Super-resolution fight club: Assessment of 2D & 3D single-molecule localization microscopy software

Daniel Sage[+1], Thanh-An Pham[+1], Hazen Babcock[2], Tomas Lukes[3,4], Thomas Pengo[5], Jerry Chao[6,7], Ramraj Velmuruga[7,8], Alex Herbert[9], Anurag Agrawal[10], Silvia Colabrese[1,11], Ann Wheeler[12], Anna Archetti[13], Bernd Rieger[14], Raimund Ober[6,7,15], Guy M. Hagen[16], Jean-Baptiste Sibarita[17,18], Jonas Ries[19], Ricardo Henriques[20], Michael Unser[1], Seamus Holden[*+21]*

*Corresponding authors: daniel.sage@epfl.ch, seamus.holden@ncl.ac.uk.

+Equal contribution

1: Biomedical Imaging Group, School of Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
2: Harvard Center for Advanced Imaging, Harvard University, Cambridge, Massachusetts, USA
3: Laboratory of Nanoscale Biology & Laboratoire d'Optique Biomédicale, STI - IBI, EPFL, Lausanne, Switzerland
4: Department of Radioelectronics, FEE, Czech Technical University, Prague, Czech Republic
5: University of Minnesota Informatics Institute, University of Minnesota Twin Cities, USA
6: Department of Biomedical Engineering, Texas A&M University, College Station, Texas, USA
7: Department of Molecular and Cellular Medicine, Texas A&M University Health Science Center, College Station, Texas, USA
8: Department of Microbial Pathogenesis and Immunology, Texas A&M University Health Science Center, Bryan, Texas, USA
9: MRC Genome Damage and Stability Centre, School of Life Sciences, University of Sussex, Brighton, UK
10 : Double Helix LLC, Boulder, Colorado, USA
11 : Istituto Italiano di Tecnologia, Genova, Italy
12: Advanced Imaging Resource, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
13 : Laboratory of Experimental Biophysics, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
14: Department of Imaging Physics, Delft University of Technology, The Netherlands
15: Centre for Cancer Immunology, University of Southampton, Southampton, UK
16: UCCS center for the Biofrontiers Institute, University of Colorado at Colorado Springs, Colorado, USA
17: Interdisciplinary Institute for Neuroscience, University of Bordeaux, Bordeaux, France
18: Interdisciplinary Institute for Neuroscience, Centre National de la Recherche Scientifique (CNRS) UMR 5297, Bordeaux, France
19: European Molecular Biology Laboratory (EMBL), Cell Biology and Biophysics Unit, Heidelberg, Germany
20: Quantitative Imaging and Nanobiophysics Group, MRC Laboratory for Molecular Cell Biology, University College London, UK
21: Centre for Bacterial Cell Biology, Institute for Cell and Molecular Biosciences, Newcastle University, UK

ABSTRACT

With the widespread uptake of 2D and 3D single molecule localization microscopy, a large set of different data analysis packages have been developed to generate super-resolution images. In a large community effort we designed a competition to extensively characterise and rank the performance of 2D and 3D single molecule localization microscopy software packages. We generated realistic simulated datasets for popular imaging modalities – 2D, astigmatic 3D, biplane 3D, and double helix 3D – and evaluated 36 participant packages against these data. This provides the first broad assessment of 3D single molecule localization microscopy software and provides a holistic view of how the latest 2D and 3D single molecule localization software perform in realistic conditions. This resource allows researchers to identify optimal analytical software for their experiments, allows 3D SMLM software developers to benchmark new software against current state of the art, and provides insight into the current limits of the field.

INTRODUCTION

Image processing software is central to single molecule localization microscopy (SMLM[1–3]). Efficient and automated image processing is essential to extract the super-resolved positions of individual molecules from thousands of raw microscope images, containing millions of blinking fluorescent spots. Improvements in SMLM image processing have been crucial in maximizing spatial resolution and reducing imaging time of SMLM for compatibly with live cell imaging[4–6]. If SMLM is to achieve a resolving power approaching that of electron microscopy, the analysis software employed needs to be robust, accurate, and performing at current algorithmic limits. This can only be achieved through rigorous quantification of SMLM software performance.

The first localization microscopy software challenge was carried out in 2013 to benchmark 2D SMLM software[7]. But biology is not just a 2D problem, and a key focus of localization microscopy is 3D imaging of nanoscale cellular processes[8,9]. 3D localization microscopy is a more difficult image processing problem than 2D SMLM. In addition to finding the center of diffraction limited spots to super-resolve lateral position, 3D SMLM algorithms must also extract axial information from the image, usually by measuring small changes in the shape of a point spread function[10] (PSF).

Despite the widespread use of 3D localization microscopy, and challenging nature of 3D SMLM image processing, the performance of software for 3D single molecule localization microscopy has previously only been assessed for 2-3 software packages at a time, and without standard test data or metrics[11–14]. In the absence of common reference datasets and reliable assessment, it is not possible to objectively assess how different software affect final image quality, or which algorithmic approaches are most successful. Crucially, end-users cannot determine which 3D SMLM software package and imaging modality is optimal for their application.

We therefore ran the first 3D localization microscopy software challenge, to assess the performance of 3D SMLM software. We assessed software performance on simulated datasets designed for maximum realism, incorporating experimentally derived point spread functions, using biologically inspired structures, using signal to noise levels based closely on common experimental conditions, and modelling fluorophore photophysics. We assessed software performance on synthetic datasets for three popular 3D SMLM modalities: astigmatic imaging[10], biplane imaging[15] and double helix point spread function microscopy[16]. We also assessed astigmatism software performance on two real STORM datasets. Furthermore, we ran a second 2D localization microscopy software challenge to assess performance of the latest 2D SMLM software.

RESULTS

Competition design

We established a broad committee from the SMLM community, including experimentalists and software developers, to define the scope of the challenge, ensure realism of the datasets and define analysis metrics. We opened this discussion to all interested parties in an online discussion forum[17].

In 2016, we ran a first round of the 3D SMLM competition with explicit submission deadlines, culminating in a special session at the 6th annual Single Molecule Localization Microscopy Symposium (SMLMS 2016). Since then, the challenge has been opened to continuously accept new entries. Thirty-six software packages have been entered in the competition thus far, including four packages used in commercial software (**Table S1**, **Supplementary Note 1**). Participation in the competition actually led at least eight teams to modify their software to support additional 3D SMLM modalities, showing how competition can foster microscopy software development.

Realistic 3D simulations

Testing super-resolution software on experimental data lacks the ground truth information required for rigorous quantification of software performance. Therefore, realistic simulated datasets are required. A critical challenge to in simulating 3D SMLM data was accurate modeling of the experimental microscope PSF for each 3D modality. 3D SMLM inherently involves addition of aberrations to the microscope PSF to encode the Z-position of the molecule. For the PSF models included in the competition: astigmatic (AS), double helix (DH), and biplane (BP), we observed that the PSFs showed complex aberrations not well described by simple analytical models (**Fig. S1**). Even experimental 2D PSFs showed significant aberrations away from the focal plane (**Fig. S1**).

We thus combined experimental 3D PSFs with simulated ground truth by performing simulations using PSFs directly derived from experimental calibration data (**Fig. 1, Methods**). We generated simulated datasets over a range of spot densities and signal to noise levels, for simulated microtubule- and endoplasmic reticulum-like structures, using a 4-state model for photophysics[18] (**Methods**).

Quantitative performance assessment of 3D software

We assessed software performance by 26 quality metrics (**Supplementary Note 2**). The complete set of summary statistics, axially resolved performance and super-resolved images is available for each competition software on the competition website. We built an interactive ranking and graphing interface for ranking and plotting software performance by any metric, including new user defined metrics (**Fig. S2**). Detailed individual software reports are also available, along with a tool for side-by-side comparison of software (**Fig. S2, S3**).

We focused our primary analysis on metrics directly assessing performance in detecting individual molecules. This was based on three key metrics *(***Methods***)*:

1. *Root mean squared localization error* (RMSE) between measured molecule position and the ground truth.
2. *Jaccard index* (JAC). This quantifies the fraction of correctly detected molecules in a dataset.
3. *Efficiency (E).* For ranking purposes, we developed a single summary statistic for overall evaluation of software performance combining RMSE and Jaccard index, which we term the *efficiency* **(Methods)**.

Choice of ranking metric is discussed in **Supplementary Note 2**, where several alternative ranking metrics are also presented.

Performance of 3D software

Complete rankings for each imaging modality and spot density are presented (**Fig. 2**), together with summary information on all competition software (**Supplementary Table 1**, **Supplementary Note 1**).

After assembling an overall summary of best performers for each competition category, we investigated the performance of software within each imaging modality.

Astigmatic localization microscopy

Astigmatic localization microscopy is probably the most popular 3D SMLM modality, reflected by the highest number of software submissions in the 3D competition (**Fig. 2**). For astigmatism, we observed a large spread of software performance, even for the most straightforward high SNR, low spot density (LD) conditions (**Fig. 3, Supplementary Table 2**). The best-in-class software (SMAP-2018[19]) has significantly better localization error and Jaccard index performance than average (lateral RMSE 26 nm best vs 38 nm average, axial RMSE 29 nm best vs 66 nm average, Jaccard index 85 % best vs 74 % average). Clearly, the quality of the image reconstruction depends strongly on choice of 3D software.

To investigate the reasons for software variation, we inspected plots of software performance as a function of axial position in the low density, high SNR dataset for best-in-class and representative middle-range software (**Fig. S4A**). We observed that a key cause of the spread in software performance is variation in software performance away from the focal plane. Near the focal plane, most software packages perform well. However, the axial and lateral RMSE away from the plane of focus is significantly higher for the best in class software, and the Jaccard index is also slightly improved

(**Fig. S4A**). This is also visibly apparent in the super-resolved images (**Fig. 4A**). We observed that best-in-class software had a Z-range (the FWHM range of axially resolved software recall, **Methods**) of 1170 nm, greater than two-thirds of the simulated range. Outside this range, the recall and Jaccard index dropped sharply, probably due the large increase in PSF size and decrease in effective SNR at large defocus (**Fig. S1**).

