nature methods

ANALYSIS

https://doi.org/10.1038/s41592-019-0364-4

Corrected: Publisher Correction

Super-resolution fight club: assessment of 2D and 3D single-molecule localization microscopy software

Daniel Sage[1,22*], Thanh-An Pham[1,22], Hazen Babcock[2], Tomas Lukes[3,4], Thomas Pengo[5], Jerry Chao[6,7], Ramraj Velmurugan[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] and Seamus Holden[21,22*]

With the widespread uptake of two-dimensional (2D) and three-dimensional (3D) single-molecule localization microscopy (SMLM), 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 characterize and rank the performance of 2D and 3D SMLM 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 SMLM software and provides a holistic view of how the latest 2D and 3D SMLM packages 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 the current state of the art, and provides insight into the current limits of the field.

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 the imaging time of SMLM for compatibility with live-cell imaging[4–6]. If SMLM is to achieve a resolving power approaching that of electron microscopy, the analysis software used needs to be robust, be accurate, and perform at current algorithmic limits. This can be achieved only through rigorous quantification of SMLM software performance.

The first localization-microscopy software challenge was carried out in 2013 to benchmark two-dimensional (2D) SMLM software[7]. But biology is not just a 2D problem, and a key focus of localization microscopy is three-dimensional (3D) imaging of nanoscale cellular processes[8,9]. Three-dimensional 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 the challenging nature of 3D SMLM image processing, the performance of software for 3D SMLM has previously been assessed only for two or three 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 softwares 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 PSFs,

[1]Biomedical Imaging Group, School of Engineering, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. [2]Harvard Center for Advanced Imaging, Harvard University, Cambridge, MA, USA. [3]Laboratory of Nanoscale Biology and Laboratoire d'Optique Biomédicale, STI - IBI, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. [4]Department of Radioelectronics, FEE, Czech Technical University, Prague, Czech Republic. [5]University of Minnesota Informatics Institute, University of Minnesota Twin Cities, Minneapolis, MN, USA. [6]Department of Biomedical Engineering, Texas A&M University, College Station, TX, USA. [7]Department of Molecular and Cellular Medicine, Texas A&M University Health Science Center, College Station, TX, USA. [8]Department of Microbial Pathogenesis and Immunology, Texas A&M University Health Science Center, Bryan, TX, USA. [9]MRC Genome Damage and Stability Centre, School of Life Sciences, University of Sussex, Brighton, UK. [10]Double Helix LLC, Boulder, CO, 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, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. [14]Department of Imaging Physics, Delft University of Technology, Delft, the Netherlands. [15]Centre for Cancer Immunology, University of Southampton, Southampton, UK. [16]UCCS Center for the Biofrontiers Institute, University of Colorado, Colorado Springs, CO, 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, Cell Biology and Biophysics Unit, Heidelberg, Germany. [20]Quantitative Imaging and Nanobiophysics Group, MRC Laboratory for Molecular Cell Biology, University College London, London, UK. [21]Centre for Bacterial Cell Biology, Institute for Cell and Molecular Biosciences, Newcastle University, Newcastle, UK. [22]These authors contributed equally: Daniel Sage, Thanh-An Pham, Seamus Holden. *e-mail: daniel.sage@epfl.ch; seamus.holden@ncl.ac.uk

using biologically inspired structures, using signal-to-noise levels based closely on common experimental conditions, and modeling fluorophore photophysics. We assessed software performance on synthetic datasets for three popular 3D SMLM modalities: astigmatic imaging[10], biplane imaging[15] and double-helix PSF 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 the performance of the latest 2D SMLM software.

Results

Competition design. We established a broad committee comprising members of 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, and this culminated 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 (Supplementary Table 1 and Supplementary Note 1). Participation in the competition actually led at least eight teams to modify their software to support additional 3D SMLM modalities, thus showing how competition can foster microscopy software development.

Realistic 3D simulations. Tests of super-resolution software on experimental data lack the ground-truth information required for rigorous quantification of software performance. Therefore, realistic simulated datasets are required. A critical challenge in simulating 3D SMLM data was accurate modeling of the experimental microscope PSF for each 3D modality. Three-dimensional 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, double helix and biplane—we observed that the PSFs showed complex aberrations not well described by simple analytical models (Supplementary Fig. 1). Even experimental 2D PSFs showed significant aberrations away from the focal plane (Supplementary Fig. 1).

We thus combined experimental 3D PSFs with simulated ground truth by carrying out simulations using PSFs directly derived from experimental calibration data (Fig. 1 and 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 four-state model for photophysics[18] (Methods).

Quantitative performance assessment of 3D software. We assessed software performance on the basis of 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 (Supplementary Fig. 2). Detailed individual software reports are also available, along with a tool for side-by-side comparison of software (Supplementary Figs. 2 and 3).

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 (r.m.s.e.) between measured molecule position and the ground truth.
2. Jaccard index. 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 r.m.s.e. and Jaccard index, which we term the efficiency (Methods).

Choice of ranking metric is discussed in Supplementary Note 2, in which 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 in software performance, even for the most straightforward high-signal-to-noise-ratio (SNR), low-spot-density conditions (Fig. 3 and Supplementary Table 2). The best-in-class software, SMAP-2018[19], had significantly better localization error and Jaccard index performance than average (lateral r.m.s.e. 26 nm best versus 38 nm average, axial r.m.s.e. 29 nm best versus 66 nm average, Jaccard index 85% best versus 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-spot-density, high-SNR dataset for best-in-class and representative middle-range software (Supplementary Fig. 4a). 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 r.m.s.e. away from the plane of focus is significantly higher for the best-in class software, and the Jaccard index is also slightly improved (Supplementary Fig. 4a). This is also visibly apparent in the super-resolved images (Fig. 4a). We observed that best-in-class software had a *z*-range (the full-width at half-maximum (FWHM) range of axially resolved software recall; Methods) of 1,170 nm, greater than two-thirds of the simulated range. Outside this range, the recall and Jaccard index dropped sharply, probably because of the large increase in PSF size and decrease in effective SNR at large defocus (Supplementary Fig. 1).

