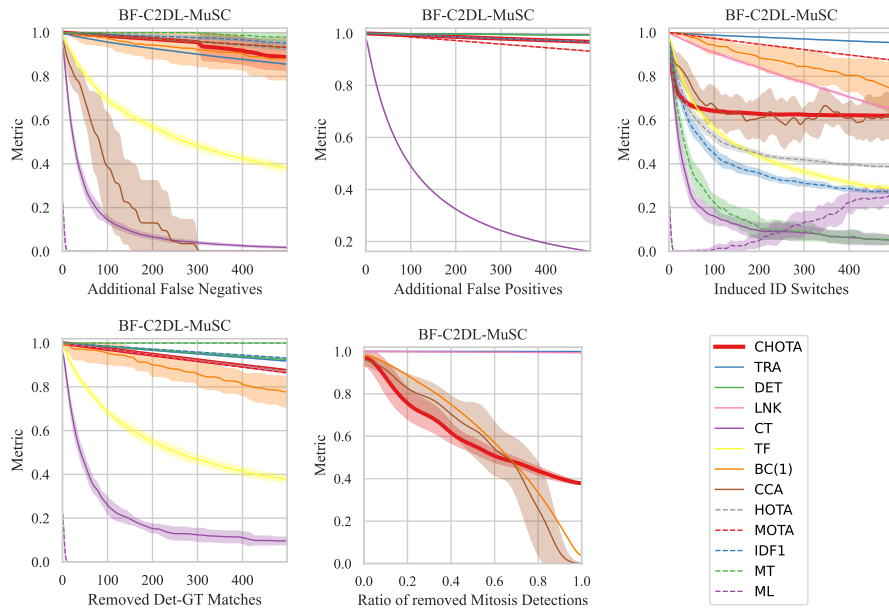


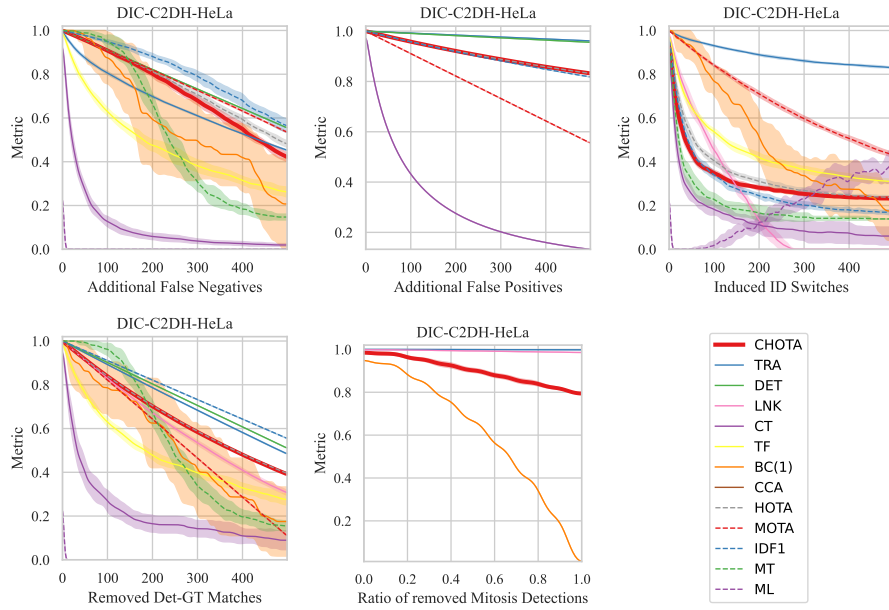**Fig. 8:** Synthetic error induction for Dataset BF-C2DL-MuSC (Sequence 02)