When we examined results for the low SNR, low density dataset (**Fig. 2A, 3F**), we found an expected two-fold degradation in best-in-class RMSE (lateral RMSE 39 nm, axial RMSE 60 nm), due to the decrease in image SNR. However, the best-in-class software (SMolPhot[20]) Jaccard index was effectively constant between the low and high SNR datasets (86 % vs 85 %), although the Z-range did drop at lower SNR (930 nm vs 1120 nm). The best astigmatism software packages were thus remarkably good at finding spots at low SNR, even away from the focal plane.

We compared best-in-class software performance to Cramér-Rao lower bound (CRLB) theoretical limits (**Fig. S5, S6, Supplementary Note 3**). Close to the focus, best-in-class software was near the CRLB (within 25 %), but significant deviations from the CRLB occurred > 200 nm (**Fig. S6**). This could be due to difficulty in distinguishing signal from false positives away from focus.

Astigmatic software performance dropped for the challenging high spot density datasets (**Fig. 2A, 3**). For the high SNR high spot density dataset (best software, SMolPhot), localization error increased and Jaccard index decreased significantly compared to the low density condition (lateral RMSE best HD 51 nm vs best LD 27 nm, axial RMSE best HD 66 nm vs best LD 29 nm, Jaccard index best HD 66 % vs best LD 85 %). Inspection of the super-resolved images (**Fig. S7**) nevertheless shows qualitatively acceptable results for the HD dataset, particularly in the lateral dimension. In some circumstances, the performance reduction at 10x higher spot density could be acceptable for 10x faster, potentially live-cell-compatible, imaging speed. We also observed a large spread of software performance for the high density datasets, probably because a significant fraction of the software packages were primarily designed for low density conditions.

We observed poor performance for the most challenging low SNR high spot density astigmatism dataset (**Fig. 2A, 3, S8**, best software SMolPhot). Best-in-class localization precision and Jaccard index decreased significantly (lateral RMSE 76 nm, axial RMSE 101 nm, Jaccard index 58 %). These data suggest that low SNR high density 3D astigmatic localization microscopy entails significant reduction in image resolution.

Double helix point spread function localization microscopy

We next analyzed the performance of the double helix software (**Fig. 3D-F, S9A**). For the software in the high SNR low spot density condition, double helix software showed more uniform performance than astigmatism. Best-in-class software (SMAP-2018) showed only a limited improvement compared with average software (**Fig. 3D-F**, lateral RMSE, 27 nm best vs 37 nm average; axial RMSE 21 nm best vs 34 nm average; Jaccard index 77 % best vs 73 % average). In general software localization performance was close to the CRLB (**Fig. S6**). We observed that performance of the software away from the focal plane is relatively uniform (**Fig. 4A, S4A**), and best-in-class Z-range at high SNR was large at 1180 nm (**Fig. S4A, Supplementary Table 2**). Double helix imaging may show less software-to-software variation and larger Z-range at low spot density than astigmatic imaging because the PSF shape and intensity are fairly constant as a function of Z; unlike astigmatic imaging, where spot size, shape and intensity vary greatly as a function of Z (**Fig. S1**).

Double helix software performance decreased significantly for the low spot density low SNR condition (best software, SMAP-2018), particularly in terms of best-in-class Jaccard index (66 % low SNR vs 77 % high SNR, **Fig. 3D-E, S8, S9A**). DH Jaccard index was also significantly worse than astigmatism results at either high or low SNR (85 % high SNR, 86 % low SNR). This poor performance in the low SNR DH dataset is likely because the large size of the DH PSF spreads emitted photons over a large area, lowering effective image SNR. DH PSF designs with reduced Z-range but more compact PSF would likely be less sensitive to this issue[21].

Double helix software performed poorly on the high spot density datasets at high SNR (best software CSpline[22]), especially in terms of the Jaccard index (**Fig. 3D-E, S9A**, best lateral RMSE 67 nm, best axial RMSE 69 nm, best Jaccard index 46 %). The poor performance at high spot density is again probably because the large DH PSF size increases spot density and decreases SNR (**Fig. S1**). DHPSF performance at high spot density and low SNR was also not reliable (**Fig. 3D-F, S9A**, best software, SMAP-2018).

Biplane localization microscopy

Best-in-class biplane software (SMAP-2018), at low spot density and for both high and low SNR, delivered the best performance in any modality (high SNR: lateral RMSE 12.3 nm, axial RMSE 21.7 nm, Jaccard 87 %), despite a slightly decreased image SNR for the biplane simulations (**Methods**). We observed a large spread in software performance in terms of lateral RMSE and Jaccard index, with the best-in-class software significantly outperforming the other competitors (**Fig. S9B, 2D**). At low spot density, best-in-class biplane software (SMAP-2018) showed good performance as a function of Z, with high Jaccard index over almost the entire Z-range of the simulations, and with a Z-range of 1200 nm at high SNR (**Fig. S4AC, Supplementary Table 2**). The axial RMSE was relatively uniform as a function of Z and close to the CRLB limit (**Fig. S6**). As axial and lateral RMSE are both averaged over the entire Z-range, the strong biplane results arise from good performance across a large Z-range (**Fig. S4**).

At high spot density and high SNR, best-in-class biplane software (SMAP-2018) showed acceptable performance (**Fig. 3D-F, S7, S9B**, best lateral RMSE 43 nm, best axial RMSE 49 nm, best Jaccard index 61 %). Uniquely among the 3D modalities, best-in-class biplane software also gave acceptable performance at high spot density and low SNR (**Fig. 3D-F, S7, S9B**, best lateral RMSE 55 nm, best axial RMSE 72 nm, best Jaccard index 61 %, best software SMAP-2018).

Performance of 2D software

We next assessed the performance of 2D SMLM software. For the pseudo-ER 2D dataset at low density, best-in-class software (ADCG[23]) performed substantially better than the class average (**Fig. S10, S11**, lateral RMSE 31 nm vs 36 nm average, Jaccard index 90 % best vs 72 %). Low density results for the brighter fluorophore microtubules dataset were similar to the dimmer pseudo-ER dataset (**Fig. S10, S12** best software SMolPhot). For the very high density 2D dataset, which had 25x higher spot density than the LD dataset, best-in-class software (ADCG) showed excellent performance (**Fig. S10**, lateral RMSE, 45.5 nm, Jaccard index 75%). Best-in-class performance (ADCG) on the dimmer fluorophore data at high spot density was also strong (**Fig. S10**, best lateral RMSE 51 nm, best Jaccard index 70 %).

Algorithms

We identified several classes of algorithms in the participant software (**Supplementary Table 1**):

1) *Non-iterative* software regroups pixels in the local neighborhood of the candidates, like interpolation, center of mass (QuickPALM[24]) or template matching (WTM[25]). These often older algorithms are fast but tend to perform poorly.

2) S*ingle emitter fitting* software is usually built on a multi-step strategy of detection, spot localization, and optional spot rejection. The detection step finds bright spots in noisy images on the pixel grid. The selection of candidates is usually performed by local maximum search after a denoising filter. Others rely on more complex algorithms like the wavelet transform (WaveTracer[26]). We did not observe software ranking to depend noticeably on the choice of optimization scheme: least-square, weighted least-square or maximum-likelihood estimator.

3) *Multi-emitter fitting* software groups clusters of overlapping spots and simultaneously fits multiple model PSFs to the data. Typically, fitted spots are added to the cluster until a stopping condition is met[4,5]. This leads to improved localization performance at high spot density, at the cost of reduced speed. This class of software (*e.g.,* 3D-DAOSTORM[11], CSpline, PeakFit, ThunderSTORM[27]) was amongst the top performers in each 2D and 3D competition category.

As expected, single- and multiple-emitter fitting methods both performed well on low density data. For the 2D challenge, multi-emitter fitting showed a clear advantage over single emitter fitting at high density. Surprisingly however, well-tuned single-emitter fitting algorithms slightly outperformed multi-emitter algorithms for 3D high density conditions (e.g., astigmatism, SMolPhot vs 3D-DAOSTORM). This result merits further investigation as it conflicts with results for 2D software, and with naïve expectation, which suggests multi-emitter fitting should be a better model for data where PSFs overlap significantly.

4) *Compressed sensing algorithms.* One subset of these algorithms utilize deconvolution with sparsity constraints to reconstruct super-resolved images[28–30]. Although deconvolution approaches can give good results, they are limited by the necessary use of a sub-pixel grid; increased localization precision requires smaller grid resolution, which must be balanced against increased computational time. Recent approaches address this issue by localizing the point sources in a gridless manner under some sparsity constraint (ADCG, SMfit, SOLAR_STORM, TVSTORM[31]). This software class consistently gave the overall best performance for 2D high-density (ADCG 1st, FALCON[30] 2nd, SMfit 3rd).

5) *Other approaches.* Of the alternative algorithmic approaches used, the annihilating filter-based method LEAP[32] gave good performance for biplane imaging. Recently, we received the first challenge submission from a deep learning SMLM software (DECODE); these promising preliminary results are available on the competition website.

Post-hoc temporal grouping

Because molecule on-time is stochastically distributed across multiple frames, a common post-processing approach to improve localization precision is to group molecules detected multiple times in adjacent frames, and average their position[33] **(Supplementary Note 4)**. Temporal grouping was used by the top performers (including SMolPhot, MIATool[34] and SMAP-2018), and is visibly apparent as a more punctate super-resolved image (**Fig. 4A**).

Choice of PSF model

Most software used a variant of Gaussian PSF model. A few participants designed more accurate PSF models. Either diffraction theory was used (MIATool, LEAP) or spline fitting of an analytical function to the experimental PSF was adopted (CSpline, SMAP-2018). Although simple Gaussian model PSFs were sufficient to obtain best-in-class performance for the 2D and astigmatic modalities (ADCG, PeakFit, SMolPhot), top results for the more optically complex biplane and double helix modalities were exclusively software using non-Gaussian PSF models (SMAP-2018, CSpline, MIATool, LEAP).

Multi-algorithm packages

Several software packages take a Swiss army knife approach of integrating multiple optional localization algorithms into one program, to be flexible enough to suit various experimental conditions[19,27]. SMAP-2018 and ThunderSTORM achieved strong across-the-board performance supporting this rationale.