When we examined results for the low-SNR, low-spot-density dataset (Figs. 2a and 3f), we found an expected twofold degradation in best-in-class r.m.s.e. (lateral r.m.s.e. 39 nm, axial r.m.s.e. 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% versus 85%), although the *z*-range did decrease at lower SNR (930 nm versus 1,120 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 (Supplementary Figs. 5 and 6 and Supplementary Note 3). Close to the focus, best-in-class software was near the CRLB (within 25%), but significant deviations from the CRLB occurred at more than 200 nm (Supplementary Fig. 6). This could be due to difficulty in distinguishing signal from false positives away from focus.

Astigmatic software performance decreased for the challenging high-spot-density datasets (Figs. 2a and 3). For the high-SNR, high-spot-density dataset (best software, SMolPhot), localization error increased and Jaccard index decreased significantly compared with those in the low-spot-density condition (best high-spot-density

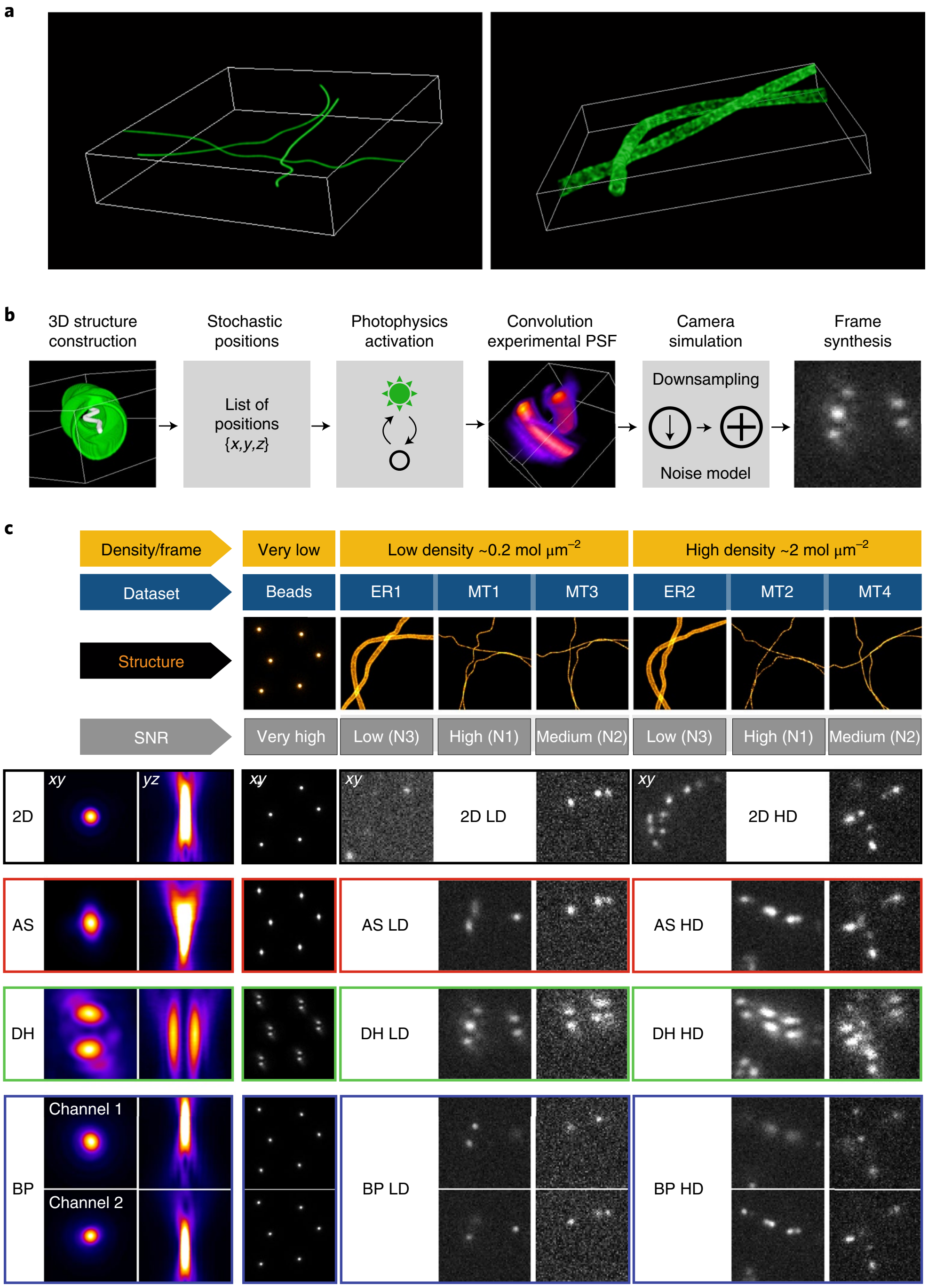

Fig. 1 | Summary of SMLM challenge simulations. a, Three-dimensional rendering of simulated microtubules and endoplasmic reticulum samples. **b**, Key simulation steps. The structure is constructed from 3D tubes continuously defined by three *B*-spline functions in the volume of interest. Membranes of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

performance of 51 nm lateral r.m.s.e., 66 nm axial r.m.s.e., 66% Jaccard index, versus best low-spot-density performance of 27 nm lateral r.m.s.e., 29 nm axial r.m.s.e., 85% Jaccard index). Inspection of the super-resolved images (Supplementary Fig. 7) nevertheless shows qualitatively acceptable results for the high-spot-density dataset, particularly in the lateral dimension. In some circumstances, the performance reduction at ten times higher spot density could be acceptable for ten times faster, potentially live-cell-compatible, imaging speed. We also observed a large spread in software performance for the high-spot-density datasets, probably because a significant fraction of the software packages were designed primarily for low-spot-density conditions.

