Benchmarking multi-object tracking for *C. elegans* whole-brain imaging

# Introduction

Recently developed fast volumetric imaging techniques enable researchers to capture neuron activity dynamics of the entire brain in small organisms, such as those of zebrafish, *C. elegans* and Drosophila, at high spatiotemporal resolution1–14. These global neuron activity data with single-cell resolution have enabled previously highly difficult studies. For example, in *C. elegans*, global neuron activity data have been used to study locomotion15–20, chemosensation21, mating22, sleep23,24, aging25, network properties26,27 etc. Combined with developments in high-throughput instrumentation for automation in microfluidics28–34, neuron activity videos under a variety of stimulation conditions can now be acquired rapidly. However, the rate at which data can be processed is still much slower than the rate of data collection.

To extract high signal-to-noise ratio (SNR) neuron activity traces from the videos, a critical step is to track nuclei across frames. However, tracking nuclei in *C. elegans* whole-brain videos is extremely challenging due to several reasons. First, fluorescent intensities of neurons vary considerably within each frame and across frames due to inherent biological variability (stochasticity in fluorophore expression). Additionally, there are technical limitations introduced by imaging such as low SNR, photobleaching of fluorophores, motion artifacts etc. Thus, the number of neurons detected in each frame of video are usually not consistent due to errors made by cell detection methods which are usually used upstream of tracking. As a result, establishing correspondence between inconsistent number of nuclei across frames is extremely challenging. Second, nuclei in *C. elegans* head ganglion are densely packed; thus correspondence estimation algorithm must be able to discriminate features of nearby cells. Finally, *C. elegans* head ganglion can undergo huge non-rigid deformations in the head making it difficult to generate correspondences between nuclei in two image volumes. In the absence of automated methods, neuron activity traces are obtained by highly manually supervised segmentation and tracking. A whole-brain video may consist of thousands of volumes of images, and each volume may contain hundreds of neurons. Thus, manual supervision in segmenting and tracking each neuron across volumes is extremely labor-intensive and may take several weeks. As a result, small sample size (typically 5-10) of individual worm datasets are analyzed in previous works, limiting the statistical power of several downstream analyses, and limiting the scope of whole-brain dynamics studies. Thus, there is a need for fast, automated, reliable, and easily-to-use analytical tools.

Custom pipelines have been developed in different model organisms such as mice35–37, Drosophila38–41, hydra42, and zebrafish1,2,43, which take advantage of properties of the data in their respective systems. For example, in zebrafish, instead of tracking individual region of interest (ROI), image stacks are registered to a common reference frame and same ROI are used across frames1,13,44,45. Similarly, in Drosophila, ROIs are defined using either PCA/ICA based grouping of voxels40 or by registering whole-brain stack to an anatomical atlas39,46. Again, ROIs are tracked across frames automatically due to registration of all frames to a common atlas. Different from these systems, *C. elegans* whole-brain videos present unique challenges . For instance, in comparison to rigid brains of zebrafish and Drosophila, head deformation in *C. elegans* recordings is considerably non-rigid. Further, activities of nearby cells can be correlated, which makes them difficult to distinguish from each other. Additionally, for *C. elegans* whole-brain imaging, single-cell resolution neuron activity traces are desired. Thus, coarse ROI as those determined in Drosophila by using an anatomical reference would not suffice. In addition, a 3D anatomical atlas for registration is not available. Finally, due to complications in registration, ROIs across frames are not automatically tracked. Thus, methods that specifically cater to the data properties and requirements of *C. elegans* are needed.

For tracking cells in whole-brain images, several methods have been developed47–56. These methods differ in terms of many aspects; 1) the objective (loss) functions that are optimized for estimating correspondence between neurons, 2) track linking strategies that are used to link tracks across frames, 3) noise characteristics of datasets on which the method is tested, and 4) accuracy metrics reported. For example. to estimate correspondence between cells mainly four types of objective functions have been used –

1. Registration – these methods either register point-cloud49,51 of neurons detected in two frames to estimate correspondence or directly register images55,56
2. Quadratic53 – these methods53 estimate correspondence by matching edges between neurons
3. Hybrid – these methods47,48,50 combine registration objective and quadratic edge matching objective
4. Similarity Prediction – these methods51,52 predict similarity between neurons across two frames to estimate correspondence.
5. Identity Prediction – these methods54 predict identity of neurons across all frames thus automatically establishing correspondence.

Similarly, to link tracks temporally across frames mainly four types of strategies have been

used –

1. Sequential – in these methods47,48,51 tracks are linked sequentially across frames
2. All-to-one – in these methods50,52,53,56 a set of frames are linked to a reference frame and final tracking is performed by linking reference frames
3. All-to-all – in these methods49,55 a set of frames are linked to many reference frames and tracking is performed by clustering the results of such mapping
4. Identity Prediction – in these methods54 there is no need to link tracks as identity prediction automatically link tracks.

While high accuracy is demonstrated for specific datasets in these works, it is not established which methods perform best in general and under what conditions. This is due to inconsistent accuracy metrics reported across works, which focus on different aspects of tracking accuracy thus making it difficult to assess the methods in terms of usefulness. Finally, tracking algorithms may have to operate on datasets across varying noise levels. The noise could arise either due to errors made by cell detection methods that are run upstream of tracking or noise in images. However, it is not clear if the methods are generalizable and robust to noises in data as it is extremely cumbersome to obtain ground-truth annotated benchmarking datasets across a variety of conditions.

To address the challenges in building optimized tracking algorithms, we developed a set of tools that enable to easily mix and match the aforementioned design choices, test them on ground-truth annotated datasets, and quantify accuracy across standardized metrics, thus allowing to compare different design choices (Figure 1). Key features of our toolbox include 1) easy generation of synthetic data, thus known ground truth tracking annotations, with varied levels of common types of noises in data, 2) a framework to benchmark over 20 correspondence estimation across Registration, Quadratic and Hybrid categories of methods 3) implementation of two commonly used track linking strategies; Sequential, and All-to-one. Apart from the strategies used in whole-brain cell tracking works, we also compare against min-cost flow optimization based tracking methods used in the general cell-tracking literature, and 4) implementation of 6 standardized multi-object tracking (MOT) accuracy metrics thus removing bias due to reporting metrics. This set of tools are easy-to-use, callable implementations of various methods and tracking strategies; they enable fast screening and benchmarking of new methods. Using the toolbox, we conducted an unbiased screen of design choices across all accuracy metrics to identify the ones that perform better and are more robust to noise for tracking neurons in *C. elegans* whole-brain videos. We further conduct a similar screen on a manually annotated whole-brain video recording and .

# Methods

## Generating synthetic data for tracking

Datasets that have ground-truth annotations available are critical for development of tracking algorithms, but they are extremely laborious to generate, for example tracking 100 neurons across just 100 frames requiring would require 100,000 manual annotations. Because of the scarcity of fully tracked and validated datasets across a variety of conditions, we took the approach of generating synthetic data. This approach allows us to generate arbitrarily long video data sets with known ground-truth tracks, and to dial in noise levels at will thus testing the robustness of methods to noise. Synthetic data was generated using the freely available 3D atlas at OpenWorm57. For each synthetic video, a random subset of ~130 (typical number of cells observed in whole-brain image stack) cells were selected from the atlas, and these group of ground-truth cells were used to generate synthetic video. Most previous tracking methods detect cells in images first using a segmentation algorithm which are then subsequently tracked. However even the most advanced cell detection methods have false positive detections and missed detections that are defined by the precision and recall metrics of the method. Additionally, inherent noise in the images such as etc. lead to missed neuron detections as well To simulate these noises, three types of noises were added to generate synthetic video. These included 1) adding false positive cells, 2) removing ground-truth cells, and 3) perturbing positions of cells in video frames. To mimic these noises, a precision and recall value was set for each synthetic video varying between 0.6 to 1. Precision value (defined as fraction of total nuclei detections that are true positive) defines the number of false positive nuclei to be added to each frame. Recall value (defined as fraction of total ground-truth nuclei that are detected) defines the number of nuclei removed from each frame. False positive nuclei were added randomly and uniformly distributed throughout the whole-brain volume. Nuclei to be removed, i.e. false negative nuclei, were also randomly and uniformly selected among the ground-truth nuclei. Finally, to mimic noise in cell positions predicted by nuclei detection algorithms used upstream, random Gaussian perturbation was applied to cell positions in each frame. This noise was sampled from a normal distribution with zero mean and fixed variance . Here and denote variances along , and image dimensions and denotes diagonalizing vector . Hence, the position of cell in synthetic data was defined as .