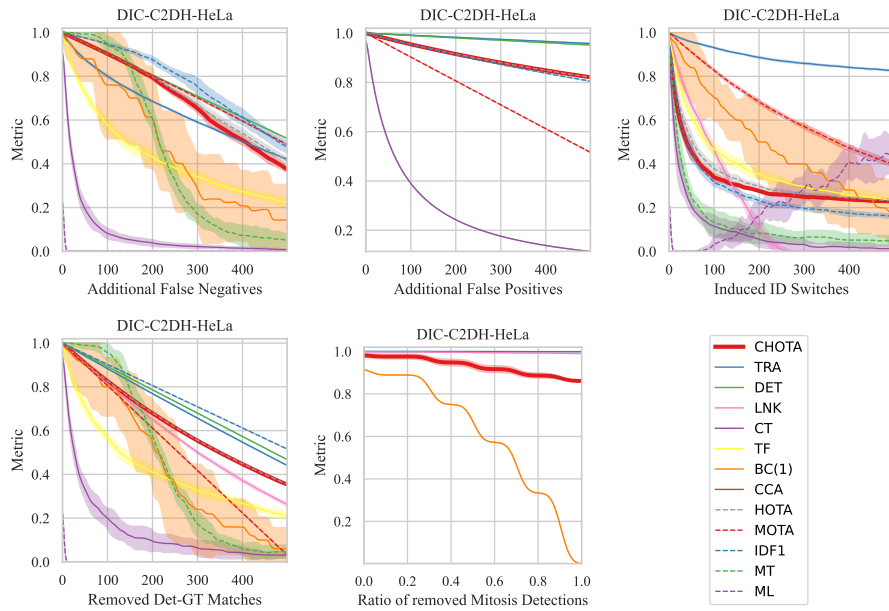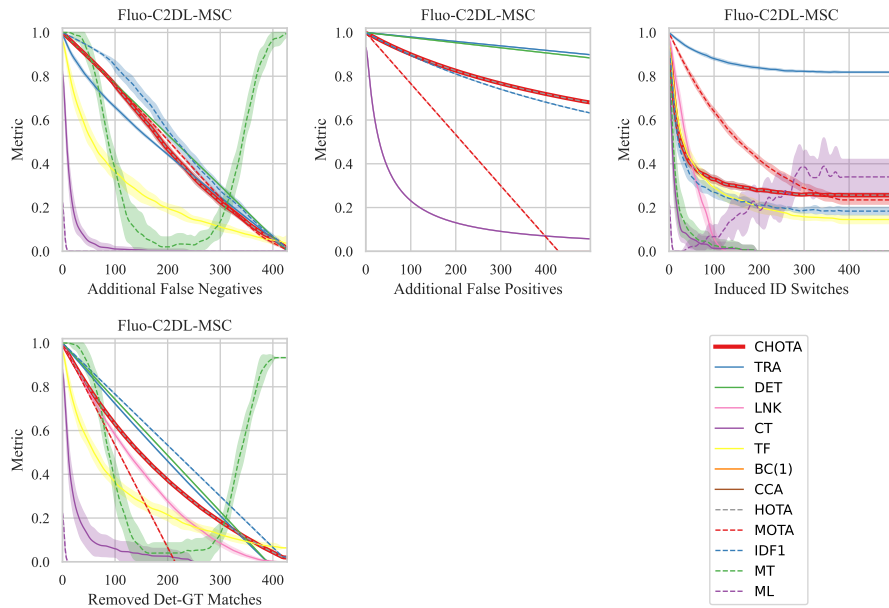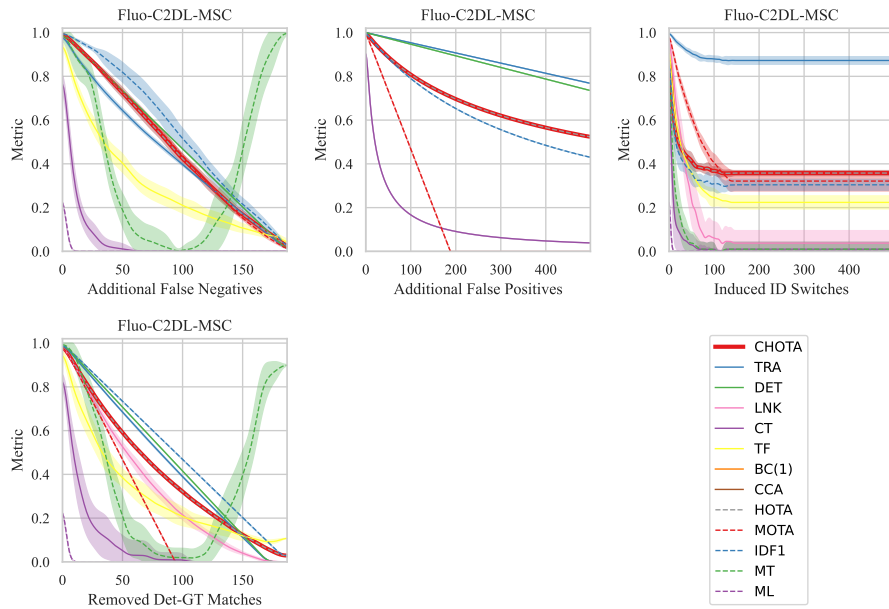