We observed poor performance for the most challenging low-SNR, high-spot-density astigmatism dataset (Figs. 2a and 3 and

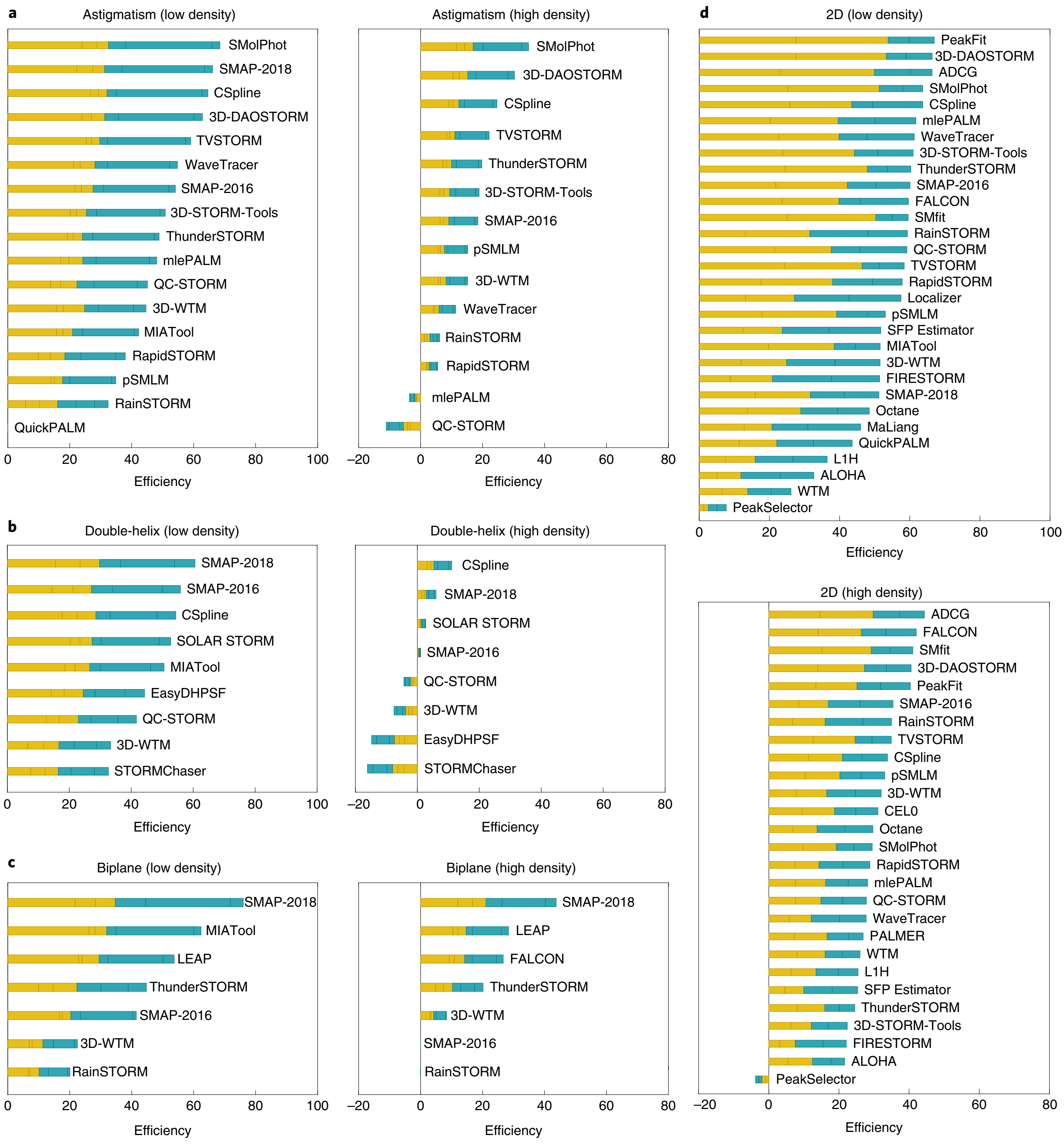

Fig. 2 | Leaderboards for each competition modality, at low and high spot density. Ranking is based on software efficiency, which combines Jaccard index and localization precision (r.m.s.e., lateral and axial). Orange, contribution of high-SNR dataset; blue, contribution of low-SNR dataset.

Supplementary Fig. 8; best software SMolPhot). Best-in-class localization precision and Jaccard index decreased significantly (lateral r.m.s.e. 76 nm, axial r.m.s.e. 101 nm, Jaccard index 58%). These data suggest that low-SNR, high-spot-density 3D astigmatic localization microscopy entails a 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 and Supplementary Fig. 9a). 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 r.m.s.e., 27 nm best versus 37 nm average; axial r.m.s.e., 21 nm best versus 34 nm average; Jaccard index, 77% best versus 73% average). In general, software localization performance was close to the CRLB (Supplementary Fig. 6). We observed that the performance of the software away from the focal plane was relatively uniform (Fig. 4a and Supplementary Fig. 4a),