### List of correspondence estimation methods available in toolbox and screened

A critical choice that determines the accuracy of cell-trackers is the underlying correspondence estimation method that is used to determine which cells correspond across frames. Our frameworks includes correspondence estimation methods across 3 categories of loss functions (Table 1):

1. Point-cloud registrations based methods where the linking cost is constrained linear as these methods link cells in one frame to cells in another frame up to a smooth deformation or other specified constraint. We compiled and standardized publicly available of implementations of 6 point-cloud registration methods. These methods differ in terms of the constraints that are included in the cost-function along with the linear cost. Further we include Hungarian (Munkres) algorithm.
2. 2) Graph-matching methods that link a pair of cells (i.e. edges) in one frame to a pair of cells (edges) in another frame, hence the linking cost is quadratic. We compiled and standardized 13 graph-matching methods. These methods differ in terms of the underlying optimization strategy used to optimize the quadratic cost-function.
3. We develop a new hybrid method that matches both cells and edges across frames to link cells thus the cost is both linear and quadratic. The method is similar to the Fused Gromov Wasserstein metric58,59. Entropic regularization term in the objective functions helps in efficient optimization using the Sinkhorn iterations60–62.

Here denotes the linear cost of matching cell in first frame to cell in second frame, denotes the quadratic cost of matching pair of cells in first frame to pair of cells in second frame, denotes the correspondence between cells in first frame and in second frame, denotes the entropic regularization term and is regularization hyperparameter.

Table 1. List of correspondence estimation methods available in toolbox

|  |  |  |
| --- | --- | --- |
| Method | Type | Reference |
| L2QPMAP | Quadratic | 63 |
| IPFPMAP | Quadratic | 64 |
| CRF | Quadratic | 65,66 |
| GW\_EOT | Hybrid | This work |
| SMAC | Quadratic | 67 |

|  |  |  |
| --- | --- | --- |
| PHM | Quadratic | 68 |
| PSM | Quadratic | 68 |
| IPFP | Quadratic | 64,69 |
| IPFPgm | Quadratic | 64,69 |
| SMIPFP | Quadratic | 64,69 |
| IPFPU | Quadratic | 64,68 |
| IPFPS | Quadratic | 64,68 |
| SM | Quadratic | 69 |
| RRWM | Quadratic | 68,70 |
| GLTP | Linear | 71 |
| CPD | Linear | 72 |
| TPSRPM | Linear | 73 |
| ECMPR | Linear | 74 |
| GMMReg | Linear | 75 |
| GLMD | Linear | 76 |
| PRGLS | Linear | 77 |

|  |  |  |
| --- | --- | --- |
| L2ERPM | Linear | 78 |
| Munkres | Linear | 79 |

## Tracking nuclei in whole-brain video of freely moving animal.

One other important challenge is that non-rigid deformations in the head make it difficult to track nuclei. Thus, behaviour recording channel was used to detect skeleton of the worm. Next the skeleton was mapped to neuron activity recording channel. To define coordinates of nuclei in each frame, a coordinate system was defined based on the centerline in each frame. X coordinate for each nuclei was defined as the distance along the centerline starting from posterior end of the centerline. Y coordinate for each cell was defined as the perpendicular distance of the cell to the centerline.

## Accuracy quantification on experimental whole-brain imaging video

Ground tracks were annotated manually for 200 frames in the video. These tracks were used to assess the accuracy of various methods.

# Results

## Toolbox for comparison tracking methods

Several methods have been developed for tracking nuclei in 3D *C. elegans* whole-brain recording datasets 42,47,51,80 showing great accuracy. However, the decision of which method performs best or may be most suitable to researchers is difficult to make due to several reasons. First, most methods report accuracy using different custom defined metrics. The absolute numbers in these metrics are not directly comparable across methods from different research groups and may even inflate performance across some methods. Second, few of these methods characterize the robustness against noises that are common in data. These noises include deviations in positions of cells frame to frame, false positive detections and missing detections generated by upstream nuclei detection methods, etc. Third, many of these methods provide accuracy on datasets collected in their lab; thus, whether accuracy is generalizable across datasets is not known. Finally, there is scarcity of ground truth tracked and labelled whole-brain video data. Such ground-truth tracking data are necessary to optimize performance of tracking methods and standardize accuracy comparisons across methods. In the absence of such data, it is difficult for researchers to build and optimize new tracking methods.

To address these issues, we developed a toolbox that enables researcher to develop, optimize and compare their new tracking methods against previous methods. There are three modular components in our toolbox. In the first component, researchers can generate synthetic video data of segmented nuclei where ground truth tracking of all cells is known by default. To recapitulate the challenges faced by tracking algorithms in real data, three kinds of noises can be added to the synthetic video; false positive detected cells, missed detections and deviations in cell positions. These noises commonly arise in experimental data due to errors in segmentation method used, transients in neuron activities, and low fluorophore expression levels, and affect tracking algorithm accuracy. Since the ground-truth tracking is known by default for synthetic data, researchers can easily optimize and compare their tracking methods across a range of noise levels. Thus, the first component addresses the issue of data scarcity and generalizability.

In the second component, researchers can choose from 7 different point-cloud matching based methods or linear methods and 14 different graph matching methods (Table 4.1) for establishing correspondence between cells across frames. Further two different tracking strategies can be used across these methods. These include 1) sequential tracking where each frame is tracked to its previous frame, and 2) tracking by matching all frames to one reference frame. We also compare accuracy for min-cost flow based tracking strategies such as uTrack81 and CINDA82 Additionally other correspondence estimation strategies developed by researchers such as deep learning based etc. can be included easily in the toolbox to enable researchers to compare their methods against baselines. Thus the second component addresses the issue of enabling researchers to generate baseline comparisons easily.