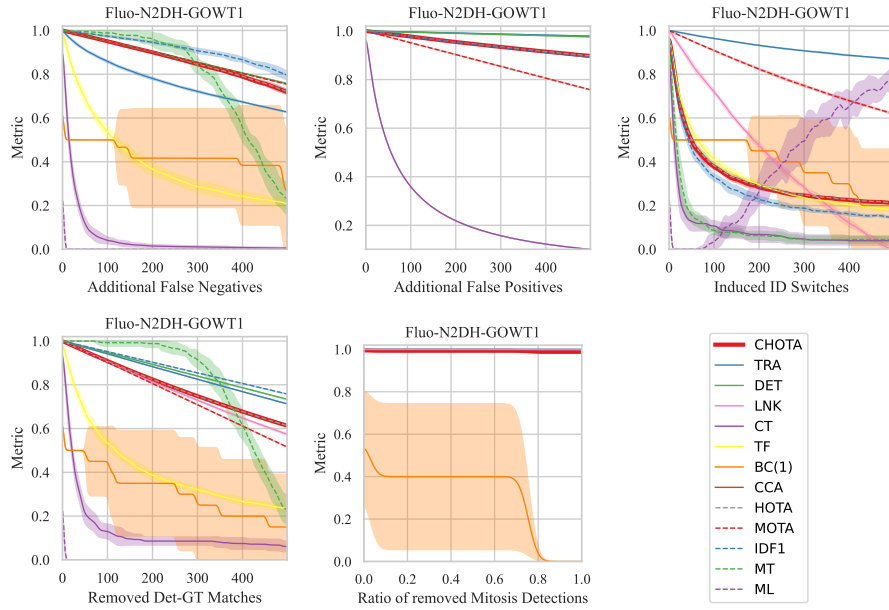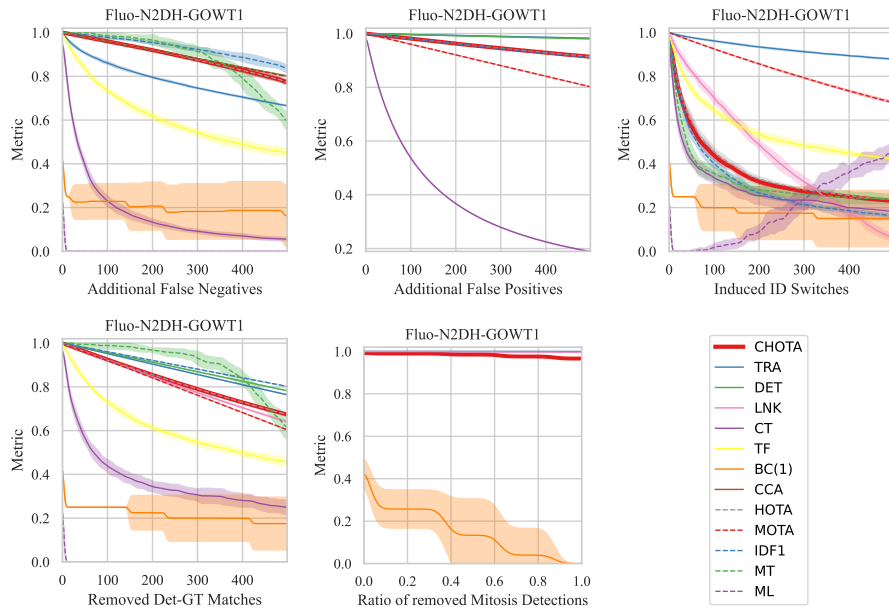