Software run time

Software run time is important both for ease of use and real time analysis. We did not observe correlation between software localization performance (Efficiency) and software run time (**Fig. S13A**). We thus created an alternative ranking metric, *Efficiency-Runtime*, which gave 25 % weighting to run time (**Supplementary Note 2.7, Fig S13B**). Many good performers in the efficiency-only ranking were relatively fast and thus retained good ranking (SMAP-2018, SMolPhot, 3D-DAOSTORM). Interestingly, two software packages highly optimized for speed gained top ranking in this analysis: pSMLM-3D[35] and QC-STORM.

Diagnostic tools for software and algorithm performance

During our analysis, we frequently noticed common types of deviation between software results and ground truth which were easily diagnosed by visual inspection (**Fig. S14, S15**). This included not only obvious issues of poor localization precision or spot averaging at high density, but also more subtle problems such as a common error of structural warping which significantly reduced software performance. On the competition website, we provide detailed diagnostic software reports including multiple examples of software performance on individual frames to help developers to identify algorithm and software limitations and maximize software performance (**Fig. S3, S16**).

Assessment on real STORM data

We investigated the performance of a representative subset of astigmatism software on real STORM datasets of well-characterized test structures, microtubules and nuclear pore complex, NPC (**Fig. 4B, S17**). This qualitative assessment was consistent with findings for simulated data. No performance difference between single and multi-emitter fitters was observed, which is not surprising since spot density in these datasets was low. Relatively poor software performance was immediately obvious from visual inspection (QuickPALM). Temporal grouping noticeably improved resolution (3D-DAOSTORM, CSpline, MIAtool, SMAP-2018). Interestingly, although Gaussian/ Bessel PSF modelling software (3D-DAOSTORM, MIATool, ThunderSTORM) gave high resolution images, software which explicitly modelled the non-ideal experimental PSF via spline fitting (CSpline, SMAP-2018) gave noticeably improved resolution of fine structural features such as the top and bottom of the NPC (**Fig. 4B**) or the hollow core of antibody-labelled microtubules (**Fig. S17**).

DISCUSSION

The strongest conclusion we draw from the 3D localization microscopy challenge is that choice of localization software greatly affects the quality of final super-resolution data, even at “easy” high SNR, low spot density conditions. Biplane performance was particularly dependent on software choice, with only one software (SMAP-2018) achieving near-Cramér-Rao lower bound performance. Double helix SMLM showed less sensitivity to choice of software than biplane, with astigmatic SMLM intermediate between the two. The best software in each modality performed close to the Cramér-Rao lower bounds over a wide focal range and successfully detected most molecules, even at low signal to noise. Average software in all three modalities was significantly worse, with the obtained axial resolution being particularly sensitive to software choice.The second major conclusion is that localization software that explicitly includes the experimental PSF in the fitting model gives a significant performance increase for 3D SMLM. For the more optically complex biplane and double helix modalities in particular, the best results were from software that incorporated non-Gaussian PSF models (SMAP-2018, CSpline, MIATool). This result also highlights the importance of accurate PSF modelling in 3D SMLM simulations. The performance advantage of experimental PSF fitting software would not have been observable had simulations been generated with a simple Gaussian PSF.

We can also make an overall comparison between 3D modalities, taking into account software performance. We stress that these comparisons apply to microscope PSFs similar to those tested here; for example, additional PSF engineering could improve results of any modality. Biplane imaging gave the best overall performance of any modality when used with best-in-class software (SMAP-2018), but performance depended surprisingly strongly on the software used. This requires further investigation; possibly it could be due to the inherent complexity of multi-channel imaging. Astigmatic imaging gave a good compromise of robustness and performance, particularly in combination with experimental PSF fitting software. For the model PSF used here, double helix imaging gave good results at high SNR and large Z-range, but performed poorly at low SNR or high emitter density. This is probably due to the large DH PSF used here; double helix designs with more compact PSF should reduce this issue[21].

Of the different algorithm classes, well-tuned single-emitter and multi-emitter fitting algorithms (each capable of dealing well with occasional molecule overlap) gave good results for low density 3D SMLM. We also found that several software packages for astigmatic or biplane imaging gave adequate performance for the challenging case of high molecule densities, as long as the image SNR was high. Current software packages gave poor performance when molecule density was high and image SNR was low. These results indicate that with current algorithms high density 3D SMLM performance is mediocre at high SNR and poor at low SNR. Surprisingly, multi-emitter fitting did not show significant improvement over well-tuned single emitter fitting for the 3D high-density datasets; this may indicate that potential for improvement remains in this category. Many software packages did not apply temporal grouping[33], resulting in reduced software performance. Since temporal grouping is a simple step for maximum precision, we urge all software developers to integrate this approach into their software as an optional final step in the localization process.

The second 2D localization microscopy challenge provided the opportunity to reassess the state of the field. The performance of best-in-class 2D software over a range of conditions, at both high and low spot density, was very strong. Interestingly, the top three performers in the 2D high density condition were all compressed sensing algorithms (ADCG, FALCON, SMfit). In low density 2D conditions, the best single-emitter, multi-emitter and compressed sensing algorithms all gave comparable, excellent, performance. We speculate that performance in the low spot density 2D category might now be near optimal levels.

We look forward to new competition submissions using approaches not yet represented in the software challenge. In addition to the elegant HAWK preprocessing technique[36], deep-learning-based SMLM algorithms show great promise[37–40], especially for modelling complex point spread functions[38] or analyzing high emitter density data[40]. However, caution is required about making direct comparisons between algorithms which use strong structural priors to increase performance[37], and algorithms which do not, as the latter may be more robust when presented with novel samples.

In future, we plan to extend the SMLM challenge into an open platform with a fully automated assessment process, and where new competition simulations and assessment metrics can easily be created and contributed by the community. It will be important to account for new technologies and developments in SMLM, such as scientific CMOS cameras[6], in future simulations. It would also be exciting to adapt the tools developed in the SMLM challenge to other classes of super-resolution microscopy, such as fluorescence-fluctuation-based super-resolution microscopies (*e.g.*, 3B[41], SOFI[42], SRRF[43]) and structured illumination microscopy[44].

The results of this competition show that the best 2D and 3D localization microscopy software have formidable algorithmic performance. However, a problem that often hinders adoption of new SMLM algorithms is that only a small subset of algorithms is packaged in, or compatible with fast, well-maintained, user-friendly software packages, which include all stages of the SMLM data analysis pipeline – analysis, visualization and quantification. This remains a key outstanding challenge for the field.

Both the 3D and 2D localization microscopy software challenges remain open and continuously updated on the competition website. This continuously evolving analysis of SMLM software performance provides software developers with a robust means of benchmarking new algorithms, and helps to ensure that super-resolution microscopists use software that gets the best out of their hard-won data.

ACKNOWLEDGEMENTS

Authors acknowledge the following funding sources: a Newcastle University Research Fellowship and a Wellcome Trust & Royal Society Sir Henry Dale Fellowship grant number 206670/Z/17/Z to SH; an European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, Grant Agreement no. 692726 to DS, TAP, MU; UK BBSRC grants BB/M022374/1, BB/P027431/1, BB/R000697/1 grant and MRC grants MC-UU-12018/2, MR/K015826/1 to RH; European Research Council (ERC) grant CoG-724489, CellStructure to JR; FranceBioImaging infrastructure ANR-10-INBS-04 to J.-B.S; National Institutes of Health grant 1R15GM128166-01 to GMH; and NSF SBIR grants 1353638, 1534745 to Double Helix LLC. We thank R. Piestun at University of Colorado for providing DH-PSF phase mask designs to Double Helix LLC. We thank all the localization microscopy challenge participants for their contribution: Hazen Babcock (3D-DAOSTORM, Cspline, L1H), Fabian Hauser (3D-STORM Tools), Shigeo Watanabe (3D-WTM,WTM), Nicholas Boyd (ADCG), Junhong Min, Kyong Jin and Jong Chul Ye (ALOHA, FALCON), Hervé Rouault (B-recs), Emmanuel Soubies (CEL0-STORM), Artur Speiser, Srinivas Turagas and Jakob Macke (DECODE), Alex von Diezmann, Camille Bayas and W. E. Moerner (Easy-DHPSF), Thomas Vomhof and Jochen Reichel (FIRESTORM), Hanjie Pan (LEAP), Ann Wheeler (Localizer), Zhen-li Huang and Yujie Wang (MaLiang), J. Chao, R. Velmurugan, A. V. Abraham and R. J. Ober (MIATool), Hendrik Deschout (mlePALM), Thomas Pengo (Octane, PeakSelector), Yi-na Wang (PALMER), Alex Herbert (PeakFit), Koen Martens and Johannes Hohlbein (pSMLM-3D), Luchang Li (QC-STORM), Ricardo Henriques (QuickPALM), G. Tamas and J. Sinko (RainSTORM), Steve Wolter and Markus Sauer (RapidSTORM), Manfred Kirchgessner and Frederik Gruell (SFP Estimator), Yiming Li and Jonas Ries (SMAP), Hayato Ikoma (SMfit), A. Loot, A. Valdmann, M. Eltermann, M. Kree and M. Pärs (SMolPhot), Yoon J. Jung, Anthony Barsic Rafael Piestun, and Nikta Fakhri (SOLAR_STORM), Anna Archetti (STORMChaser), Martin Ovesny, Guy Hagen and Pavel Krizek (ThunderSTORM), Jiaqing Huang (TVSTORM), Adel Kechkar, Corey Butler and Jean-Baptiste Sibarita (WaveTracer) and Benoît Lelandais (ZOLA-3D). We thank the SMLMS 2016 organizers (S. Manley and A. Radenovic, EPFL) for hosting a localization microscopy challenge special session. We also thank Double Helix LLC and Molecular Devices LLC for sponsoring the SMLMS 2016 special session. The sponsors had no input or influence on the research.