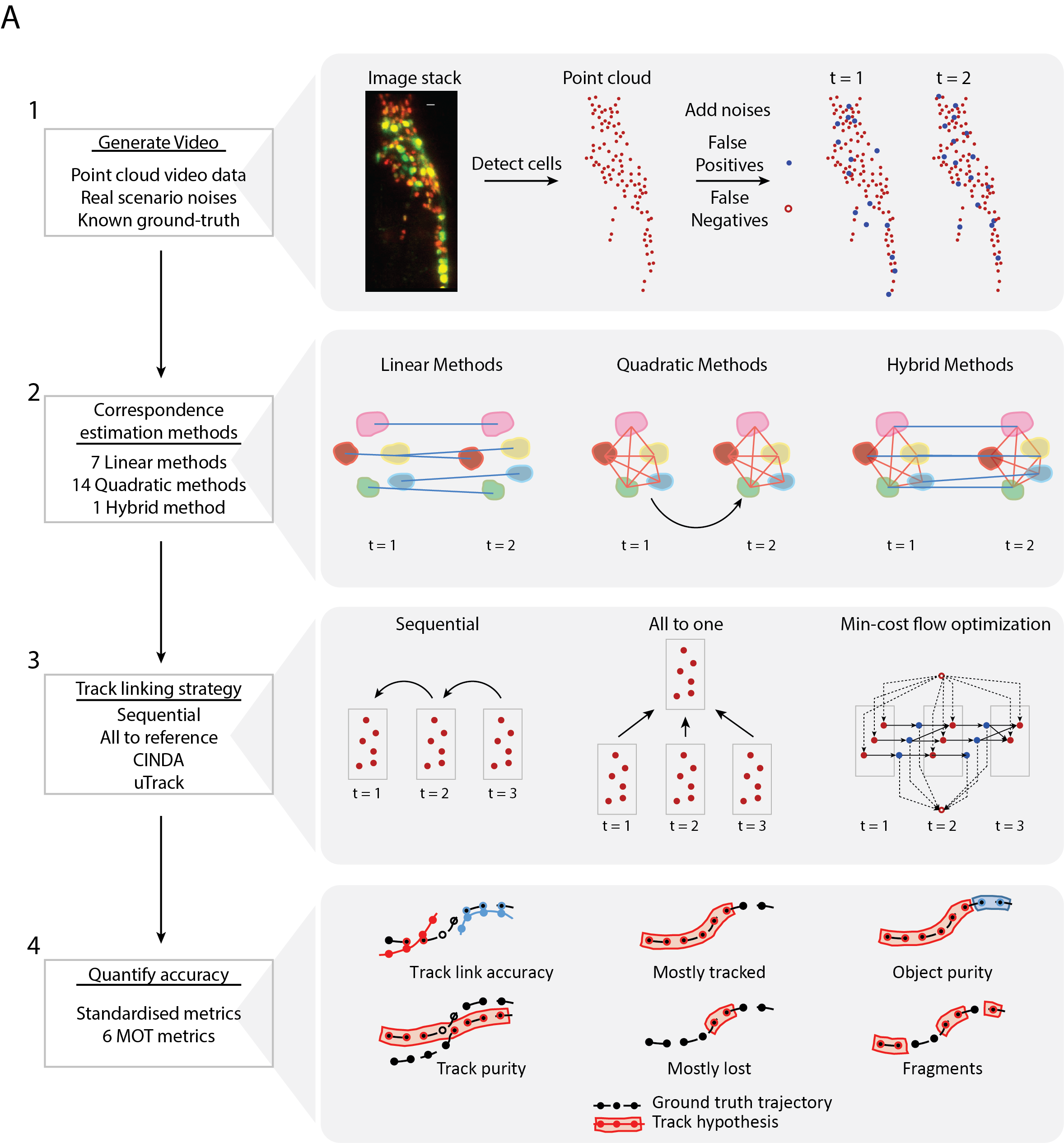


Figure 1. Overview of multi-object tracking toolbox. A) Steps performed in the toolbox to analyze and compare tracking methods. 1) Synthetic data with noises common in experimental data can be generated using OpenWorm atlas or experimental recordings. 2) Toolbox provides implementations of 13 Quadratic, 7 Linear (point-registration), and 1 Hybrid method for correspondence estimation between nuclei in different frames. 3) Two different track linking strategies can be used by correspondence estimation method selected previously. These include – ‘Sequential’ and ‘All to One strategies. Additionally, min-cost flow optimization based tracking strategies such as CINDA, uTrack can be compared. 4) 6 standardized accuracy metrics commonly used in MOT challenges are output based on ground-truth trackings.

Finally, in the third component, six different accuracy metrics are output by the toolbox. These include Object purity, Track Purity, Track Link Accuracy, Fragments per ground-truth track, Mostly Lost, Mostly Tracked (Table 4.2, Figure 4.1A, 4.2A). The accuracy metrics are standardized and commonly used in multi-object tracking literature83–86 as they assess different aspects of tracking that may be critical for different applications. Examples in Figure 4.2B illustrate that examining a tracking method with just one metric can be deceptive. For example, a small decrease in track-link accuracy may lead to dramatic decrease in Track Purity and Object Purity (Figure 4.2B top panel). In contrast in another example, a dramatic decrease in track link accuracy causes no change in object purity and track purity. Thus, a multidimensional comparison across methods using all accuracy metrics is essential to truly grasp the performance of a tracking method. A systematic comparison across methods to screen methods that perform highly across all metrics may reveal robust and generalizable trackers. Thus, this component addresses the issue of standardizing the comparison by providing a common footing to compare various tracking methods.

Table 4.2. Definition of MOT accuracy metrics used to compare methods

|  |  |  |
| --- | --- | --- |
| **Accuracy metric** | **Description** | **Reference** |
| Object purity | Each ground truth object may be tracked by multiple track hypothesis. Thus, a track hypothesis is assigned to ground truth object based on majority voting i.e. the track hypothesis that tracks the ground truth object for maximum number of frames. | 83 |

|  |  |  |
| --- | --- | --- |
|  | Subsequently object purity for a ground truth object is defined as the fraction of timepoints that the ground truth object is tracked by the matched track hypothesis |  |
| Track Purity | Each track hypothesis may track multiple ground truth objects at different time points. Thus, a ground truth object is assigned to track hypothesis based on majority voting i.e. the ground truth object tracked by the track hypothesis for maximum number of frames. Subsequently track purity for a track hypothesis is defined as the fraction of timepoints that the assigned ground truth object is tracked by the track hypothesis | 83 |
| Track Link Accuracy | For each track hypothesis, it is defined as the fraction of frames where the track hypothesis tracks the same ground-truth object in the next frame. | 85 |
| Fragments per ground truth track | Number of track hypothesis fragments for each ground truth track. A track hypothesis is fragmented if it does not track the ground-truth object continuously. | 84 |
| Mostly Lost | For each ground-truth object, if the fraction of frames correctly tracked (i.e. same track hypothesis tracks current and next frame) are lower than 20% then the ground-truth object is marked as mostly lost. | 84 |

|  |  |  |
| --- | --- | --- |
| Mostly Tracked | For each ground-truth object, if the fraction of frames correctly tracked (i.e. same track hypothesis tracks current and next frame) are higher than 80% then the ground-truth object is marked as mostly tracked. | 84 |

Along with these accuracy metrices, we defined an additional metric for multi-dimensional comparison of correspondence estimation methods called Robustness Score. This metric measures how robust the methods are, in terms of Object Purity accuracy, when the noise level in data is increased (Precision and Recall value is decreased from 0.9 to 0.6)



Figure 2. Standardized accuracy metrics and some examples of non-linear relationship between metrics. A) Examples depict calculation of accuracy metrics – Track link accuracy, Object purity, Track Purity, Fragments per ground truth, Mostly lost, Mostly tracked. B) Examples depicting judging accuracy with one tracking method can be deceptive. Top left panel – all ground-truth objects are tracked accurately with track hypothesis thus achieving perfect Track link accuracy, Object purity and Track Purity. Top right panel – With incorrect tracking, track linking accuracy is decreased by only 12% but object purity and track purity decrease by ~45%. Object purity and Track purity are calculated by averaging across ground-truth objects and track hypothesis respectively.

## Identifying optimal strategies for nuclei tracking using synthetic data

Using our toolbox we sought to determine, which combinations of correspondence estimation methods and tracking linking strategies perform best. Since for whole-brain imaging application we are most interested in quality of neuron activity traces extracted from videos, we chose object purity metric for comparisons because this metric measures the purity of each track. Thus, it is correlated with the purity of calcium signals that we can expect to get from tracked videos.

First, we compared accuracy of all methods across three different tracking strategies. These include 1) sequential matching of each frame to previous frame, 2) matching all frames to a randomly selected frame from the video denoted as ‘all\_to\_random’, and 3) matching all frames to an atlas frame denoted as ‘all\_to\_atlas’. Here perfect atlas frame is a frame that does not contain any false positive segmentation and missed detections of cells. Such an atlas frame is available to us for synthetic data case however it may not be available for real experimental datasets. Comparison across these strategies show that sequential matching achieves much lower accuracy compared to ‘all\_to\_atlas’ and ‘all\_to\_random’ strategies (Figure 4.3A). This is because, as mentioned in previous studies, tracking errors can accumulate in sequential tracking. Further, ‘all\_to\_random’ strategy performs worse than ‘all\_to\_atlas’ strategy (Figure 4.3A). This is because a randomly selected frame from the video contains falsely detected cells and missing cells which are possibly different from the falsely detected or missing cells in other frames. Thus, estimating correspondence between cells in frame to be tracked and random reference frame is inaccurate. In contrast, such errors do not arise when an atlas frame is used as reference. We note, that the modular nature of the toolbox allows to explore other track linking strategies as they can be easily included in the toolbox as well. E.g. we tested two additional strategies uTrack81 and CINDA82 (Figure 4.3B). However, these strategies performed much worse compared to other methods. In conclusion, tracking using atlas frame is advantageous and a methodology to build such an atlas frame using imperfect information in individual video frames will be useful.