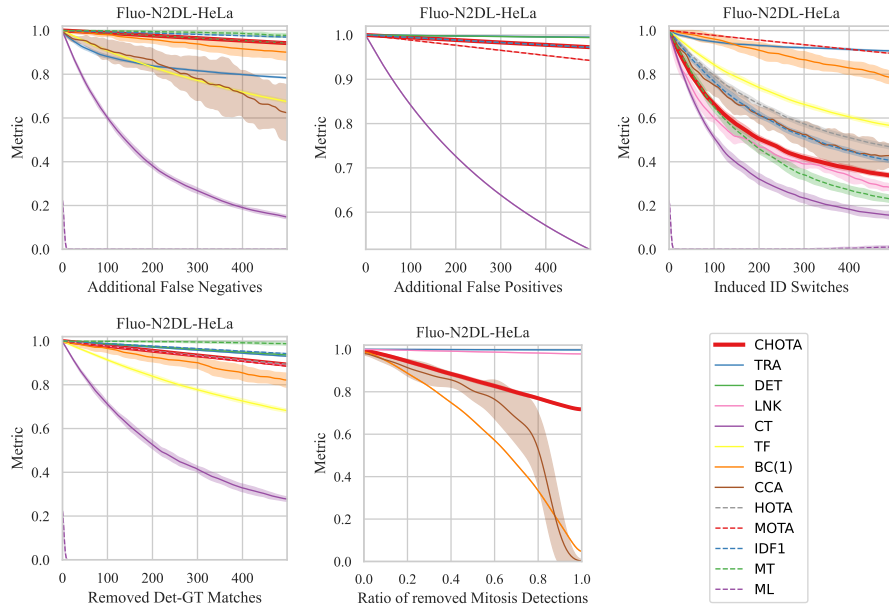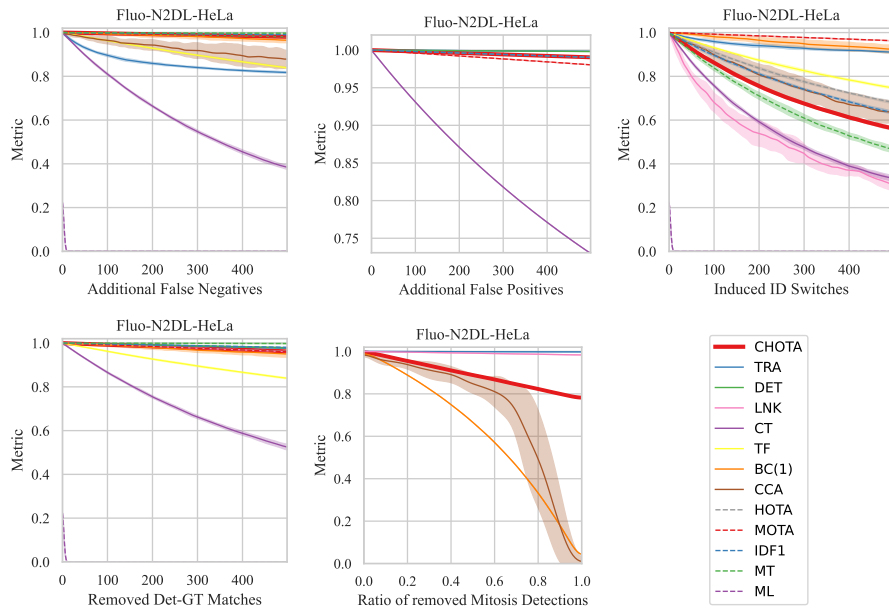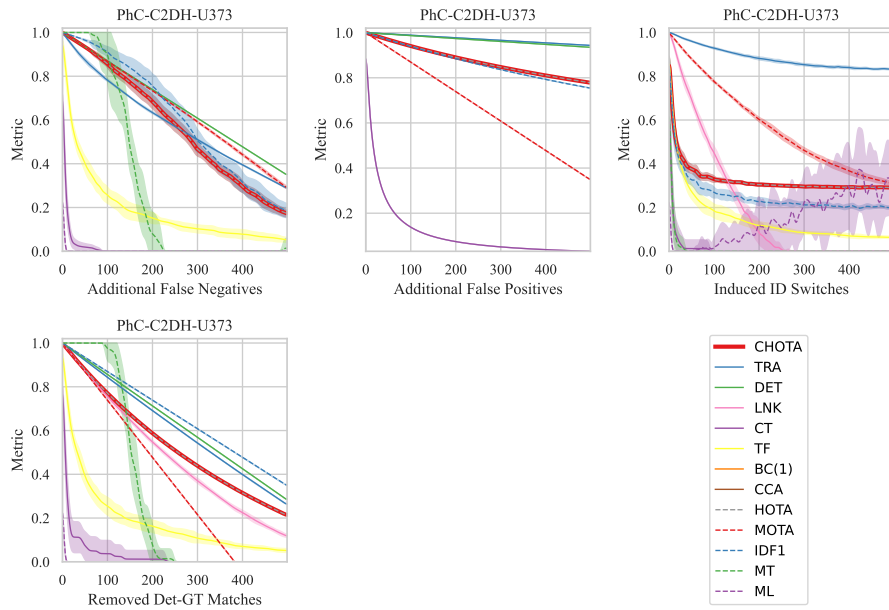