AUTHOR CONTRIBUTIONS

DS and SH conceived and coordinated the study. DS, SH, TAP, AAr, HB, SC, AW, GMH, RH, TL, TP, JBS designed the study. SH, AAg, RH, JBS collected experimental PSFs. DS, TAP, SH, TL wrote simulation code. BR shared unpublished software. DS generated simulated datasets. JR shared experimental STORM data. AH, JR, JC, RV provided feedback and quality control on simulations and analysis methods. TAP carried out the assessment of software performance. TAP, DS, SH analysed and interpreted the results. DS, HB, RO, BR, GMH, JBS, JR, RH, MU, SH directed research. SH, DS, TAP wrote the manuscript with feedback from all authors.

Editor's Summary

This study reports results from the second community-wide single molecule localization microscopy software challenge, which tested over thirty software packages on realistic simulated data for multiple popular 3D image acquisition modes as well as 2D localization microscopy.

REFERENCES

1. Betzig, E. *et al.* Imaging Intracellular Fluorescent Proteins at Nanometer Resolution. *Science* **313**, 1642–1645 (2006).

2. Hess, S. T., Girirajan, T. P. K. & Mason, M. D. Ultra-High Resolution Imaging by Fluorescence Photoactivation Localization Microscopy. *Biophys. J.* **91**, 4258–4272 (2006).

3. Rust, M. J., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat Methods* **3**, 793–795 (2006).
4. Holden, S. J., Uphoff, S. & Kapanidis, A. N. DAOSTORM: an algorithm for high- density super-resolution microscopy. *Nat Meth* **8**, 279–280 (2011).
5. Huang, F., Schwartz, S. L., Byars, J. M. & Lidke, K. A. Simultaneous multiple-emitter fitting for single molecule super-resolution imaging. *Biomed. Opt. Express* **2**, 1377–1393 (2011).
6. Huang, F. *et al.* Video-rate nanoscopy using sCMOS camera-specific single-molecule localization algorithms. *Nat. Methods* **10**, 653–658 (2013).
7. Sage, D. *et al.* Quantitative evaluation of software packages for single-molecule localization microscopy. *Nat. Methods* **12**, 717–724 (2015).
8. Huang, B., Jones, S. A., Brandenburg, B. & Zhuang, X. Whole-cell 3D STORM reveals interactions between cellular structures with nanometer-scale resolution. *Nat Meth* **5**, 1047–1052 (2008).
9. Shtengel, G. *et al.* Interferometric fluorescent super-resolution microscopy resolves 3D cellular ultrastructure. *Proc. Natl. Acad. Sci.* **106**, 3125–3130 (2009).
10. Huang, B., Wang, W., Bates, M. & Zhuang, X. Three-Dimensional Super-Resolution Imaging by Stochastic Optical Reconstruction Microscopy. *Science* **319**, 810–813 (2008).
11. Babcock, H., Sigal, Y. M. & Zhuang, X. A high-density 3D localization algorithm for stochastic optical reconstruction microscopy. *Opt. Nanoscopy* **1**, 1–10 (2012).
12. Ovesný, M., Křížek, P., Švindrych, Z. & Hagen, G. M. High density 3D localization microscopy using sparse support recovery. *Opt. Express* **22**, 31263–31276 (2014).
13. Min, J. *et al.* 3D high-density localization microscopy using hybrid astigmatic/ biplane imaging and sparse image reconstruction. *Biomed. Opt. Express* **5**, 3935–3948 (2014).
14. Zhang, S., Chen, D. & Niu, H. 3D localization of high particle density images using sparse recovery. *Appl. Opt.* **54**, 7859–7864 (2015).
15. Juette, M. F. *et al.* Three-dimensional sub–100 nm resolution fluorescence microscopy of thick samples. *Nat. Methods* **5**, 527–529 (2008).
16. Pavani, S. R. P. *et al.* Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function. *Proc. Natl. Acad. Sci.* **106**, 2995–2999 (2009).
17. Collaboration through competition. *Nat. Methods* **11**, 695 (2014).
18. Annibale, P., Vanni, S., Scarselli, M., Rothlisberger, U. & Radenovic, A. Quantitative Photo Activated Localization Microscopy: Unraveling the Effects of Photoblinking. *PLOS ONE* **6**, e22678 (2011).
19. Li, Y. *et al.* Real-time 3D single-molecule localization using experimental point spread functions. *Nat. Methods* (2018). doi:10.1038/nmeth.4661
20. Loot A. , Valdmann A., Eltermann M., Kree M., Pärs M. SMolPhot Software. Available at: https://bitbucket.org/ardiloot/. (Accessed: 28th January 2019)
21. Grover, G., DeLuca, K., Quirin, S., DeLuca, J. & Piestun, R. Super-resolution photon-efficient imaging by nanometric double-helix point spread function localization of emitters (SPINDLE). *Opt. Express* **20**, 26681–26695 (2012).
22. Babcock, H. P. & Zhuang, X. Analyzing Single Molecule Localization Microscopy Data Using Cubic Splines. *Sci. Rep.* **7**, 552 (2017).
23. Boyd, N., Schiebinger, G. & Recht, B. The Alternating Descent Conditional Gradient Method for Sparse Inverse Problems. *SIAM J. Optim.* **27**, 616–639 (2017).
24. Henriques, R. *et al.* QuickPALM: 3D real-time photoactivation nanoscopy image processing in ImageJ. *Nat Meth* **7**, 339–340 (2010).
25. Takeshima, T., Takahashi, T., Yamashita, J., Okada, Y. & Watanabe, S. A multi-emitter fitting algorithm for potential live cell super-resolution imaging over a wide range of molecular densities. *J. Microsc.* **271**, 266–281 (2018).

26. Kechkar, A., Nair, D., Heilemann, M., Choquet, D. & Sibarita, J.-B. Real-Time Analysis and Visualization for Single-Molecule Based Super-Resolution Microscopy. *PLOS ONE* **8**, e62918 (2013).
27. Ovesný, M., Křížek, P., Borkovec, J., Švindrych, Z. & Hagen, G. M. ThunderSTORM: a comprehensive ImageJ plug-in for PALM and STORM data analysis and super-resolution imaging. *Bioinformatics* **30**, 2389–2390 (2014).
28. Soubies, E., Blanc-Féraud, L. & Aubert, G. A Continuous Exact l0 Penalty (CEL0) for Least Squares Regularized Problem. *SIAM J. Imaging Sci.* **8**, 1607–1639 (2015).
29. Babcock, H. P., Moffitt, J. R., Cao, Y. & Zhuang, X. Fast compressed sensing analysis for super-resolution imaging using L1-homotopy. *Opt. Express* **21**, 28583–28596 (2013).
30. Min, J. *et al.* FALCON: fast and unbiased reconstruction of high-density super-resolution microscopy data. *Sci. Rep.* **4**, 4577 (2014).
31. Huang, J., Sun, M., Ma, J. & Chi, Y. Super-Resolution Image Reconstruction for High-Density Three-Dimensional Single-Molecule Microscopy. *IEEE Trans. Comput. Imaging* **3**, 763–773 (2017).
32. Pan, H., Simeoni, M., Hurley, P., Blu, T. & Vetterli, M. LEAP: Looking beyond pixels with continuous-space EstimAtion of Point sources. *Astron. Astrophys.* **608**, A136 (2017).
33. Durisic, N., Laparra-Cuervo, L., Sandoval-Álvarez, Á., Borbely, J. S. & Lakadamyali, M. Single-molecule evaluation of fluorescent protein photoactivation efficiency using an in vivo nanotemplate. *Nat. Methods* **11**, 156–162 (2014).
34. Chao, J., Ward, E. S. & Ober, R. J. A software framework for the analysis of complex microscopy image data. *IEEE Trans. Inf. Technol. Biomed. Publ. IEEE Eng. Med. Biol. Soc.* **14**, 1075–1087 (2010).
35. Martens, K. J. A., Bader, A. N., Baas, S., Rieger, B. & Hohlbein, J. Phasor based single-molecule localization microscopy in 3D (pSMLM-3D): An algorithm for MHz localization rates using standard CPUs. *J. Chem. Phys.* **148**, 123311 (2017).
36. Marsh, R. J. *et al.* Artifact-free high-density localization microscopy analysis. *Nat. Methods* **15**, 689 (2018).
37. Ouyang, W., Aristov, A., Lelek, M., Hao, X. & Zimmer, C. Deep learning massively accelerates super-resolution localization microscopy. *Nat. Biotechnol.* **36**, 460 (2018).
38. Zhang, P. *et al.* Analyzing complex single-molecule emission patterns with deep learning. *Nat. Methods* **15**, 913 (2018).
39. Boyd, N., Jonas, E., Babcock, H. P. & Recht, B. DeepLoco: Fast 3D Localization Microscopy Using Neural Networks. *bioRxiv* 267096 (2018). doi:10.1101/267096
40. Nehme, E., Weiss, L. E., Michaeli, T. & Shechtman, Y. Deep-STORM: super-resolution single-molecule microscopy by deep learning. *Optica* **5**, 458–464 (2018).
41. Cox, S. *et al.* Bayesian localization microscopy reveals nanoscale podosome dynamics. *Nat. Methods* **9**, 195–200 (2012).
42. Dertinger, T., Colyer, R., Iyer, G., Weiss, S. & Enderlein, J. Fast, background-free, 3D super-resolution optical fluctuation imaging (SOFI). *Proc. Natl. Acad. Sci.* **106**, 22287–22292 (2009).
43. Gustafsson, N. *et al.* Fast live-cell conventional fluorophore nanoscopy with ImageJ through super-resolution radial fluctuations. *Nat. Commun.* **7**, (2016).
44. Gustafsson, M. G. L. Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy. SHORT COMMUNICATION. *J. Microsc.* **198**, 82–87 (2000).

METHODS

1. CHALLENGE ORGANIZATION

We first ran the 3D SMLM software challenge as a time limited competition, with a results session hosted as a special session of the 6th Annual Single Molecule Localization Microscopy Symposium in August 2016. The competition has now been converted to a permanent software challenge accepting new submissions. Special thanks is due to the software SMAP and 3D-WTM[25] that participated in all eight categories (*density* x *modality*). The current list of participants is at:

http://bigwww.epfl.ch/smlm/challenge2016/index.html?p=participants

All datasets, methods, participations, and results of the challenge 2016 made available at http://bigwww.epfl.ch/smlm/challenge2016/. Software for simulation and analysis is hosted on the competition GitHub repository: https://github.com/SMLM-Challenge/Challenge2016/

A Life Sciences Reporting Summary is associated with this manuscript on the Nature Methods website.