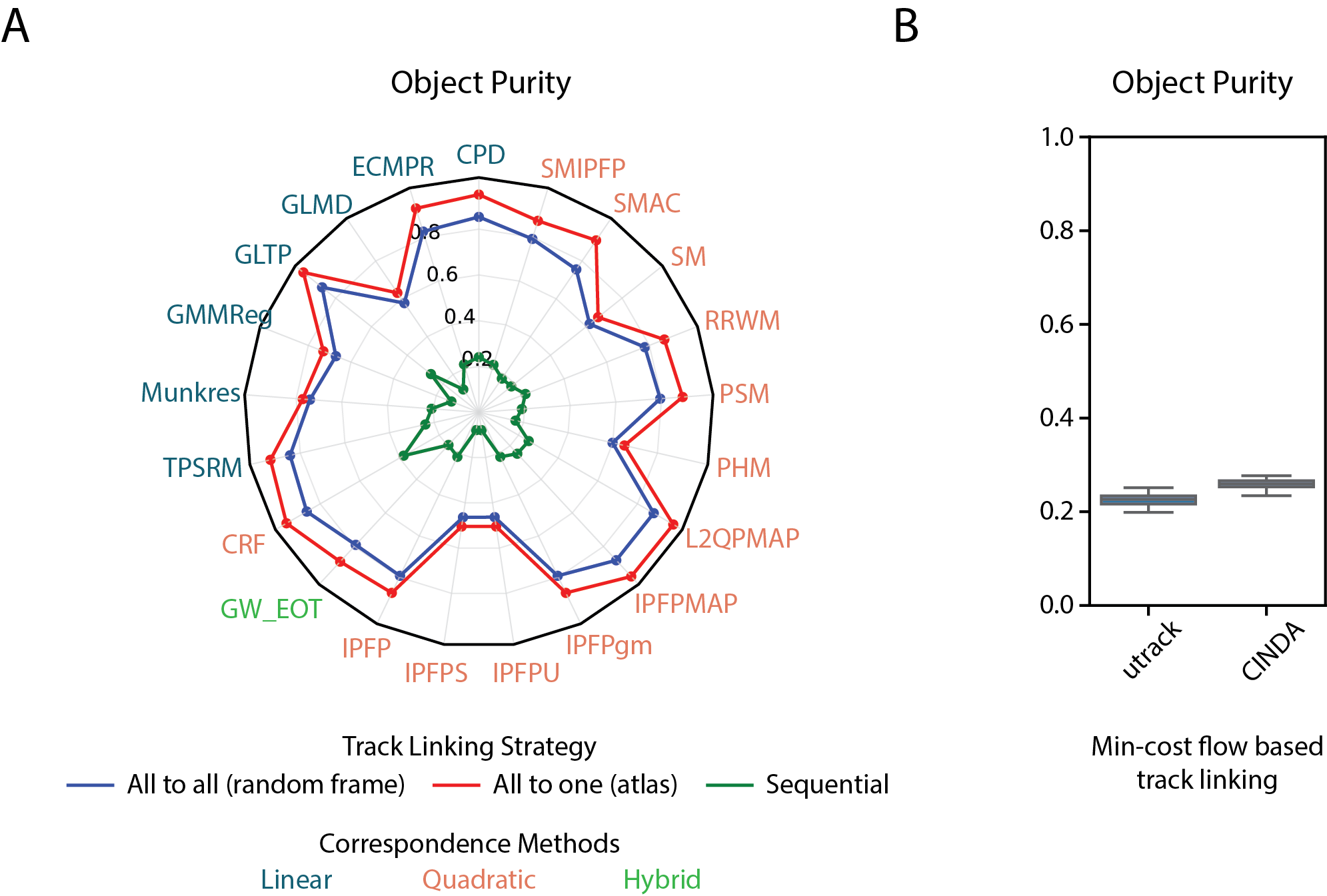


Figure 3. Comparison of object purity across track linking strategies paired with various correspondence estimation methods on synthetic data. Noise in the video (precision and recall values were kept fixed at 0.9). Sequential track lining strategy performs worst and ‘All to reference’ track linking strategy performs best. Min-cost optimization based strategies also achieve low accuracy.

A tracking method that achieves high object purity while having a small runtime would be ideal for tracking nuclei in long whole-brain imaging datasets. Thus next, we compared the runtime and object purity achieved by all 21 methods (Figure 4.4A) while keeping the precision and recall based noise levels fixed at 0.9 in the video. Further we used ‘all\_to\_atlas’ tracking strategy for each run. Simulation results revealed that several methods such as L2QPMAP, CRF, GLTP, CPD achieved object purity accuracy greater than 90%. Among these methods, graph matching methods L2QPMAP, IPFPMAP, and CRF. achieved higher accuracy compared to registration based methods. However, the runtime of graph matching methods is orders of magnitude higher than registration based methods such as CPD, GLTP and ECMPR. Thus, researchers should choose methods considering trade-offs between runtime and accuracy. Along with runtime, an important consideration while choosing the best tracking method for whole-brain videos is robustness to common types of noise in data such as false positive cell detections and missed detections. This is because the noise level in whole-brain video data may vary significantly based on imaging conditions used such as exposure time, laser power, axial resolution etc. Choice of imaging conditions significantly affect the signal-to-noise ratio in images thus affecting the performance of cell segmentation algorithms used to detect cells before tracking. Thus next, we compared runtime and object purity across all methods for high noise conditions that is while keeping the precision and recall values fixed to 0.6 (Figure 4.4B). In this case, accuracy of all methods decreased. In particular registration based methods CPD, ECMPR, TPSRPM, and graph matching methods SMAC, PSM, IPFP that achieved ~90% object purity for low noise condition (Figure 4.4A) diminished in accuracy. Remarkably, some methods such as GLTP, IPFPMAP, L2QPMAP and CRF were robust to high noise levels, still achieving high accuracy.

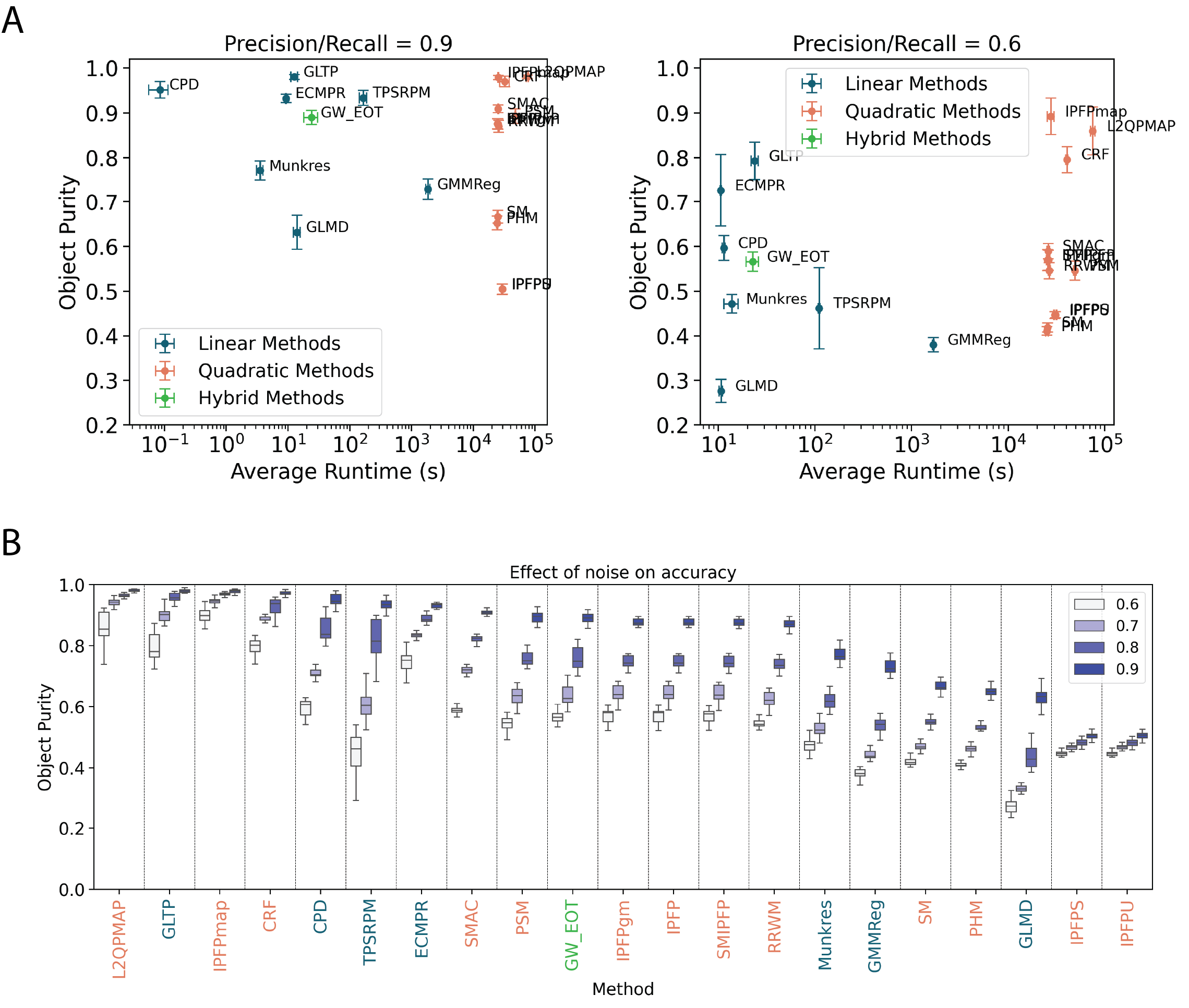


Figure 4. Comparison of runtime and object purity across correspondence estimations methods. A) Runtime and obj purity of graph-matching and linear methods for two noise levels in synthetic data. Left panel - Low noise level (precision and recall set to 0.9), right panel – high noise level (precision and recall set to 0.6). Each dot corresponds to average runtime and object purity over ~20 synthetic videos with 100 frames each. Whiskers indicate standard deviation. B) Effect of noise on object purity across correspondence estimation methods. Methods such as IPFPMAP, CRF are more robust to increase in noise level compared to other methods.