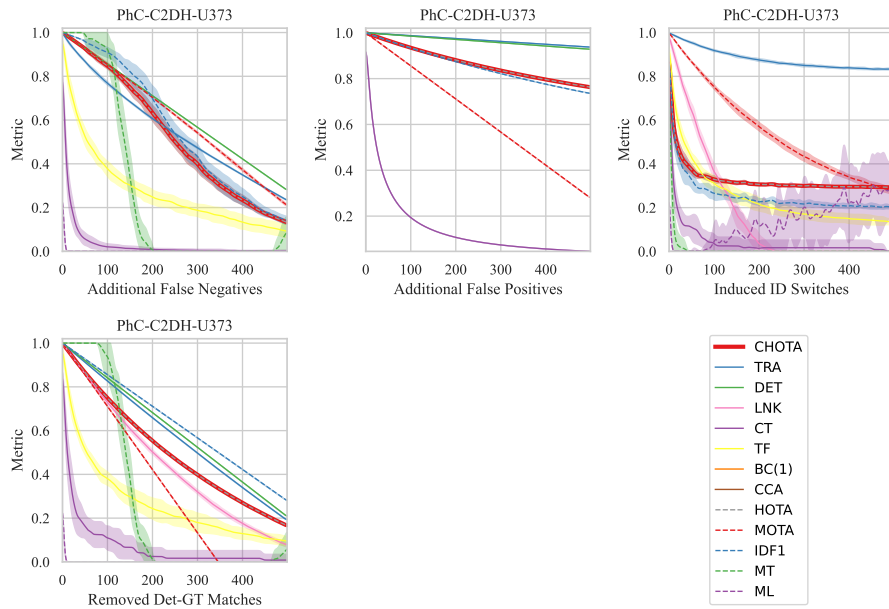**Fig. 9:** Synthetic error induction for Dataset DIC-C2DH-HeLa (Sequence 01)