2. LOCALIZATION MICROSCOPY SIMULATIONS

2.1. Structure, noise levels and spot densities

Structure. The synthetic datasets were designed to be similar to images derived from real cellular structures . We defined mathematical models for cellular structures that imitate cytoskeletal filaments such as microtubules and larger tubular structures such as the endoplasmic reticulum or mitochondria (**Fig. S18A**). These structures have a tubular shape in the 3D space. For the 3D competition, we simulated synthetic 25 nm diameter microtubules (**Fig. 1**). Psuedo-microtubules are defined with their central axis elongating in a 3D space having an average outer diameter of 25 nm with an inner, hollow tube of 15 nm diameter. For the 2D competition, in addition to synthetic microtubules (MT), we simulated larger diameter 150 nm cylinders, called pseudo-endoplasmic reticulum (pseudo-ER), designed to approximate larger cellular structures such as mitochondria and the endoplasmic reticulum (ER) (**Fig. 1**).

The underlying sample structure is formalized in a continuous space which allows rendering of digital images at any scale, from very high resolution (up to 1 nm/pixel) to low resolution (camera resolution: 100 nm/ pixel). The continuous-domain 3D curve is represented by means of a polynomial spline. The sample is imaged in a 6.4 × 6.4 μm^2 field of view, and the center lines of the microtubules have limited variation along the z (vertical) axis, *i.e.*, less than 1.5 µm. The fluorescent markers are uniform randomly distributed over the structure according to the required density. The photon emission rate of each fluorophore is controlled by a photo-activation model (see below). The exact locations of all fluorophores are stored at high precision floating-point numbers expressed in nanometers. This ground-truth file is used for conducting objective evaluations without human bias.

Noise levels. We generated data at three different signal-to-noise ratio (SNR) levels, based on real signal to noise levels encountered under common SMLM experimental scenarios: *N1*, fixed cells antibody labelled with organic dye[10], high signal, medium background; *N2,* fluorescent protein labelling[1], low signal, low background; and *N3,* live cell affinity dye labelling[45,46], high signal, high background.

Spot density. As performance at different density of active emitters is a key challenge for SMLM software, we generated 3D competition datasets at both sparse emitter density (0.25 mol. [molecule] μm^{-2}), *3D LD* and high emitter density (2.5 mol. μm^{-2}), *3D HD*. For the 2D competition, we generated a sparse (0.5 mol. μm^{-2}), *2D LD*, and very high density dataset (5 mol. μm^{-2}), *2D HD*.

Together, these simulated conditions closely resemble experimental 3D and 2D data under a range of challenging conditions of SNR, spot density, axial thickness and structure summarized in **Supplementary Table 3**. In addition, we provide simulated z-stacks of bright beads for software calibration. The competition datasets (**Supplementary Table 4**) are available online on the competition website.

2.2. Photophysics activation model

We incorporated a 4-state model of fluorophore photophysics[18], including a transient dark state (dye blinking) and a bleaching pathway (**Fig. S18C**). Given a list of source locations from the structure simulator, fluorophore blinking was simulated by a 4-states Markov chain model. The states are ON, OFF, BLEACH, DARK and the transitions are Poisson distributed (**Fig. S18C**), except for the OFF to ON transitions which follow a uniform random distribution to reflect that in typical experimental conditions, constant imaging density is maintained by tuning the photoactivation rate during the experiment. All switching is calculated at sub-frame resolution and then total fluorophore on-time was integrated over each frame.

Due to two decay paths, the actual mean lifetime of the state ON is

$$T_{LIFETIME} = \frac{1}{\frac{1}{T_{ON}} + \frac{1}{T_{BLEACH}}}$$

Switching rates were chosen to approximate photoactivatable fluorescent proteins T_{ON}=3 frames, T_{DARK}=2.5 frames, and T_{BLEACH}=1.5 frames.

Fractional fluorophore ON-times per frame (between 0 and 1) were multiplied by the mean flux of photon emission. The flux of photons expressed in photons/seconds was given by the relation

$$\boldsymbol{F} = \frac{\emptyset\, P\, \sigma}{e}$$

Φ is the quantum yield of the dye, *P* is power of the laser in W/cm^2, e = h c / λ is the energy of one photon, σ = 1000 ln(10) ε / N_A is the absorption cross section in cm^2 and ε is the molar extinction coefficient (EC) or absorptivity in cm^2/mol which is a characteristic of a given fluorophore. The laser power was Gaussian distributed over the field of view. At the end of this process a list of XY positions, on-frames and (noise-free) intensities for all activated fluorophores was obtained.

Analysis of the resulting simulated photon counting distribution is presented in **Supplementary Note 5** and **Figure S23**.

2.3. Experimental Point Spread Function

Model PSFs, stored as high resolution look up tables, were derived from experimentally measured PSFs. Although the algorithmic approach is distinct, the concept of accurately modelling the experimental PSF based on calibration data bears relation to the PSF phase retrieval approach previously employed by Hanser and coworkers[47].

Images of fluorescent beads were recorded for each modality (**Supplementary Table 5**). Signal to noise ratio of recorded PSFs was maximized in all cases by maximizing exposure time and averaging over several frames to increase dynamic range.

To acquire experimental PSFs, we took 100 nm Tetraspek beads (Invitrogen) adsorbed to #1.5 (170 µm thick) coverglass, imaged in water. The excitation wavelength was between 640 nm and 647 nm, and a Cy5 emission filter was used. Data acquisition parameters for each modality are listed in **Supplementary Table 5**.

The experimental PSFs used to generate the simulated data are available on the competition website. As the goal of this study was to compare software obtained on typical SMLM microscopes, we deliberately chose PSFs representative of common implementations of each 3D modality. However, additional PSF engineering should improve results of any specific modality, for example adaptive-optics corrected astigmatism[48], or reduced Z-range, higher SNR DH-PSF designs[21].

The experimental point spread functions used here were measured for fluorescent beads adsorbed to the microscope cover slip, and should be appropriate simulations of SMLM data acquired within a few microns of the cover slip. Performing SMLM imaging at greater depths, *e.g.,* in tissue or even deep within single cells, with oil immersion objectives will cause spherical aberration due to refractive index mismatch[49]. In order to accurately simulate SMLM data acquired at depth, the experimental PSFs could be acquired at a matching depth, by embedding fluorescent beads in agarose. Alternatively, the PSF for beads at the coverslip could be measured and explicitly calculated via phase retrieval, and then convolved with the appropriate degree of spherical aberration[49].

2.4. Simulation PSF construction

For each modality, 3-6 beads were selected within a small (< 32 µm) region, to minimize PSF variation due to spherical aberration. Images for each selected bead were interpolated in XY to a pixel size of 10 nm. Beads were then coaligned by cross-correlation on the in-focus frame. Coaligned beads were averaged in XY to minimize pixel quantization artefacts and to increase SNR. Where necessary, Z-stacks were interpolated to a Z-step size of 10 nm. A central Z-range of 1.5 µm was selected that represents 151 optical planes with a Z-step of 10 nm. The Z-range covers -750 nm to +750 nm. The plane of best focus was chosen as the simulation 0 nm plane. Each model PSF was normalized such that the total intensity of the PSF in the in-focus frame within a diameter of 3 FWHM from the PSF center was equal to 1.

For the DH PSF, the transmission of the combined phase mask system was measured as 96 %, which was approximated as 100 % brightness relative to the 2D and astigmatic PSFs.

In biplane super-resolution microscopy, emitted fluorescence is split into two simultaneously imaged channels, with a small (500-1000 nm) defocus introduced between the two channels[15]. As the small defocus should introduce minimal additional aberration into an optical system, we semi-synthetically constructed a realistic biplane PSF from the experimental 2D PSF. The two defocused PSFs were constructed by duplicating the 2D PSF and offsetting it by -250 nm and 250 nm for each Z-plane.

This yielded five high SNR model PSFs with an isotropic voxel size of 10x10x10 nm^3.

The ground truth XY=0 was defined as the image center of mass of the in-focus frame of the model PSF, and Z=0 was defined as the in-focus frame. Accounts for shifts in the fitted XY center of the model PSF by localization software due to systematic offsets and Z-dependent variation of the model PSF center of mass are dealt with below (wobble correction).

2.5. Noise model

A constant mean autofluorescent background was added to the noise-free simulated images, and these images were then fed through the noise model representing Poisson distributed fluorescence emission recorded on a high quantum efficiency back-illuminated EMCCD[50,51].

The proposed noise model assumed as main contributions to the stochastic noise:

- σ_S , the shot noise produced by the fluorescence background and signal and the spurious charge. Shot noise can be derived from the second moment of the Poisson distribution
- σ_R, the read noise of EMCCD camera, which is described by second moment of the Gaussian distribution

- σ_{EM}, the electron multiplication noise introduced by the gain process, which is described by the second moment of the Gamma distribution[51].

We assumed as camera parameters the ones specified for the Photometrics Evolve Delta 512 EMCCD camera (values for other manufacturer's EMCCDs are similar):

- QE = 0.9, Evolve quantum efficiency at 700 nm absorption wavelength.
- σ_R= 74.4 electrons, manufacturer measured root mean square noise for Evolve 512 camera
- c = 0.002 electrons, manufacturer quoted spurious charge (clock induced charge only, dark counts negligible)
- EM_{gain} = 300
- e_{adu} = 45 electron per analog to digital unit (ADU), analog to digital conversion factor
- G = 0.9*300/45 = 6, total system gain
- BL = 100 ADU

The final simulated photon electrons will thus be given by:

$$n_{ie} = \mathcal{P}(QE \cdot n_{photIn} + c)$$

$$n_{oe} = \Gamma\left(n_{ie}, EM_{gain}\right) + \mathcal{G}(0, \sigma_R)$$

which leads to the final pixel counts:

$$ADU_{out} = min\left(\frac{n_{oe} - n_{oe} mod\ e_{ADU}}{e_{per_{adu}}} + BL, 65535\right)$$

2.6. Depth-dependent lateral distortion/ wobble

As the PSF models are experimentally derived, the 3D estimated localizations exhibit a depth-dependent lateral distortion, here called *wobble*. This optical distortion is due to a combination of a systematic offset (arbitrary definition of PSF center) and optical aberrations[52]. In order to compare estimated and true localizations, we correct this effect during the assessment (**Methods 3.1**).