Next, we systematically assessed the robustness of methods to increasing noise in synthetic video data. We created synthetic videos across a range of noise levels by varying precision and recall parameters of synthetic video (that is percentage of false positive detections and missed detections) from 60% to 90%. Across all correspondence estimation methods, accuracy falls with increasing noise in data. Interestingly, some methods such as L2QPMAP, GLTP, IPFPMAP, CRF are more robust to noise levels compared to others as their accuracy fell by least amount when noise was increased from 0.9 to 0.6. Similar trends were observed when tracking was performed using ‘all\_to\_random’ strategy. Taken together these results indicate that accuracy of tracking methods is highly contingent on the type of correspondence estimation method, track linking strategy used and noise levels in the video.

Finally, we conducted a screen by evaluating all correspondence estimation methods on a multi-dimensional criteria. Along with standard accuracy metrics, we also included a robustness score for each method defined as the percentage fall in accuracy of methods when noise level is increased from 0.9 to 0.6. Multi-dimensional comparison shows that the methods achieve varied performance profiles (Figure 4.5A) thus inspecting just one accuracy metric may not be enough to identify top performing methods on variety of tracking tasks. Next, we ranked all methods in each accuracy criterion and calculated a final rank to identify methods that perform best across all criteria (Figure 4.5B). Interestingly, some methods such as IPFPS, IPFPU, SMIPFP that performed well in robustness score did not perform well in other metrics like object purity, track link accuracy, etc. Further, many methods such as GMMReg, TPSRPM, SMIPFP, IPFPgm etc. performed adequately in one of the metrics while not performing effectively across other metrics. Finally, our objective screen identified methods such as L2QPMAP, IPFPMAP, CRF, GLTP that achieve high accuracy across all metrics. Among these methods, GLTP is the only linear method that has small runtime and performs well across all accuracy metrics. Thus, an objective multi-dimensional method screening with our toolbox enabled identification of several previously unexplored methods such as GLTP, L2QPMAP, IPFPMAP etc. with potential of achieving high accuracy in tracking nuclei in *C. elegans* whole-brain videos.

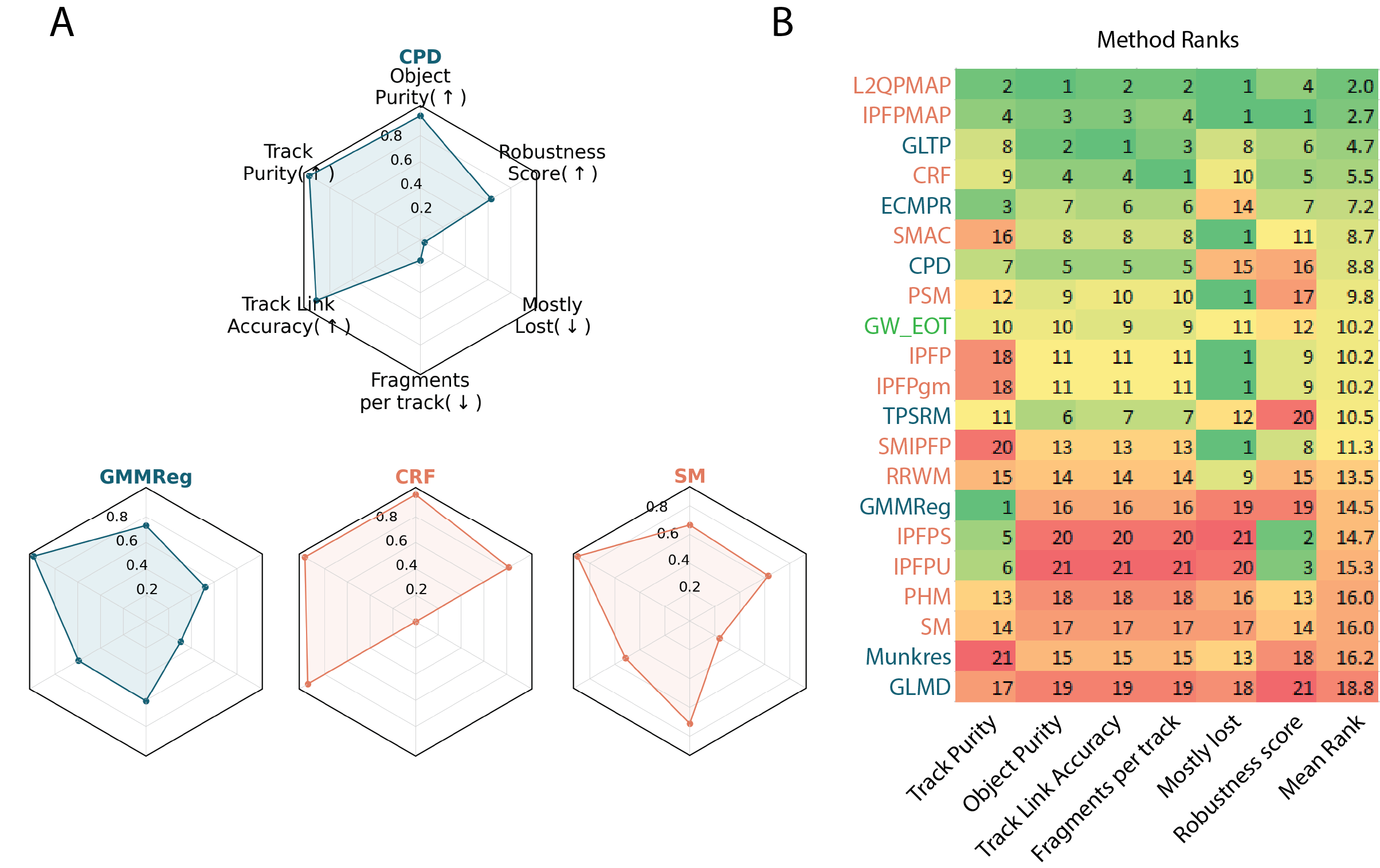


Figure 5. Multidimensional screen of tracking methods using synthetic data. A) Examples of MOT accuracy achieved by some correspondence estimation methods. B). Ranking of methods across Track Purity, Object Purity, Track Link Accuracy, Fragments per track, Mostly Lost and Robustness Score. Methods are sorted based on mean rank across all metrics.

## Multi-dimensional accuracy screen on whole-brain recording of freely moving worms

Multi-dimensional accuracy seen using synthetic data generated several insights about high performing track linking strategies, accurate and robust correspondence estimation methods. Next, we examined if the insights generated using synthetic datasets are repeated in experimental data as well. We tracked nuclei in pre-segmented whole-brain recording of *C. elegans* roaming freely on agar pad. The modular architecture of our toolbox allows us to mix and match correspondence estimation methods with different track linking strategies and compare accuracy. Similar to results in synthetic data, ‘all to random’ track linking strategy performed better than ‘sequential’ (Figure 4.6A). Here again we found a range of object purity accuracies across different methods.

To explore the robustness of methods on noise levels, we synthetically introduced false positive cells and randomly removed true positive cells from ground-truth cells tracked in the experimental data and compared object purity accuracy achieved by various methods (Figure 4.6B). Similar to results obtained on synthetic data, accuracy of all methods decreased as noise levels were increased (Figure 4.6B). Some methods were more robust compared to others such as TPSRPM, IPFPS, PSM, however these methods achieved lower object purity compared to other methods. Interestingly, CRF method achieved high accuracy compared to the robust methods (TPSRPM, IPFPS, PSM) and higher robustness compared to other high performing methods (GLTP, GMMReg, CPD).

Finally, similar to screen conducted using synthetic data, we evaluated all correspondence estimation methods on multi-dimensional criteria that included MOT accuracy metrics and robustness score. Sorting methods based on overall ranks achieved across these accuracy criteria revealed high performing methods (Figure 4.6C). GLTP and CRF methods that were ranked highly in synthetic data screen were ranked highly in experimental data screen as well. However, some methods such as L2QPMAP, IPFPT were ranked highly in synthetic data but not in experimental whole-brain data. This could be due to differences in properties of synthetic data and whole-brain video data, such as fewer cells were tracked in experimental data compared to synthetic data. Additionally, non-rigid deformations were not modeled in synthetic data.