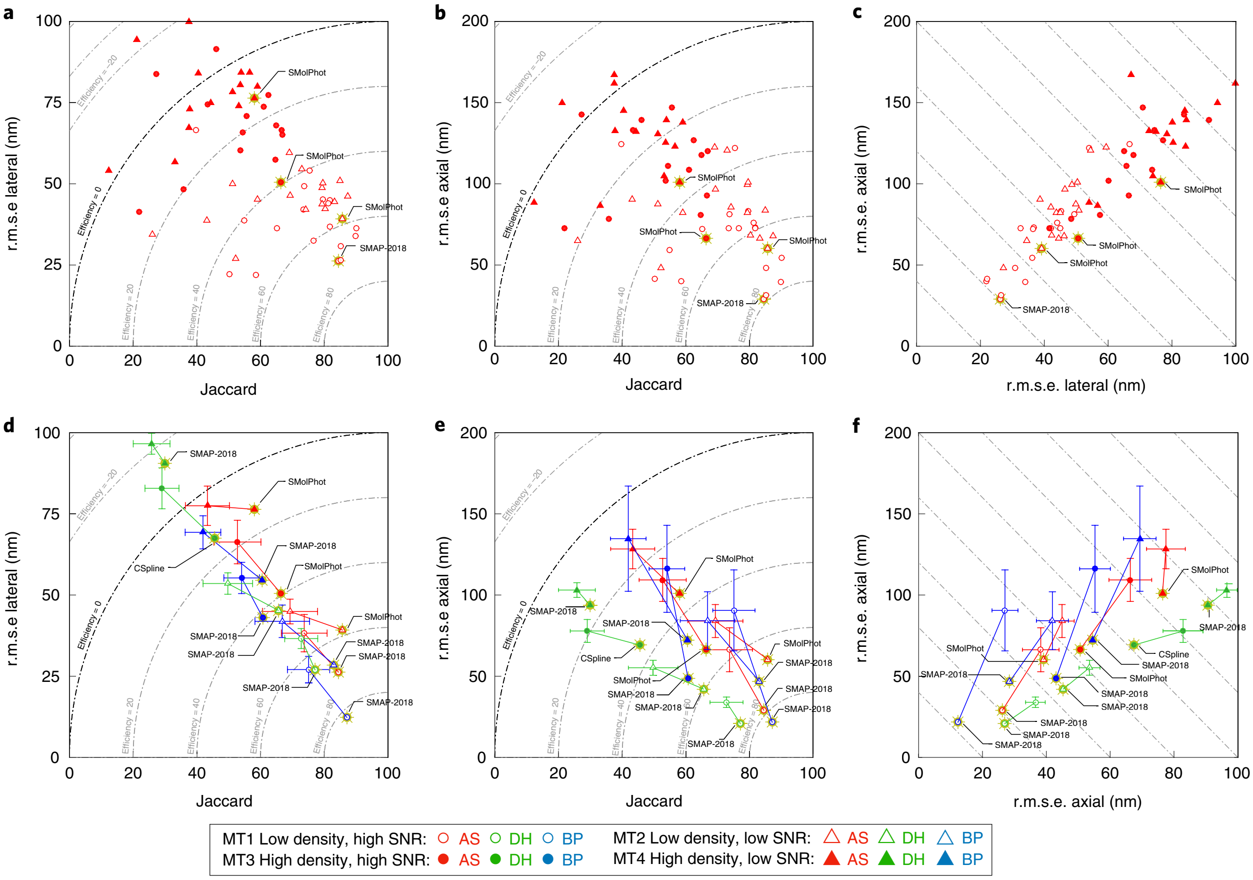

Fig. 3 | Comparison of 3D software performance. a–c, Localization error and spot detection performance of all astigmatic SMLM software. **d–f** Average (error bars are s.d.; sample sizes for each category are indicated in Supplementary Table 2) and best-in-class (marked with gold star) software performance for all competition modalities. Dashed lines in **a**,**b**,**d**,**e** indicate overall efficiency (higher is better).

and the best-in-class *z*-range at high SNR was large at 1,180 nm (Supplementary Fig. 4a and 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 in astigmatic imaging, where spot size, shape and intensity vary greatly as a function of *z* (Supplementary Fig. 1).

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 versus 77% high SNR; Fig. 3d,e and Supplementary Figs. 8 and 9a). Double-helix Jaccard index was also significantly worse than astigmatism results at either high or low SNR (85% high SNR astigmatism, 86% low SNR astigmatism). This poor performance in the low-SNR double-helix dataset is likely because the large size of the double-helix PSF spreads emitted photons over a large area, lowering effective image SNR. Double-helix PSF designs with reduced *z*-range but more compact PSF would probably 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 and Supplementary Fig. 9a; best lateral r.m.s.e. 67 nm, best axial r.m.s.e. 69 nm, best Jaccard index 46%). The poor performance at high spot density is again probably because the large double-helix PSF size increases spot density and decreases SNR (Supplementary Fig. 1). Double-helix PSF performance at high spot density and low SNR was also not reliable (Fig. 3d–f and Supplementary Fig. 9a; 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 r.m.s.e. 12.3 nm, axial r.m.s.e. 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 r.m.s.e. and Jaccard index, with the best-in-class software significantly outperforming the other competitors (Fig. 2c and Supplementary Fig. 9b). 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 1,200 nm at high SNR (Supplementary Fig. 4a,c and Supplementary Table 2). The axial r.m.s.e. was relatively uniform as a function of *z* and close to the CRLB limit (Supplementary Fig. 6). As axial and lateral r.m.s.e. are both averaged over the entire *z*-range, the strong biplane results arise from good performance across a large z-range (Supplementary Fig. 4).