2.7 Comparison of software results between different modalities.

The intensities of the PSF in each imaging modality were normalized to facilitate comparison of results between different modalities. Software results between 2D, 3D AS and 3D DH modalities are expected to be directly comparable.

For the biplane model PSF, as the emitted fluorescence is split into two channels, the intensity in each of the two simulated biplane channels was additionally reduced by 50 %. We note that a simulation bug meant that the fluorescence background was not reduced by 50 % as intended, leading to artificially high background for the biplane simulation. *I.e.*, the background in each of the two biplane channels is the same as in the single channel of the other modalities. However, due to the low background level in the 3D simulations, the effect on image SNR and thus localization error is small (see **Fig. S5, S6**), less than 5 nm near the plane of focus. Therefore, as long as the small drop in image SNR is taken into account, approximate comparisons of the biplane data to the other modalities can still be made.

3. SOFTWARE ASSESSMENT

3.1 Protocol

Each localization file submitted by the participants was manually checked for erroneous systematic errors in the definition of the dataset coordinate system, such as offsets, XY axis flips or clear scaling errors. Datasets were then programmatically standardized into a consistent output format. All modifications are publicly available. If required, the modifications consisted of columns reordering, reversing axes, XY axis swap, and shifting the lateral positions by a half camera pixel.

The assessment pipeline includes three main parts: localization processing, the pairing between true and estimated localization and the metrics calculations. The first one depends on the assessment settings. There are two switchable properties: photon thresholding and wobble correction. Their combinations yield four different assessment settings. Up to 64 assessment runs per software were possible (*i.e.,* 4 modalities, 4 datasets per modality). For any setting, we excluded the fluorophores within a lateral distance of 450 nm from the border. This value corresponds to the radius of the largest PSF, *i.e.,* Double Helix. The activations too close from the border are more difficult to localize and could bias the results.

The pairing between true and estimated localizations was performed frame by frame. For every frame, we identified the localizations that are close enough to a ground-truth position as true-positives (TP), the spurious localizations as false-positives (FP) and the undetected molecules as false-negatives (FN). The procedure matches two sets of localizations. We deployed the presorted nearest-neighbor search for its efficiency, with a linking threshold of 250 nm. The results are effectively similar to the computationally intensive Hungarian algorithm[7].

Photon thresholding

A photon threshold was required primarily due to the use of a realistic fluorophore blinking model. Since a fluorophore could activate/ bleach at any point in a simulated frame, this led to many frames containing very dim, undetectable localizations, *e.g.,* where a molecule had been active for one or more frames previously, and then bleached during the first 5 % of a frame. These fractional localizations should also be present but practically undetectable in an experimental dataset.

We decided to focus the software analysis on the localizations where the molecule was active for the majority of a frame, to be consistent with experimental expectations. Therefore, we implemented a photon threshold means where we kept the 75% brightest ground truth fluorophore activations. Because this was performed *after* the pairing step, observed localizations that were paired to discarded ground truth activations were also removed from the metric calculations.

Wobble correction

The centroid of experimental point spread functions shifts laterally by as much as 50 nm, as a function of axial position[10,52]. This is most often ignored by localization software, and instead corrected post-hoc by reference to a calibration curve[37]. Since our simulated PSF is experimentally derived, it was necessary to correct for these artefactual shifts between the observed localizations and ground truth, as part of the assessment process. This correction was performed using calibration data uploaded by competitors, similar to the correction typically performed on experimental data[52].

Three scenarios were proposed to the participants: no correction was applied during the assessment; the correction was based on a file provided by the participant itself or the correction was calculated by ourselves. The latter nevertheless requires the participant to localize a stack of beads we provided. Since the true positions of the beads are known, the difference between the estimated and true positions could be calculated and averaged. It thus yields the values for wobble correction.

In certain specific cases (identified on the competition website), at the request of authors, we did not apply this correction, for example because the software explicitly considered the whole 3D PSF during fitting and was thus immune to this lateral shift artefact. For accurate results, application of lateral shift correction is critical for analysis of localization microscopy simulations using experimentally derived PSFs, as can be seen by comparison of typical software results with and without wobble correction (**Fig. S19**).

3.2 Metrics

We calculated a large number of analysis metrics to quantify the performance of software relative to ground truth. These are discussed in detail in **Supplementary Note 2**. The metrics are split into two categories: localization based and image based metrics.

Localization based metrics. This directly relies on the localizations positions and notably includes the Recall, the Precision, the Jaccard Index, the RMSE (axial and lateral) and the consolidated Z-range. For the calculation of average software performance (**Fig. 3D-F, S10**) outlier software with an efficiency less than *eff=0 (eff=-30 for 3D high density dataset)* were excluded from the measurement. The key metrics of assessment were:

1. *Root mean squared localization error (RMSE).* The foremost consideration for localization software is how accurately it finds the position of labelled molecules. This was quantified as the root mean squared difference between the measured molecule position, x_i^s, and the ground truth position, x_i^t, in both the lateral (XY) and axial (Z) dimensions.

   RMSE lateral (RMSE Lateral) [nm]: $\sqrt{\frac{1}{\mathrm{TP}}\sum_{i\in \mathrm{S}\cap\mathrm{T}} (x_i^s - x_i^t)^2 + (y_i^s - y_i^t)^2}$.

   RMSE axial (RMSE Axial) [nm]: $\sqrt{\frac{1}{\mathrm{TP}}\sum_{i\in \mathrm{S}\cap\mathrm{T}} (z_i^s - z_i^t)^2}$.

2. *Jaccard index (JAC, %).* In addition to localization precision, SMLM image resolution depends critically on number of localized molecules[53], so it is crucial for SMLM software to accurately detect a large fraction of molecules in a dataset, and minimize false localizations. For every frame, we identified the localizations that are close enough to a ground-truth position as true-positives (TP), the spurious localizations as false-positives (FP) and the undetected molecules as false-negatives (FN). We then computed the *Jaccard index* (JAC, %), which measures the fraction of correctly detected molecules in a dataset,

$$JAC = 100\frac{TP}{TP+FP+FN}$$

3. *Efficiency (E).* For ranking purposes, we developed a single summary statistic for overall evaluation of software performance, which we term the *efficiency (E)*, encapsulating both the software's ability to find molecules, measured by the Jaccard index, and the software's ability to precisely localize molecules.

$$E = 100 - \sqrt{(100-JAC)^2 + \propto^2 \ RMSE^2}$$

   The trade-off between these two metrics is controlled by a parameter α. In a retrospective analysis, we chose $\alpha = 1\ \mathrm{nm}^{-1}$ for the lateral efficiency E_{lat}, $\alpha = 0.5\ \mathrm{nm}^{-1}$ for the axial efficiency E_{ax}, based on the linear regression slope between the localization errors and Jaccard index (**Fig. S20J-K**). Using this definition, an average software performance has an efficiency in the range 25-75, a perfect software would have the maximum efficiency of 100. Overall 3D efficiency was calculated as the average of lateral and axial efficiencies. Overall software rankings (**Fig. 2**) were calculated as the sum of rankings for high and low SNR datasets.

Image based metrics. The image based metrics are computed from a rendered image and includes the Signal-to-Noise Ratio (SNR) and the Fourier Ring / Shell Correlation (FRC/FSC). To render the image, we added the contribution of each localized molecule at the corresponding pixels. A contribution takes the form of a 3D additive Gaussian with a Full-Width Half Maximum (FWHM) of 20 nm. A complete list of all computed metrics is presented in the **Supplementary Note 2**.

We also calculated localization based metric results as a function of axial position. We proceeded by considering a subset of activations lying within an interval of axial positions (*i.e.,* from the true localizations). Then, most of the metrics (*e.g.,* Recall) are locally computed. This yields a curve providing information on the depth performance of each software / modality.

In order to summarize software axial performance, we analyzed how the recall varied as a function of Z. A typical recall versus axial position curve (**Fig. S4**) will drop at positions far from the focal plane, *i.e.,* where software can no longer detect spots to defocus. We first smoothed the curve using a sliding window. Then we computed the software Z-range, defined as the full width half maximal Recall of the smoothed curve (**Fig. S21**). This quantity is visually intuitive and useful for discussion of the recall performance if considered alongside a plot of recall vs axial position. However, because FHWM recall depends on the maximal recall, ranking based on this procedure would promote a software which poorly performed everywhere (*i.e.,* flat curve), whereas a software which performed well in the focal plane but less well outside would obtain a worse FWHM recall. This observation leads us to produce a so-called consolidated Z-range, by multiplying the Z-range value by the maximal Recall, which should provide a robust metric that avoids the previous case scenario.

Principal component analysis. In order to analyse the relationship between analysis metrics we computed the covariance matrix between each metric (**Fig. S22A**) and the principal component analysis (PCA) on the metrics (**Fig. S22B-D**). Each metric was standardized before applying the covariance and the PCA. For convenience, we took the additive inverse of the metrics for which lower values are best (*i.e.,* FP, FN, RMSE, FRC, FSC).

Summary statistics and detailed results for each software are available on the competition website (http://bigwww.epfl.ch/smlm/challenge2016/index.html?p=results), which also includes a tool for side-by-side comparison of the results of multiple software packages

3.3 Baseline Localization Software

We developed a minimalist Java tool software that performs localizations of bright emitters on the 4 modalities of the challenge 2016: 2D, Astigmatism, Double-Helix, and Biplane. This SMLM_BaselineLocalization software is only designed to establish the performance baseline for the SMLM challenge. It has intentionally limited lines of code and relies only on few threshold parameters to localize particles. It has basic calibration tool that has to run on a z-stack of beads to find the linear f(x) relation between the axial position Z and the shape of the bead.

- Astigmatism: $Z = f(W_X - W_Y)$, where W_X and W_Y are respectively an estimation of the size in X and Y.
- Double-Helix: $Z = f(\theta)$, where θ is the angle formed the pairing of two close points.
- Biplane: $Z = f(W_{left} - W_{right})$, where W_{left} and W_{right} are respectively an estimation of the size of the spots in left and the right plane.