**Fig. 10:** Synthetic error induction for Dataset DIC-C2DH-HeLa (Sequence 02)
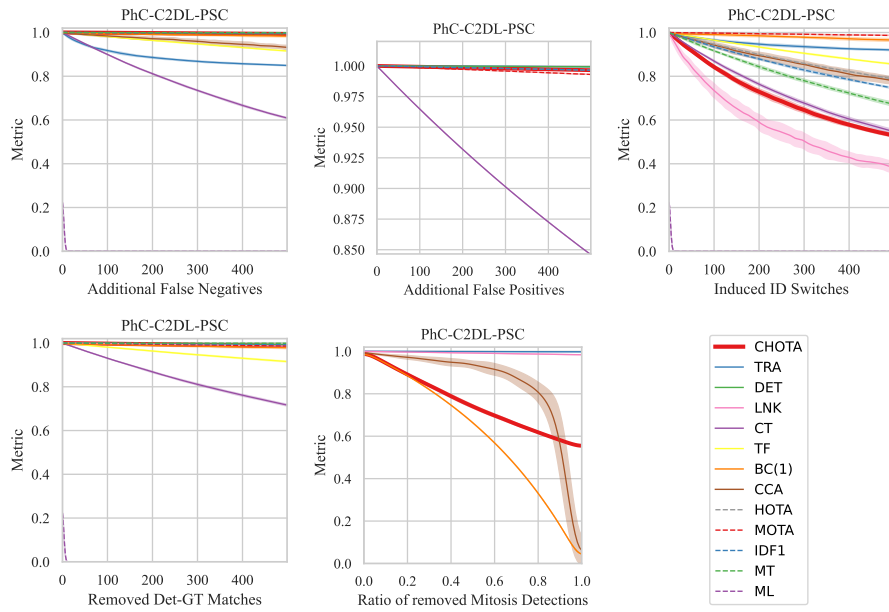
**Fig. 11:** Synthetic error induction for Dataset Fluo-C2DL-MSC (Sequence 01)