At high spot density and high SNR, best-in-class biplane software (SMAP-2018) showed acceptable performance (Fig. 3d–f and Supplementary Figs. 7 and 9b; best lateral r.m.s.e. 43 nm, best axial r.m.s.e. 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 and Supplementary Figs. 7 and 9b; best lateral r.m.s.e. 55 nm, best axial r.m.s.e. 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-endoplasmic-reticulum 2D dataset at low density, best-in-class software (ADCG[23]) performed

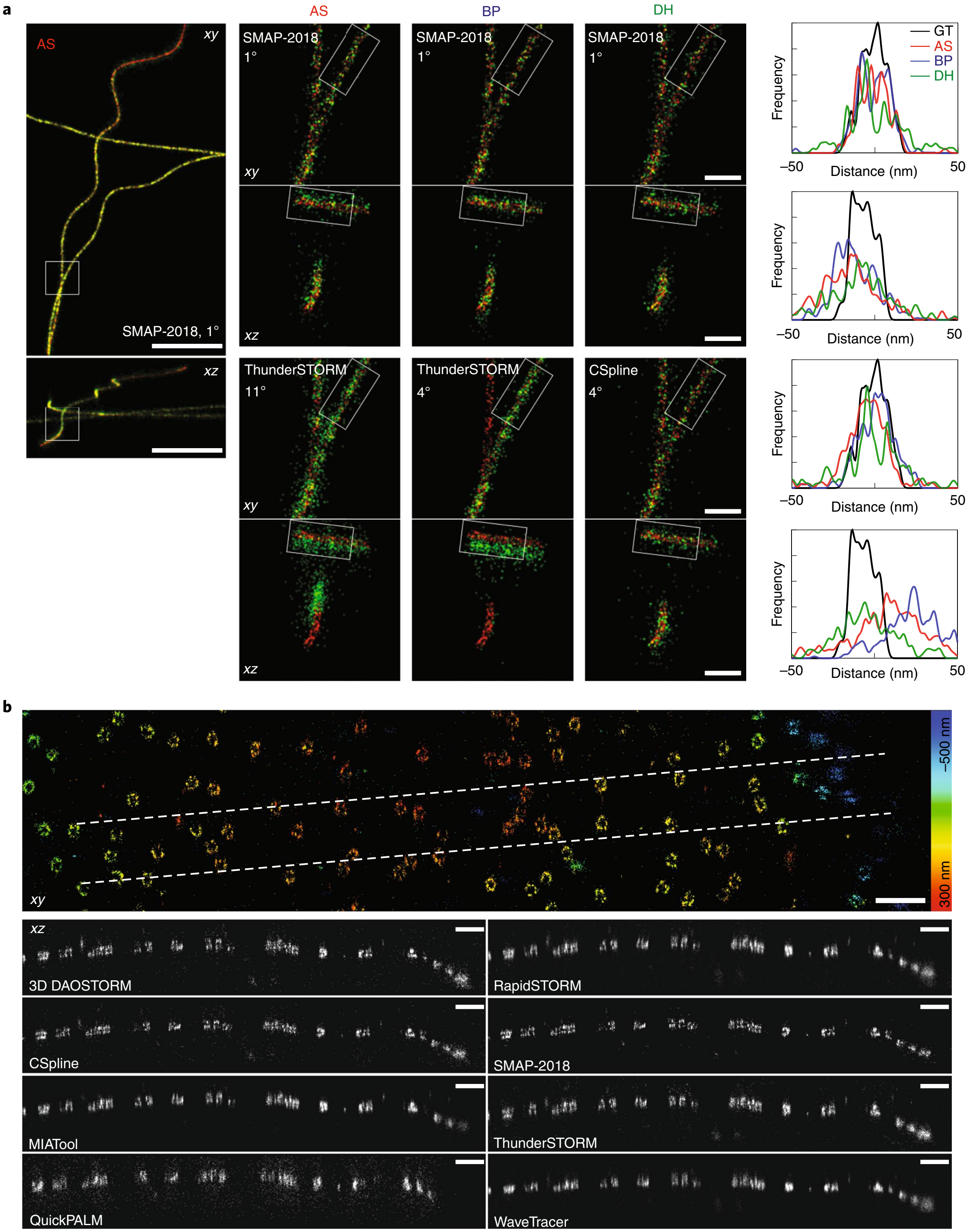

Fig. 4 | Super-resolved images of software results for simulated and real competition datasets. a, Projection images (*xy* and *xz*) of 3D competition datasets for representative software. Top, best-in-class software in each modality, for high-SNR low-spot-density dataset. Bottom, representative average software. Left, *xy* and *xz* overview images for winning astigmatism software. Middle, *xy* and *xz* zoom images of boxed regions in left, for winning and mid-range software, each modality. For each software, the dataset ranking is indicated below. Right, *xy* and *xz* line profiles of winning and mid-range software for each modality, for boxed regions in middle. Image colors: red, ground truth; green, software results. Line profiles: ground truth (GT), black; astigmatism, red; biplane, blue; double helix, green. Scale bars: 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 at the top for eight astigmatism software packages. Scale bars, 500 nm.

substantially better than the class average (Supplementary Figs. 10 and 11; lateral r.m.s.e. 31 nm versus 36 nm average, Jaccard index 90% best versus 72%). Low-spot-density results for the brighter fluorophore microtubules dataset were similar to those for the dimmer pseudo-endoplasmic-reticulum dataset (Supplementary Figs. 10 and 12; best software SMolPhot). For the 2D dataset with very high spot density, which had 25 times higher spot density than the low-spot-density dataset, best-in-class software (ADCG) showed excellent performance (Supplementary Fig. 10; lateral r.m.s.e., 45.5 nm, Jaccard index 75%). Best-in-class performance (ADCG) on the dimmer fluorophore data at high spot density was also strong (Supplementary Fig. 10; best lateral r.m.s.e. 51 nm, best Jaccard index 70%).