The Java code is available: https://github.com/SMLM-Challenge/Challenge2016

4 REAL DATA ASSESSMENT

Astigmatism software was tested on previously published real 3D STORM datasets of microtubules and nuclear pore complex[19]. The tubulin dataset corresponds to the raw data for **Fig. S6** in Ref [19], and the nuclear pore complex dataset corresponds to raw data for **Fig. S9** in Ref [19]. Key acquisition parameters for data analysis are summarized on the competition website.

Data were analyzed by software authors or expert users, and submitted via the competition website. All data were drift corrected via cross-correlation. STORM images were rendered with a constant Gaussian blur with 3 nm standard deviation and saturated by 0.1 – 0.5 %. The complete scripts used for assessment and image rendering are available on the competition GitHub page.

5 DATA AVAILABILITY

5.1 Data availability statement

Simulated competition datasets are available at http://bigwww.epfl.ch/smlm/challenge2016/, together with the parameters used to generate the data. The ground truth list of simulated molecule positions for each competition dataset remains secret in order to allow the software challenge to remain continuously open to new submissions. However, ground truth data are available for the simulated training datasets.

Raw data for this study are uploaded on the Nature Methods website. The data corresponding to specific figures are listed with the Supplementary information.

5.2 Software availability statement

All software is available at https://github.com/SMLM-Challenge/Challenge2016

REFERENCES, ONLINE METHODS

45. Carlini, L. & Manley, S. Live Intracellular Super-Resolution Imaging Using Site-Specific Stains. ACS Chem. Biol. 8, 2643–2648 (2013).
46. Shim, S.-H. et al. Super-resolution fluorescence imaging of organelles in live cells with photoswitchable membrane probes. Proc. Natl. Acad. Sci. 109, 13978–13983 (2012).
47. Hanser B. M., Gustafsson M. G. L., Agard D. A. & Sedat J. W. Phase-retrieved pupil functions in wide-field fluorescence microscopy. J. Microsc. 216, 32–48 (2004).
48. Izeddin, I. et al. PSF shaping using adaptive optics for three-dimensional single-molecule super-resolution imaging and tracking. Opt. Express 20, 4957–4967 (2012).
49. McGorty, R., Schnitzbauer, J., Zhang, W. & Huang, B. Correction of depth-dependent aberrations in 3D single-molecule localization and super-resolution microscopy. Opt. Lett. 39, 275–278 (2014).
50. Hirsch, M., Wareham, R. J., Martin-Fernandez, M. L., Hobson, M. P. & Rolfe, D. J. A Stochastic Model for Electron Multiplication Charge-Coupled Devices – From Theory to Practice. PLOS ONE 8, e53671 (2013).
51. Basden, A. G., Haniff, C. A. & Mackay, C. D. Photon counting strategies with low-light-level CCDs. Mon. Not. R. Astron. Soc. 345, 985–991 (2003).
52. Carlini, L., Holden, S. J., Douglass, K. M. & Manley, S. Correction of a Depth-Dependent Lateral Distortion in 3D Super-Resolution Imaging. PLoS ONE 10, e0142949 (2015).
53. Baddeley, D. & Bewersdorf, J. Biological Insight from Super-Resolution Microscopy: What We Can Learn from Localization-Based Images. Annu. Rev. Biochem. 87, 965–989 (2018).

FIGURES

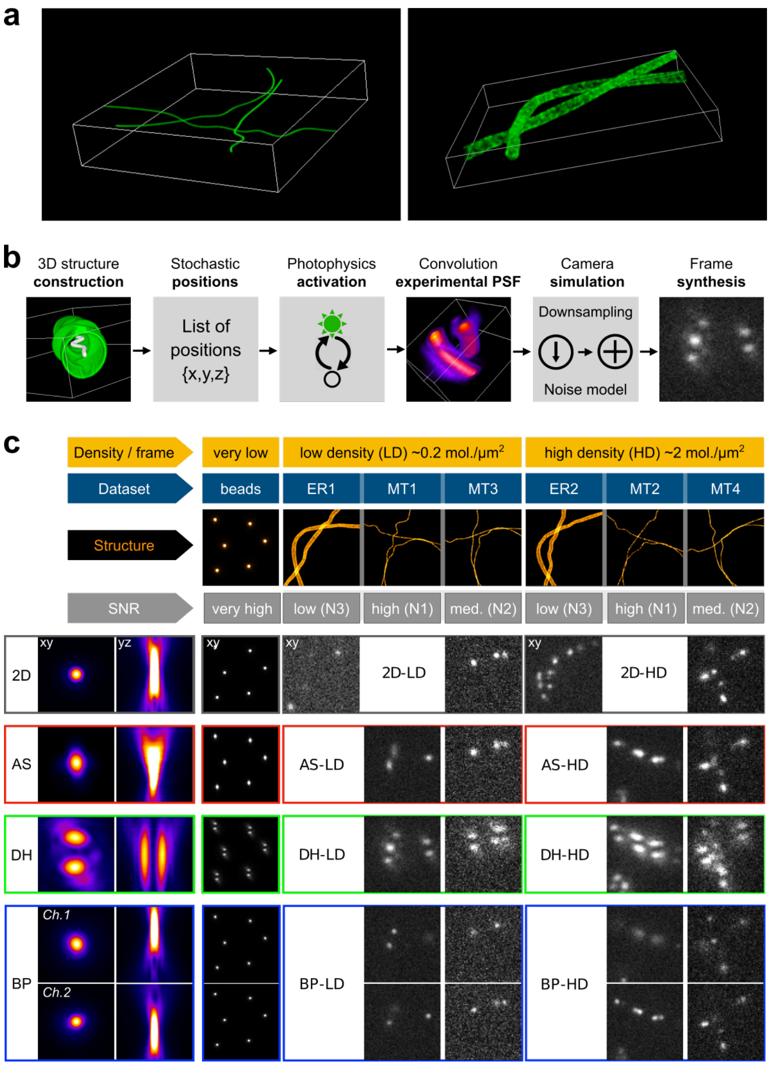

Figure 1: *Summary of SMLM challenge simulations*. ***A***. 3D rendering of simulated microtubules and endoplasmic reticulum samples. ***B***. *Key simulation steps.* The structure is constructed from 3D tubes continuously defined by three B-spline functions in the volume of interest. Membranes of the tubes are densely populated with possible positions. Fluorophores follow a 4-state photophysics model. Activations of a given frame are convolved with the experimental PSF and shot & camera noise is added. ***C.*** Summary of all 16 challenge datasets, calibration data and experimental PSFs. Left column: orthogonal projections of the experimentally-derived PSF. Right column: exemplar frame for each competition dataset, characterized by structure (endoplasmic reticulum, E; microtubules, MT), modality (2D; astigmatism, AS; double helix, DH; biplane, BP), density (low density, LD; high density, HD) and SNR (noise level N1, N2, N3). *BP Ch. 1,2*, indicates two biplane channels with a relative focal shift of 500 nm.

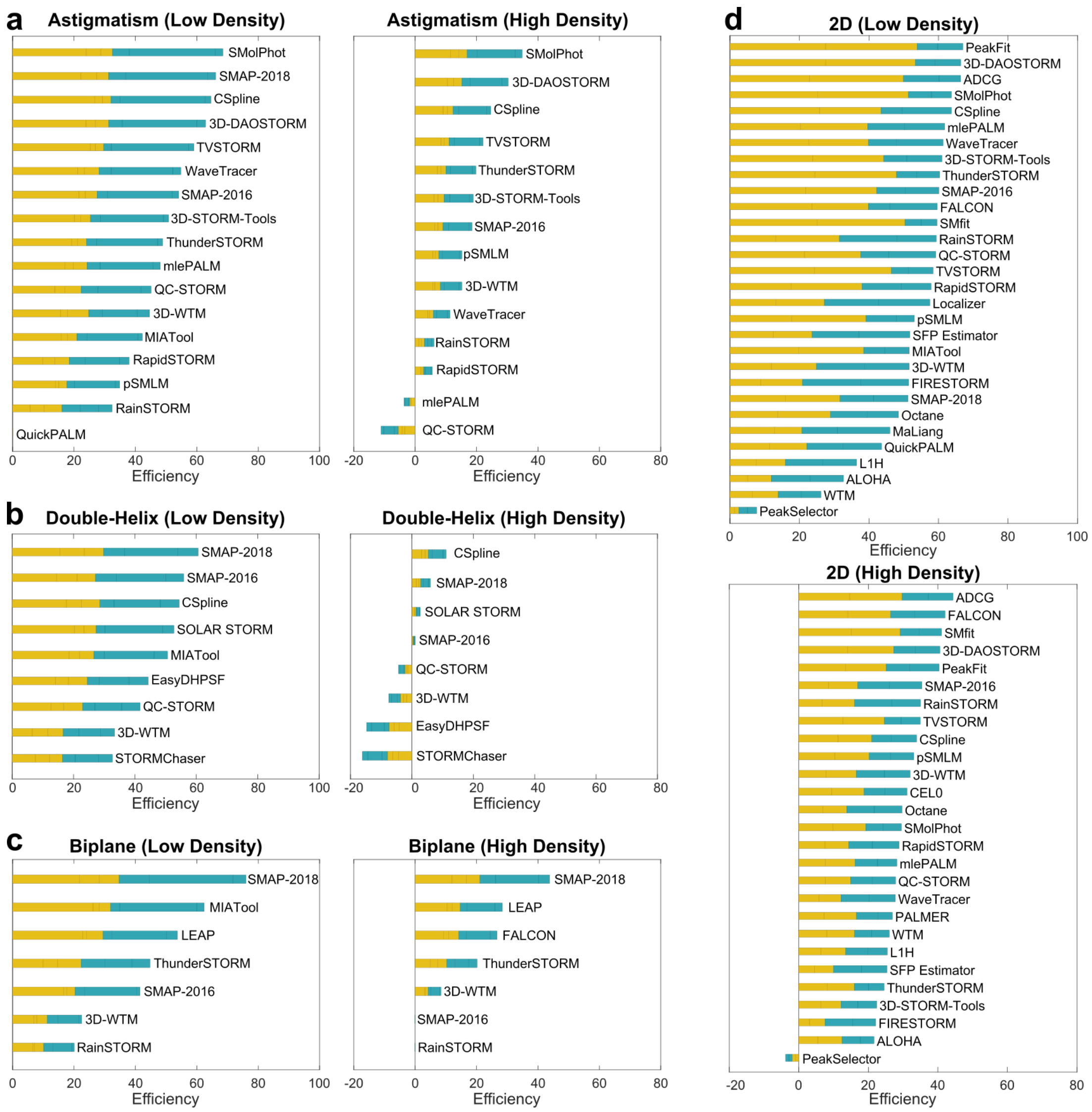

Figure 2: *Leaderboards for each competition modality, at low and high spot density*. Ranking is based on software Efficiency, which combines Jaccard index (fraction of successfully detected molecules) and localization precision (RMSE, root mean square error, lateral & axial). Orange, contribution of high SNR dataset; blue, contribution of low SNR dataset.