Thus, in general the insights generated using synthetic videos are recapitulated highlighting the importance of simulations conducted using synthetic data. Multi-dimensional screen using standardized metrics identified methods that achieve high accuracy in tracking nuclei in freely moving animals. Further, components in the toolbox will enable researchers to benchmark and optimize new tracking methods against previous methods.

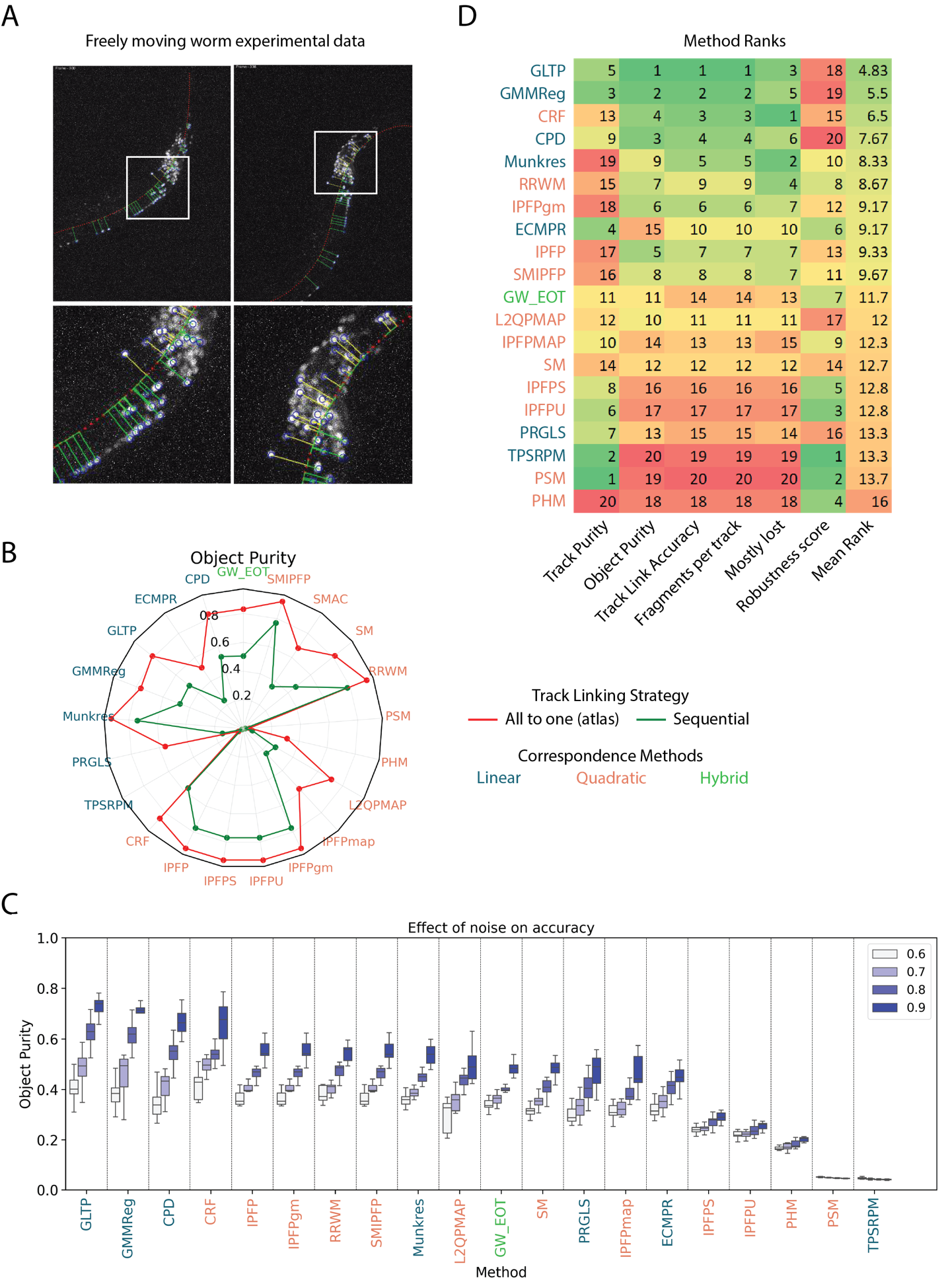


Figure 6. Robustness comparison and multi-dimensional comparison across methods on experimental whole-brain videos. A) Example frames showing non-rigid deformations in head of freely moving worm. Red dotted line indicates centerline. Yellow and green lines perpendicular to the centerine that orginate from individual nuclei were used to defined straigtened coordinates across time. B) Object purity achieved by methods as synthetic noise added to the video were increased. Boxes indicate 25th and 75th percentile. Whiskers indicate 5-95th percentile. C) Ranking of methods across Track Purity, Object Purity, Track Link Accuracy, Fragments per track, Mostly Lost and Robustness Score. Methods are sorted based on mean rank across all metrics.

## Conclusion

In this work we presented a set of tools to optimize that enable researcher to overcome barriers in developing new and optimized tracking algorithms for whole-brain imaging in *C. elegans*. These barriers include lack of ground-truth tacking data needed to optimize methods, difficulty in benchmarking new methods by comparing against previous correspondence estimation methods and track linking strategies, and inconsistent accuracy metrics used across works which make it difficult to compare methods. To solve these barriers our toolbox 1) enables generation of realistic synthetic video data corrupted with noise commonly found in experimental data and know ground-truth tracks, 2) provides callable implementations of 21 correspondence estimation methods for linking nuclei across frames to enable faster benchmarking and adoption of useful methods, and 3) outputs 6 standardized accuracy metrics used in MOT tasks thus facilitating objective comparison of methods.

Using the toolbox, we conducted extensive comparisons using synthetic data to compare methods in terms of runtime, robustness to noise in videos, and performance on MOT accuracy metrics. We identified optimal track linking strategies and new correspondence estimation methods such as GLTP, CRF that achieve accuracy across all accuracy metrics. We validated the insights using experimental data of freely moving *C. elegans*. Overall, accuracy comparisons across quadratic and linear methods using synthetic and experimental datasets revealed that in general, quadratic methods achieve higher accuracy and are more robust to noise compared to linear methods. GLTP is one of linear methods that achieved high accuracy on synthetic and experimental data screens. It performs better compared to other linear methods such as CPD because of an additional topology preservation71 cost added to the objective function. This cost enforces topology preserving constraint as point cloud of frame is deformed to match the reference point cloud for finding correspondence between cells in two frames. Further, it was revealed that accuracy performance across quadratic methods varies, although same edge features were used across all methods. This could be due to difference in optimization techniques used across methods to minimize the quadratic energy function.

An interesting result revealed by comparisons on synthetic showed that ‘all\_to\_reference’ track linking strategy, where each frame in the video is matched to an atlas frame, performs better than ‘all\_to\_random’ tracking strategy where each frame in the video is matched to a randomly selected frame in the video. Thus, a method that can build a reference or atlas frame that indicates positions of all nuclei in the head using noisy and incomplete information present in individual frames of the video will be extremely beneficial for tracking. Joint point cloud registration87 could be one approach to build such an atlas frame.

Methods available in our toolbox can be used to optimize tracking methods in other domains as well such as cell tracking in fluorescent images etc. For example, easy callable functions for correspondence estimation methods and track linking strategies will enable researcher to quickly compare methods and set up tracking pipelines. These pipelines can serve as baselines for comparisons with custom designed tracking methods. Further, tracking methods can be easily benchmarked with use of standardized accuracy metrics.

# References

1. Ahrens, M. B. *et al.* Brain-wide neuronal dynamics during motor adaptation in zebrafish. *Nature* **5**, 471–477 (2012).

2. Ahrens, M. B., Orger, M. B., Robson, D. N., Li, J. M. & Keller, P. J. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nat. Methods* **10**, 413–420 (2013).

3. Dunn, T. W. *et al.* Brain-wide mapping of neural activity controlling zebrafish exploratory locomotion. *Elife* **5**, (2016).

4. Randlett, O. *et al.* Whole-brain activity mapping onto a zebrafish brain atlas. *Nat. Methods* **12**, 1–12 (2015).

5. Naumann, E. A. *et al.* From Whole-Brain Data to Functional Circuit Models: The Zebrafish Optomotor Response. *Cell* **167**, 947-960.e20 (2016).