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

1. Non-iterative algorithms regroup 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. Single-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 done by local maximum search after application of a denoising filter. Others rely on more complex algorithms such as the wavelet transform (WaveTracer[26]). We did not find 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 (for example, 3D-DAOSTORM[11], CSpline, PeakFit, and ThunderSTORM[27]) was among the top performers in each 2D and 3D competition category. As expected, single- and multiple-emitter fitting methods both performed well on low-spot-density data. For the 2D challenge, multi-emitter fitting showed a clear advantage over single-emitter fitting at high density. However, well-tuned single-emitter fitting algorithms slightly outperformed multi-emitter algorithms for 3D high-spot-density conditions (for example, astigmatism, SMolPhot versus 3D-DAOSTORM). This result merits further investigation, as it conflicts with results for 2D software, and with naive expectation, which suggests that multi-emitter fitting should be a better model for data where PSFs overlap substantially.
4. Compressed sensing algorithms. One subset of these algorithms uses 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 spot density (ADCG first, FALCON[30] second, SMfit third).
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 from 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 for both ease of use and real-time analysis. We did not observe correlation between software localization performance (efficiency) and software run time (Supplementary Fig. 13a). We thus created an alternative ranking metric, 'Efficiency–Run time', which gave 25% weighting to run time (Supplementary Note 2 and Supplementary Fig. 13b). Many good performers in the efficiency-only ranking were relatively fast and thus retained good ranking (SMAP-2018, SMolPhot, 3D-DAOSTORM). Notably, 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 (Supplementary Figs. 14 and 15). 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 reduced software performance considerably. 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 (Supplementary Figs. 3 and 16).

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 (Fig. 4b and Supplementary Fig. 17). 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, as the 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). Although Gaussian/Bessel PSF modeling software (3D-DAOSTORM, MIATool, ThunderSTORM) gave high-resolution images, software that explicitly modeled 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 nuclear pore complex (Fig. 4b) or the hollow core of antibody-labeled microtubules (Supplementary Fig. 17).

Discussion

The strongest conclusion we draw from the 3D localization microscopy challenge is that the 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-CRLB 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 CRLBs 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 modeling 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 double-helix PSF used here; double-helix designs with more compact PSFs 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-spot-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-spot-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-spot-density datasets; this may indicate that potential for improvement remains in this category. Many software packages did not apply temporal grouping[33], and this resulted in reduced software performance. Because 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. Notably, the top three performers in the 2D high-spot-density condition were all compressed sensing algorithms (ADCG, FALCON, SMfit). In low-spot-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 modeling complex PSFs[38] or analyzing high-emitter-density data[40]. However, caution is required about making direct comparisons between algorithms that use strong structural priors to increase performance[37] and algorithms that do not, as the latter may be more robust when presented with novel samples.

In the 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 (complementary metal-oxide semiconductor) 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 (for example, 3B[41], SOFI[42], and 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 the 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.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41592-019-0364-4.

Received: 9 July 2018; Accepted: 26 February 2019;
Published online: 8 April 2019

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. Methods* **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. Methods* **5**, 1047–1052 (2008).
9. Shtengel, G. et al. Interferometric fluorescent super-resolution microscopy resolves 3D cellular ultrastructure. *Proc. Natl Acad. Sci. USA* **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. USA* **106**, 2995–2999 (2009).
17. Anonymous. 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* **15**, 367–369 (2018).
20. Loot, A., Valdmann, A., Eltermann, M., Kree, M. & Pärs, M. SMolPhot software. *BitBucket* https://bitbucket.org/ardiloot/ (2016).
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. Methods* **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, A., 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.* **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 (2018).
36. Marsh, R. J. et al. Artifact-free high-density localization microscopy analysis. *Nat. Methods* **15**, 689–692 (2018).
37. Ouyang, W., Aristov, A., Lelek, M., Hao, X. & Zimmer, C. Deep learning massively accelerates super-resolution localization microscopy. *Nat. Biotechnol.* **36**, 460–468 (2018).
38. Zhang, P. et al. Analyzing complex single-molecule emission patterns with deep learning. *Nat. Methods* **15**, 913–916 (2018).
39. Boyd, N., Jonas, E., Babcock, H. P. & Recht, B. DeepLoco: fast 3D localization microscopy using neural networks. *bioRxiv* Preprint at https://www.biorxiv.org/content/10.1101/267096v1 (2018).
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 (2011).
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. USA* **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**, 12471 (2016).
44. Gustafsson, M. G. L. Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy. *J. Microsc.* **198**, 82–87 (2000).