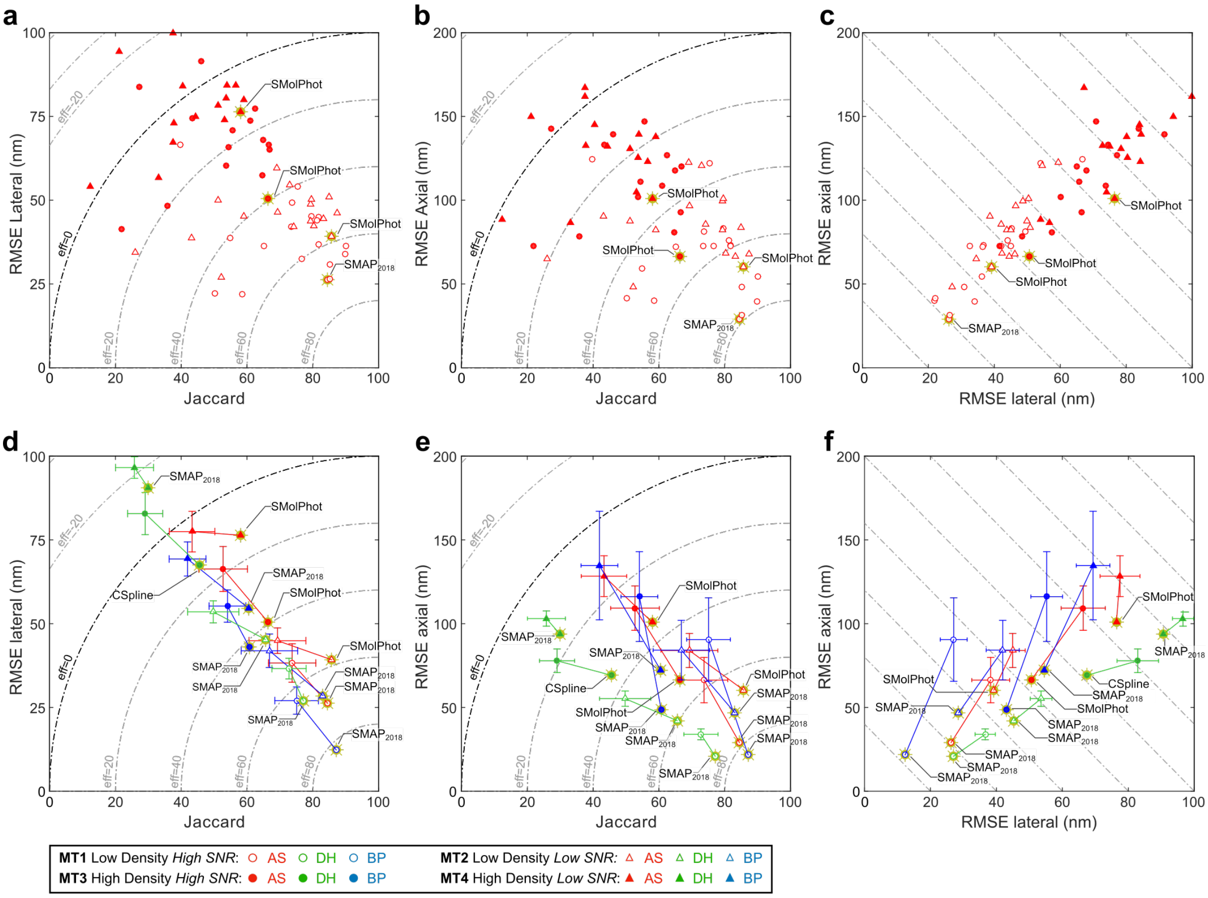

Figure *3*: *Comparison of 3D software performance*. Gold stars indicate top performers for each dataset. Dashed lines in top, middle panels indicate overall efficiency (higher is better). ***A-C.*** Localization error and spot detection performance of all astigmatic SMLM software. ***D-E.*** Average (colored marker with *s.d.* error bars, sample sizes for each category indicated in **Supplementary Table 2**) and best-in-class (colored marker with gold star) software performance for all competition modalities. *AS, astigmatism; DH, double helix; BP, biplane.*

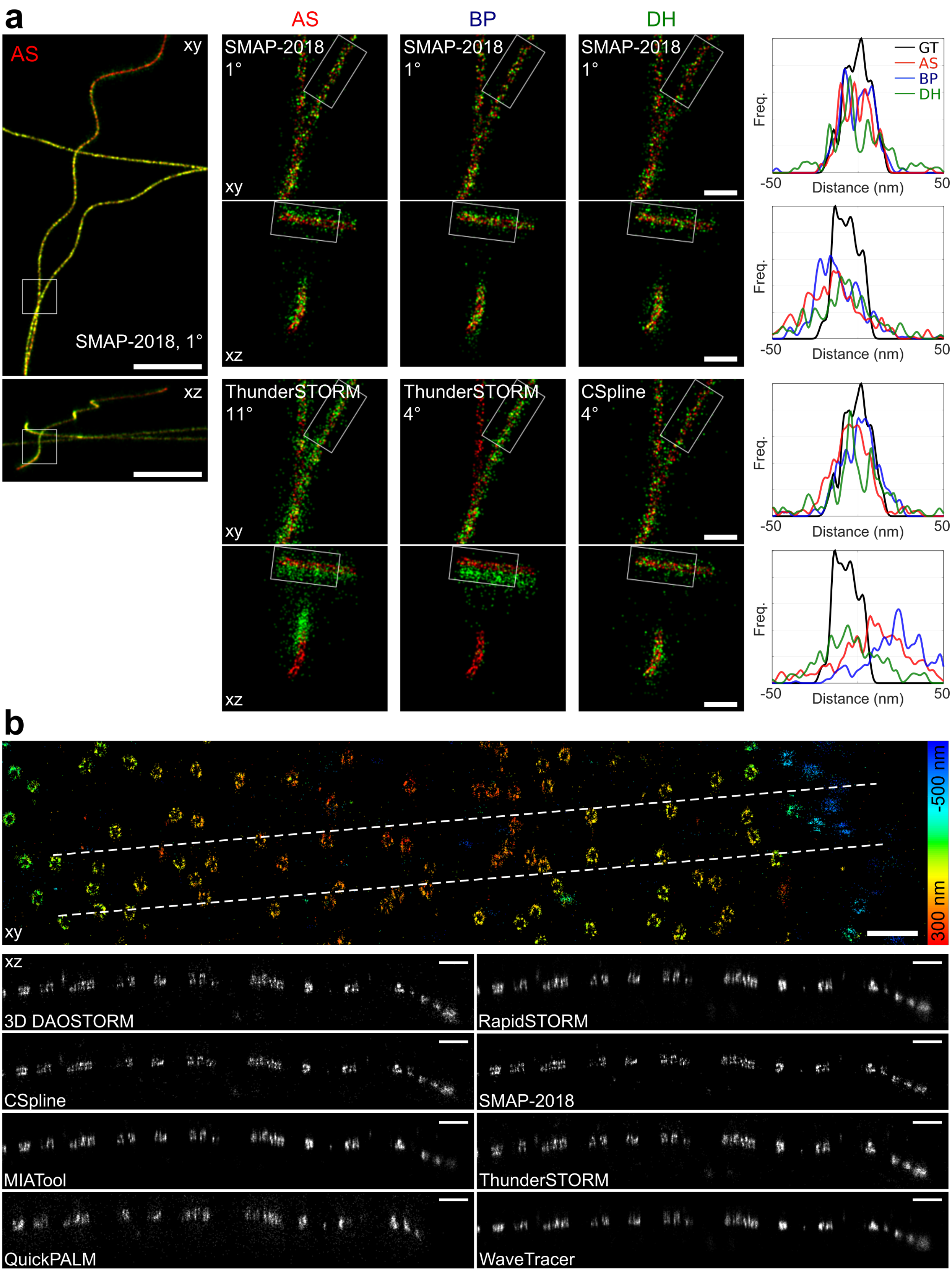

Figure 4: *Super-resolved images of software results for simulated and real competition datasets.* ***A.*** *Xy and xz projection images of 3D competition datasets for representative software. Top:* best-in-class software in each modality, for high SNR low density dataset. *Bottom*: representative average software. *Left*: *xy* and *xz* overview images for winning AS software. *Middle*: *xy* and *xz* zoom images of boxed regions in left panel, for winning and mid-range software, each modality. *Right*: *xy* and *xz* line profiles of winning and mid-range software for each modality, for boxed regions in middle panel. *Image colors*:

red, ground truth; green, software results. *Line profiles*: GT, ground truth, black; AS, astigmatism, red; BP, biplane, blue; DH, double helix, green. *Panel key:* Software-name Dataset-ranking°. *Scale bar*: full image, 1 µm, magnified regions, 100 nm. ***B**. Astigmatism software results for real nuclear pore complex 3D STORM data. Top:* Super-resolved overview image in *xy* for 3D-DAOSTORM software, color coded for depth. *Bottom: xz* orthoslices along 600 nm wide dashed region indicated in top panel for 8 astigmatism software packages. *Scale bar*s, 500 nm.