**Fig. 12:** Synthetic error induction for Dataset Fluo-C2DL-MSC (Sequence 02)

**Fig. 13:** Synthetic error induction for Dataset Fluo-N2DH-GOWT1 (Sequence 01)



**Fig. 14:** Synthetic error induction for Dataset Fluo-N2DH-GOWT1 (Sequence 02)

**Fig. 15:** Synthetic error induction for Dataset Fluo-N2DL-HeLa (Sequence 01)



**Fig. 16:** Synthetic error induction for Dataset Fluo-N2DL-HeLa (Sequence 02)

**Fig. 17:** Synthetic error induction for Dataset PhC-C2DH-U373 (Sequence 01)



**Fig. 18:** Synthetic error induction for Dataset PhC-C2DH-U373 (Sequence 02)

**Fig. 19:** Synthetic error induction for Dataset PhC-C2DL-PSC (Sequence 01)
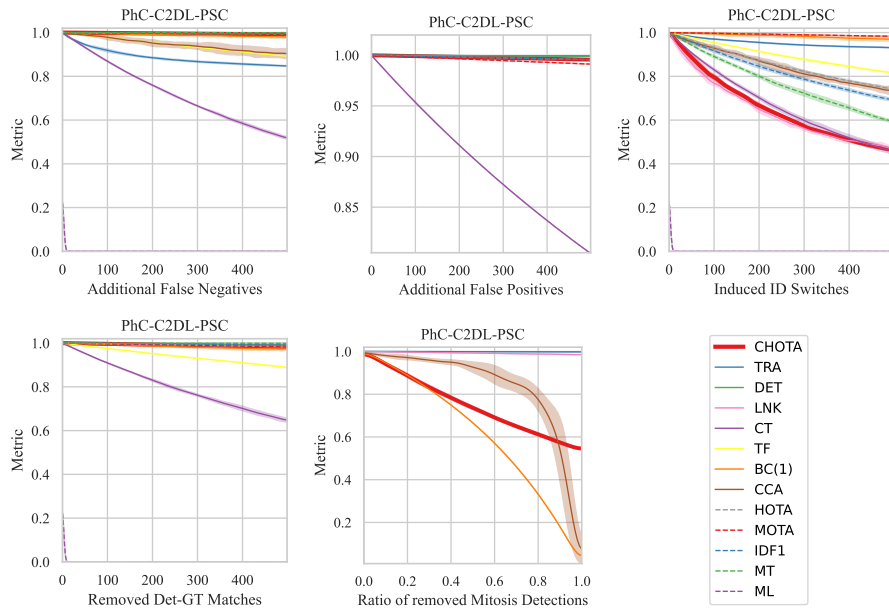


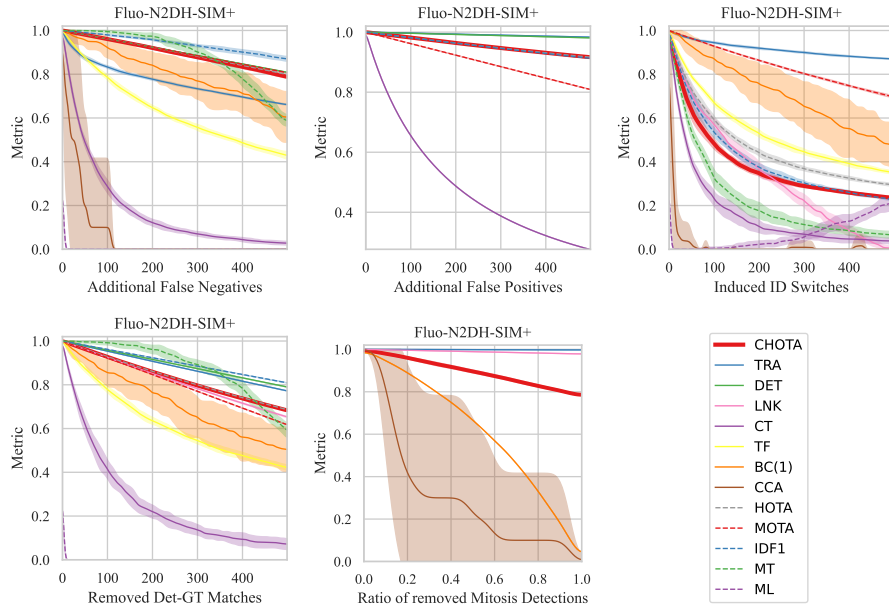**Fig. 20:** Synthetic error induction for Dataset PhC-C2DL-PSC (Sequence 02)

**Fig. 21:** Synthetic error induction for Dataset Fluo-N2DH-SIM+ (Sequence 01)
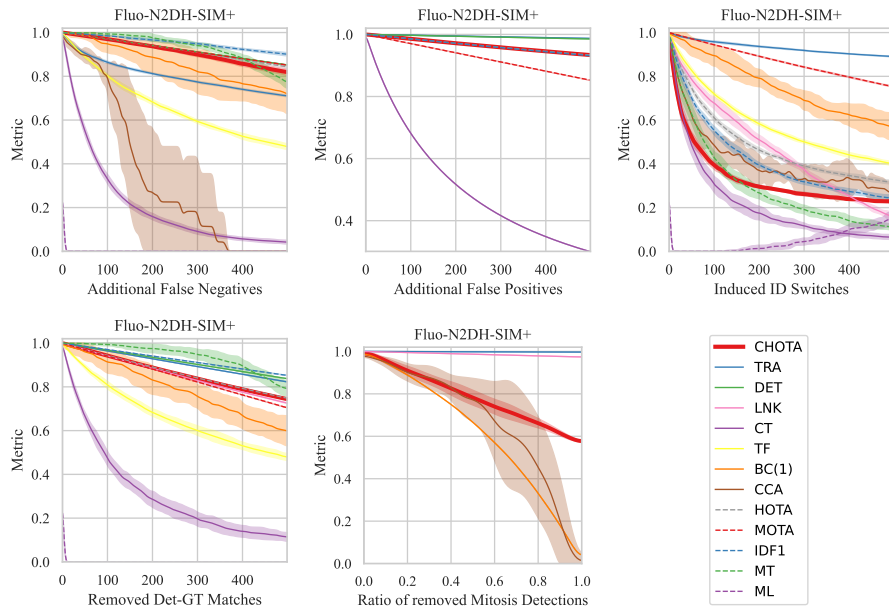


**Fig. 22:** Synthetic error induction for Dataset Fluo-N2DH-SIM+ (Sequence 02)