Acknowledgements

The authors acknowledge the following funding sources: a Newcastle University Research Fellowship and a Wellcome Trust and Royal Society Sir Henry Dale Fellowship grant number 206670/Z/17/Z to S.H.; a European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program, Grant Agreement number 692726 to D.S., T.A.P., and M.U.; UK BBSRC grants BB/M022374/1, BB/P027431/1, BB/R000697/1 grant and MRC grants MC-UU-12018/2, MR/K015826/1 to R.H.; ERC grant CoG-724489, CellStructure to J.R.; FranceBioImaging infrastructure ANR-10-INBS-04 to J.-B.S.; National Institutes of Health grant 1R15GM128166-01 to G.M.H.; and NSF SBIR grants 1353638, 1534745 to Double Helix LLC. We thank R. Piestun at University of Colorado for providing double-helix PSF phase mask designs to Double Helix LLC. We thank all the localization microscopy challenge participants for their contribution: H. Babcock (3D-DAOSTORM, Cspline, L1H); F. Hauser (3D STORM Tools); S. Watanabe (3D-WTM,WTM); N. Boyd (ADCG); J. Min, K. Jin and J.C. Ye (ALOHA, FALCON); H. Rouault (B-recs); E. Soubies (CEL0-STORM); A. Speiser, S. Turagas and J. Macke (DECODE); A. von Diezmann, C. Bayas and W.E. Moerner (Easy-DHPSF); T. Vomhof and J. Reichel (FIRESTORM); H. Pan (LEAP); A. Wheeler (Localizer); Z.-l. Huang and Y. Wang (MaLiang); J. Chao, R. Velmurugan, A.V. Abraham and R.J. Ober (MIATool); H. Deschout (mlePALM); T. Pengo (Octane, PeakSelector); Y.-n. Wang (PALMER); A. Herbert (PeakFit); K. Martens and J. Hohlbein (pSMLM-3D); L. Li (QC-STORM); R. Henriques (QuickPALM); G. Tamas and J. Sinko (RainSTORM); S. Wolter and M. Sauer (RapidSTORM); M. Kirchgessner and F. Gruell (SFP Estimator); Y. Li and J. Ries (SMAP); H. Ikoma (SMfit); A. Loot, A. Valdmann, M. Eltermann, M. Kree and M. Pärs (SMolPhot); Y.J. Jung, A. Barsic, R. Piestun and N. Fakhri (SOLAR_STORM); A. Archetti (STORMChaser); M. Ovesny, G. Hagen and P. Krizek (ThunderSTORM); J. Huang (TVSTORM); A. Kechkar, C. Butler and J.-B. Sibarita (WaveTracer) and B. 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 and Molecular Devices for sponsoring the SMLMS 2016 special session. The sponsors had no input or influence on the research.

Author contributions

D.S. and S.H. conceived and coordinated the study. D.S., S.H., T.-A.P., A. Archetti, H.B., S.C., A.W., G.M.H., R.H., T.L., T.P., and J.-B.S. designed the study. S.H., A. Agrawal, R.H., and J.-B.S. collected experimental PSFs. D.S., T.-A.P., S.H., and T.L. wrote simulation code. B.R. shared unpublished software. D.S. generated simulated datasets. J.R. shared experimental STORM data. A.H., J.R., J.C., and R.V. provided feedback and quality control on simulations and analysis methods. T.-A.P. carried out the assessment of software performance. T.-A.P., D.S., and S.H. analyzed and interpreted the results. D.S., H.B., R.O., B.R., G.M.H., J.-B.S., J.R., R.H., M.U., and S.H. directed research. S.H., D.S., and T.-A.P. wrote the manuscript with feedback from all other authors.

Competing interests

A. Agrawal is an employee of Double Helix LLC, USA.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0364-4.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to D.S. or S.H.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

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 owed to the software SMAP and 3D-WTM[25] that participated in all eight categories (density × modality). The current list of participants is available at http://bigwww.epfl.ch/smlm/challenge2016/index.html?p=participants.

All datasets, methods, participations and results of the challenge 2016 have been 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/).

Localization microscopy simulations. *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 and mitochondria (Supplementary Fig. 18a). These structures have a tubular shape in the 3D space. For the 3D competition, we simulated synthetic 25-nm-diameter microtubules (Fig. 1). Pseudo-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, we simulated larger-diameter 150-nm cylinders, called pseudo–endoplasmic reticulum, designed to approximate larger cellular structures such as mitochondria and the endoplasmic reticulum (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 per pixel) to low resolution (camera resolution: 100 nm per 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, that is, 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 photoactivation 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 SNR levels, based on real signal-to-noise levels encountered under common SMLM experimental scenarios: N1, fixed cells antibody-labeled with organic dye[10], high signal, medium background; N2, fluorescent protein labeling[1], low signal, low background; and N3, live-cell affinity-dye labeling[45,46], high signal, high background.

Spot density. As performance at different densities of active emitters is a key challenge for SMLM software, we generated 3D competition datasets at both sparse emitter density (0.25 molecules per μm^2), 3D low spot density and high emitter density (2.5 molecules per μm^2), 3D high spot density. For the 2D competition, we generated datasets at sparse emitter density (0.5 molecules per μm^2), 2D low spot density, and very high spot density (5 molecules per μm^2), 2D high spot density.

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.

Photophysics activation model. We incorporated a four-state model of fluorophore photophysics[18], including a transient dark state (dye blinking) and a bleaching pathway (Supplementary Fig. 18c). Given a list of source locations from the structure simulator, fluorophore blinking was simulated by a four-states Markov chain model. The states are ON, OFF, BLEACH, DARK and the transitions are Poisson distributed (Supplementary Fig. 18c), except for the OFF to ON transitions, which follow a uniform random distribution to reflect 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

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

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

$$F = \frac{\Phi \ P\sigma}{e}$$

where Φ is the quantum yield of the dye, P is the power of the laser in W cm^{-2}, $e = hc/\lambda$ is the energy of one photon, h is Planck's constant, c is the speed of light, λ is the wavelength, $\sigma = 1{,}000\ \ln(10)\varepsilon/N_{\mathrm{A}}$ is the absorption cross-section in cm^2, and ε is the molar extinction coefficient or absorptivity in cm^2 per molecule, 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 Supplementary Fig. 23.

Experimental PSF. Model PSFs, stored as high-resolution lookup tables, were derived from experimentally measured PSFs. Although the algorithmic approach is distinct, the concept of accurately modeling the experimental PSF on the basis of calibration data bears relation to the PSF phase-retrieval approach previously used by Hanser et al.[47].