6. Lovett-Barron, M. *et al.* Ancestral Circuits for the Coordinated Modulation of Brain State. *Cell* **171**, 1411-1423.e17 (2017).

7. Keller, P. J. & Ahrens, M. B. Visualizing whole-brain activity and development at the single-cell level using light-sheet microscopy. *Neuron* vol. 85 462–483 at https://doi.org/10.1016/j.neuron.2014.12.039 (2015).

8. Schrödel, T., Prevedel, R., Aumayr, K., Zimmer, M. & Vaziri, A. Brain-wide 3D imaging of neuronal activity in Caenorhabditis elegans with sculpted light. *Nat. Methods* **10**, 1013–1020 (2013).

9. Prevedel, R. *et al.* Simultaneous whole-animal 3D-imaging of neuronal activity using light-field microscopy. *Nat. Methods* **11**, 727–730 (2015).

10. Nguyen, J. P. *et al.* Whole-brain calcium imaging with cellular resolution in freely behaving Caenorhabditis elegans. *Proc. Natl. Acad. Sci. U. S. A.* 33 (2015) doi:10.1073/pnas.1507110112.

11. Venkatachalam, V. *et al.* *Pan-Neuronal Imaging in Roaming Caenorhabditis Elegans.* *Proceedings of the National Academy of Sciences of the United States of America* vol. 113 (2016).

12. Kim, D. H. *et al.* Pan-neuronal calcium imaging with cellular resolution in freely swimming zebrafish. *Nat. Methods* **14**, 1107–1114 (2017).

13. Cong, L. *et al.* Rapid whole brain imaging of neural activity in freely behaving larval zebrafish (Danio rerio). *Elife* **6**, (2017).

14. Symvoulidis, P. *et al.* NeuBtracker - Imaging neurobehavioral dynamics in freely behaving fish. *Nat. Methods* **14**, 1079–1082 (2017).

15. Kato, S. *et al.* Global Brain Dynamics Embed the Motor Command Sequence of Caenorhabditis elegans. *Cell* **163**, 656–669 (2015).

16. Kaplan, H. S., Salazar Thula, O., Khoss, N. & Zimmer, M. Nested Neuronal Dynamics Orchestrate a Behavioral Hierarchy across Timescales. *Neuron* (2019) doi:https://doi.org/10.1016/j.neuron.2019.10.037.

17. Scholz, M. *et al.* Predicting natural behavior from whole-brain neural dynamics. *bioRxiv* 445643 (2018) doi:10.1101/445643.

18. Atanas, A. A. *et al.* Brain-wide representations of behavior spanning multiple timescales and states in C. elegans. *Cell* (2023) doi:10.1016/j.cell.2023.07.035.

19. Kramer, T. S. *et al.* Neural Sequences Underlying Directed Turning in C. elegans. *bioRxiv* (2024).

20. Fieseler, C. *et al.* An intrinsic neuronal manifold underlies brain-wide hierarchical organization of behavior in C. elegans. *bioRxiv* 2003–2025 (2025).

21. Seyedolmohadesin, M. *et al.* Whole-brain chemosensory responses of both C. elegans sexes. *bioRxiv* 2005–2025 (2025).

22. Susoy, V. *et al.* Natural sensory context drives diverse brain-wide activity during C. elegans mating. *Cell* **184**, 5122-5137.e17 (2021).

23. Nichols, A. L. A., Eichler, T., Latham, R. & Zimmer, M. A global brain state underlies C. elegans sleep behavior. *Science* **356**, eaam6851 (2017).

24. Skora, S., Mende, F. & Zimmer, M. Energy Scarcity Promotes a Brain-wide Sleep State Modulated by Insulin Signaling in C. elegans. *Cell Rep.* **22**, 953–966 (2018).

25. Wirak, G. S., Florman, J., Alkema, M. J., Connor, C. W. & Gabel, C. V. Age-associated changes to neuronal dynamics involve a loss of inhibitory signaling in &lt;em&gt;C. elegans&lt;/em&gt; *bioRxiv* 2021.07.07.451497 (2021) doi:10.1101/2021.07.07.451497.

26. Dag, U. *et al.* Dissecting the functional organization of the C. elegans serotonergic system at whole-brain scale. *Cell* (2023) doi:10.1016/j.cell.2023.04.023.

27. Randi, F., Sharma, A. K., Dvali, S. & Leifer, A. M. Neural signal propagation atlas of Caenorhabditis elegans. *Nature* (2023) doi:10.1038/s41586-023-06683-4.

28. Cho, Y., Oakland, D. N., Lee, S. A., Schafer, W. R. & Lu, H. On-chip functional neuroimaging with mechanical stimulation in: Caenorhabditis elegans larvae for studying development and neural circuits. *Lab Chip* **18**, 601–609 (2018).

29. Cho, Y. *et al.* Multimodal Stimulation in a Microfluidic Device Facilitates Studies of Interneurons in Sensory Integration in C. elegans. *Small* **n/a**, 1905852 (2020).

30. Rouse, T., Aubry, G., Cho, Y., Zimmer, M. & Lu, H. A programmable platform for sub-second multichemical dynamic stimulation and neuronal functional imaging in: C. elegans. *Lab Chip* **18**, 505–513 (2018).

31. Kwon, Y. *et al.* Advanced Neural Functional Imaging in C. elegans Using Lab-on-a-Chip Technology. *Micromachines* **15**, 1027 (2024).

32. Lee, S. A., Cho, Y., Schafer, W. R. & Lu, H. Dynamic temperature control in microfluidics for in vivo imaging of cold-sensing in C. elegans. *Biophys. J.* **123**, 947–956 (2024).

33. Lee, H. J., Vallier, J. & Lu, H. Microfluidic localized hydrogel polymerization enables simultaneous recording of neural activity and behavior in C. elegans. *React. Chem. Eng.* **9**, 666–676 (2024).

34. Gomez, K., Yarmey, V. R., Mane, H. & San-Miguel, A. Microfluidic and Computational Tools for Neurodegeneration Studies. *Annu. Rev. Chem. Biomol. Eng.* **16**, (2025).

35. Giovannucci, A. *et al.* CaImAn an open source tool for scalable calcium imaging data analysis. *Elife* **8**, (2019).

36. Pnevmatikakis, E. A. *et al.* Simultaneous Denoising, Deconvolution, and Demixing of Calcium Imaging Data. *Neuron* **89**, 299 (2016).

37. Pachitariu, M. *et al.* Suite2p: beyond 10,000 neurons with standard two-photon microscopy. *bioRxiv* 061507 (2016) doi:10.1101/061507.

38. Pacheco, D. A., Thiberge, S. Y., Pnevmatikakis, E. & Murthy, M. Auditory activity is diverse and widespread throughout the central brain of Drosophila. *Nat. Neurosci.* **24**, 93–104 (2021).

39. Mann, K., Gallen, C. L. & Clandinin, T. R. Whole-Brain Calcium Imaging Reveals an Intrinsic Functional Network in Drosophila. *Curr. Biol.* **27**, 2389-2396.e4 (2017).

40. Aimon, S. *et al.* Fast near-whole brain imaging in adult Drosophila during responses to stimuli and behavior. *bioRxiv* 033803 (2018) doi:10.1101/033803.

41. Brezovec, L. E., Berger, A. B., Druckmann, S. & Clandinin, T. R. Mapping the Neural Dynamics of Locomotion across the &lt;em&gt;Drosophila&lt;/em&gt; Brain. *bioRxiv* 2022.03.20.485047 (2022) doi:10.1101/2022.03.20.485047.

42. Lagache, T., Hanson, A., Pérez-Ortega, J. E., Fairhall, A. & Yuste, R. Tracking calcium dynamics from individual neurons in behaving animals. *PLOS Comput. Biol.* **17**, e1009432 (2021).

43. Vladimirov, N. *et al.* Brain-wide circuit interrogation at the cellular level guided by online analysis of neuronal function. *Nat. Methods* **15**, 1117–1125 (2018).

44. Migault, G. *et al.* Whole-Brain Calcium Imaging during Physiological Vestibular Stimulation in Larval Zebrafish. *Curr. Biol.* (2018) doi:10.1016/j.cub.2018.10.017.

45. Haesemeyer, M., Robson, D. N., Li, J. M., Schier, A. F. & Engert, F. A Brain-wide Circuit Model of Heat-Evoked Swimming Behavior in Larval Zebrafish. *Neuron* **98**, 817-831.e6 (2018).

46. Pacheco, D. A., Thiberge, S. Y., Pnevmatikakis, E. & Murthy, M. Auditory Activity is Diverse and Widespread Throughout the Central Brain of Drosophila. *bioRxiv* (2019).

47. Tokunaga, T. *et al.* Automated detection and tracking of many cells by using 4D live-cell imaging data. *Bioinformatics* **30**, i43-51 (2014).

48. Hirose, O. *et al.* SPF-CellTracker: Tracking Multiple Cells with Strongly-Correlated Moves Using a Spatial Particle Filter. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **15**, 1822–1831 (2018).

49. Nguyen, J. P., Linder, A. N., Plummer, G. S., Shaevitz, J. W. & Leifer, A. M. Automatically tracking neurons in a moving and deforming brain. (2016).

50. Chaudhary, S. & Lu, H. Point-set registration framework with Conditional Random Fields for automatic tracking of neurons in C . elegans whole-brain videos Point-set registration framework with Conditional Random Fields for automatic tracking of neurons in C . elegans whole-brai. *Work. Worm Neural Inf. Process. 31st Conf. Neural Inf. Process. Syst. (NIPS 2017), Long Beach, CA, USA* (2018).

51. Wen, C. *et al.* 3DeeCellTracker, a deep learning-based pipeline for segmenting and tracking cells in 3D time lapse images. *Elife* **10**, e59187 (2021).

52. Yu, X. *et al.* Wfast deep neural correspondence for tracking and identifying neurons in c. Elegans using semi-synthetic training. *Elife* (2021) doi:10.7554/eLife.66410.

53. Jones, C., Barzegar-Keshteli, M., Gross, A., Obozinski, G. & Rahi, S. J. A Graph Matching Approach to Tracking Neurons in Freely-Moving C. elegans. *bioRxiv* 2011–2023 (2023).

54. Park, C. F. *et al.* Automated neuron tracking inside moving and deforming C. elegans using deep learning and targeted augmentation. *Nat. Methods* **21**, 142–149 (2024).

55. Atanas, A. A. *et al.* Deep Neural Networks to Register and Annotate Cells in Moving and Deforming Nervous Systems. *bioRxiv* 2007–2024 (2024).

56. Ryu, J. *et al.* Versatile multiple object tracking in sparse 2D/3D videos via deformable image registration. *PLOS Comput. Biol.* **20**, e1012075 (2024).

57. Szigeti, B. *et al.* OpenWorm: an open-science approach to modeling Caenorhabditis elegans. *Front. Comput. Neurosci.* **8**, (2014).

58. Vayer, T., Chapel, L., Flamary, R., Tavenard, R. & Courty, N. Fused gromov-wasserstein distance for structured objects. *Algorithms* (2020) doi:10.3390/a13090212.

59. Vayer, T., Chapel, L., Flamary, R., Tavenard, R. & Courty, N. Optimal Transport for structured data with application on graphs. in *36th International Conference on Machine Learning, ICML 2019* (2019).

60. Cuturi, M. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. in *Advances in Neural Information Processing Systems* (eds. Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z. & Weinberger, K. Q.) vol. 26 (Curran Associates, Inc., 2013).

61. Solomon, J., Peyré, G., Kim, V. G. & Sra, S. Entropic metric alignment for correspondence problems. in *ACM Transactions on Graphics* vol. 35 (2016).

62. Peyre, G., Cuturi, M. & Solomon, J. Gromov-wasserstein averaging of kernel and distance matrices. in *33rd International Conference on Machine Learning, ICML 2016* vol. 6 3927–3935 (2016).

63. Leordeanu, M. & Hebert, M. Efficient MAP Approximation for Dense Energy Functions. in *Proceedings of the 23rd International Conference on Machine Learning* 545–552 (Association for Computing Machinery, New York, NY, USA, 2006). doi:10.1145/1143844.1143913.

64. Leordeanu, M., Hebert, M. & Sukthankar, R. An Integer Projected Fixed Point Method for Graph Matching and MAP Inference. in *Advances in Neural Information Processing Systems* (eds. Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. & Culotta, A.) vol. 22 (Curran Associates, Inc., 2009).

65. Schmidt, M. UGM: A Matlab toolbox for probabilistic undirected graphical models. http://www.cs.ubc.ca/~schmidtm/Software/UGM.html, 2007. at https://www.cs.ubc.ca/~schmidtm/Software/UGM.html (2007).

66. Chaudhary, S., Lee, S. A., Li, Y., Patel, D. S. & Lu, H. Graphical-model framework for automated annotation of cell identities in dense cellular images. *Elife* **10**, e60321 (2021).

67. Cour, T., Srinivasan, P. & Shi, J. Balanced graph matching. in *Advances in Neural Information Processing Systems* 313–320 (2007). doi:10.7551/mitpress/7503.003.0044.

68. Zhou, F. & De La Torre, F. Factorized Graph Matching. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**, 1774–1789 (2016).

69. Leordeanu, M. & Hebert, M. A spectral technique for correspondence problems using pairwise constraints. in *Proceedings of the IEEE International Conference on Computer Vision* vol. II 1482–1489 (2005).

70. Cho, M., Lee, J. & Lee, K. M. Reweighted random walks for graph matching. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol. 6315 LNCS 492–505 (2010).

71. Ge, S., Fan, G. & Ding, M. Non-rigid point set registration with global-local topology preservation. in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* 245–251 (2014). doi:10.1109/CVPRW.2014.45.

72. Myronenko, A. & Song, X. Point-Set Registration: Coherent Point Drift. *Adv. Neural Inf. Process. Syst. 19* 1009–1016 (2009) doi:10.1109/TPAMI.2010.46.

73. Chui, H. & Rangarajan, A. A new point matching algorithm for non-rigid registration. *Comput. Vis. Image Underst.* **89**, 114–141 (2003).

74. Horaud, R., Forbes, F., Yguel, M., Dewaele, G. & Zhang, J. Rigid and articulated point registration with expectation conditional maximization. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**, 587–602 (2011).

75. Jian, B. & Vemuri, B. C. A Robust Algorithm for Point Set Registration Using Mixture of Gaussians. *Proc. IEEE Int. Conf. Comput. Vis.* **2**, 1246–1251 (2005).

76. Yang, Y., Ong, S. H. & Foong, K. W. C. A robust global and local mixture distance based non-rigid point set registration. *Pattern Recognit.* **48**, 156–173 (2015).

77. Ma, J., Zhao, J. & Yuille, A. L. Non-rigid point set registration by preserving global and local structures. *IEEE Trans. Image Process.* **25**, 53–64 (2016).

78. Ma, J. *et al.* Robust L2E estimation of transformation for non-rigid registration. *IEEE Trans. Signal Process.* **63**, 1115–1129 (2015).

79. Kuhn, H. W. The Hungarian Algorithm for the Assignment Problem. *Nav. Res. Logist. Q.* **2**, 83–97 (1955).

80. Nguyen, J. P., Linder, A. N., Plummer, G. S., Shaevitz, J. W. & Leifer, A. M. Automatically tracking neurons in a moving and deforming brain. *PLoS Comput. Biol.* **13**, (2017).

81. Jaqaman, K. *et al.* Robust single-particle tracking in live-cell time-lapse sequences. *Nat. Methods* **5**, 695–702 (2008).

82. Wang, C., Wang, Y. & Yu, G. Efficient Global MOT under Minimum-Cost Circulation Framework. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 1888–1904 (2022).

83. Smith, K., Gatica-Perez, D., Odobez, J. M. & Ba, S. Evaluating multi-object tracking. in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* vols 2005-Septe (2005).

84. Wu, B. & Nevatia, R. Tracking of multiple, partially occluded humans based on static body part detection. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* vol. 1 951–958 (2006).

85. Bernardin, K., Elbs, A. & Stiefelhagen, R. Multiple object tracking performance metrics and evaluation in a smart room environment. *Sixth IEEE Int. Work. Vis. Surveillance, conjunction with ECCV* **90**, 91 (2006).

86. Milan, A., Leal-Taixe, L., Reid, I., Roth, S. & Schindler, K. MOT16: A Benchmark for Multi-Object Tracking. *arXiv e-prints* arXiv:1603.00831 (2016).

87. Evangelidis, G. D., Kounades-Bastian, D., Horaud, R. & Psarakis, E. Z. A generative model for the joint registration of multiple point sets. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol. 8695 LNCS 109–122 (2014).