Images of fluorescent beads were recorded for each modality (Supplementary Table 5). We maximized the SNR of recorded PSFs 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 number 1.5 (170 μm thick) coverslips, 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 the results of any specific modality, for example, adaptive-optics-corrected astigmatism[48], or reduced z-range, higher SNR double-helix PSF designs[21].

The experimental PSFs used here were measured for fluorescent beads adsorbed to the microscope coverslip, and should be appropriate for simulations of SMLM data acquired within a few micrometers of the coverslip. Performing SMLM imaging at greater depths, for example, in tissue or even deep within single cells, with oil-immersion objectives will cause spherical aberration owing to refractive index mismatch[49]. To accurately simulate SMLM data acquired at depth, the experimental PSFs could be acquired at a matching depth, with fluorescent beads embedded 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].

Simulation PSF construction. For each modality, three to six beads were selected within a small (less than 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 artifacts 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 double-helix 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 (500–1,000 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. We constructed the two defocused PSFs 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 10 × 10 × 10 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).

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 electron-multiplying charge-coupled device (EMCCD) camera[50,51].

The proposed noise model assumed the following 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 the 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)
- $\sigma_R = 74.4$ 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)
- $\mathrm{EM}_{\mathrm{gain}} = 300$ (electron-multiplying gain)
- $e_{\mathrm{adu}} = 45$ electrons per analog-to-digital unit (ADU) (analog-to-digital conversion factor)
- $G = 0.9 \times 300/45 = 6$ (total system gain)
- BL = 100 ADU (baseline)

The final simulated photon electrons will thus be given by

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

$$n_{\mathrm{oe}} = \Gamma\,(n_{\mathrm{ie}}, \mathrm{EM}_{\mathrm{gain}}) + \mathcal{G}(0, \sigma_{\mathrm{R}})$$

where n_{ie} is the number of input electrons, n_{photIn} is the number of input photons, n_{oe} is the number of output electrons, $\Gamma(\cdot)$ is the gamma function, and $\mathcal{G}(\cdot)$ is the Gaussian distribution. This leads to the final pixel count, $\mathrm{ADU}_{\mathrm{out}}$:

$$\mathrm{ADU}_{\mathrm{out}} = \min\left(\mathrm{floor}\left(\frac{n_{\mathrm{oe}}}{e_{\mathrm{ADU}}}\right) + \mathrm{BL}, 65535\right)$$

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]. To compare estimated and true localizations, we correct this effect during the assessment ('Software assessment').

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 for 2D, 3D astigmatism and 3D double-helix 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. That is, the background in each of the two biplane channels is the same as in the single channel of the other modalities. However, because of the low background level in the 3D simulations, the effect on image SNR and thus localization error is small (Supplementary Figs. 5 and 6), 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.

Software assessment. 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 column reordering, reversal of axes, *xy*-axis swap and shifting of the lateral positions by half a 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 (that is, four modalities, four 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, that is, 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, the spurious localizations as false positives and the undetected molecules as false negatives. 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 those of the computationally intensive Hungarian algorithm[7].

Photon thresholding. A photon threshold was required primarily because of the use of a realistic fluorophore blinking model. A fluorophore can activate/bleach at any point in a simulated frame, and this led to many frames containing very dim, undetectable localizations, for example, 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 PSFs 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 through reference to a calibration curve[37]. Because our simulated PSF is experimentally derived, it was necessary to correct for these artifactual shifts between the observed localizations and ground truth as part of the assessment process. This correction was done 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; or the correction was calculated by us. The latter required the participant to localize a stack of beads we provided. Because 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 artifact. 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 (Supplementary Fig. 19).

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 localization positions and notably includes the recall, the precision, the Jaccard index, the r.m.s.e. (axial and lateral) and the consolidated *z*-range. For the calculation of average software performance (Fig. 3d–f and Supplementary Fig. 10), outlier software with an efficiency less than 0 (efficiency = –30 for the 3D high-spot-density dataset) were excluded from the measurement. The key metrics of assessment were as follows:

1. Root mean squared error. The foremost consideration for localization software is how accurately it finds the position of labeled 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 (TP indicates true positives).

$$\text{r.m.s.e. lateral (nm)} = \sqrt{\frac{1}{\mathrm{TP}} \sum\nolimits_{i \in S} (x_i^s - x_i^t)^2 + (y_i^s - y_i^t)^2}$$

$$\text{r.m.s.e. axial (nm)} = \sqrt{\frac{1}{\mathrm{TP}} \sum\nolimits_{i \in S \cap T} (z_i^s - z_i^t)^2}$$

2. Jaccard index (%). In addition to localization precision, SMLM image resolution depends critically on the 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 were close enough to a ground-truth position as true positives, the spurious localizations as false positives (FP) and the undetected molecules as false negatives (FN). We then computed the Jaccard index, which measures the fraction of correctly detected molecules in a dataset:

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

3. Efficiency. 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.

$$E = 100 - \sqrt{(100 - \mathrm{JAC})^2 + \alpha^2\ \mathrm{r.m.s.e.}^2}$$

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}, on the basis of the linear regression slope between the localization errors and Jaccard index (Supplementary Fig. 20j,k). Using this definition, an average software performance has an efficiency in the range 25–75, and 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 include the SNR and the Fourier ring/Fourier shell correlation. 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 FWHM of 20 nm. A complete list of all computed metrics is presented in 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 (that is, from the true localizations). Then, most of the metrics (for example, recall) are locally computed. This yields a curve providing information on the depth performance of each software/modality.

To summarize software axial performance, we analyzed how the recall varied as a function of z. A typical recall versus axial position curve (Supplementary Fig. 4) will drop at positions far from the focal plane, that is, 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 FWHM recall of the smoothed curve (Supplementary Fig. 21). This quantity is visually intuitive and useful for discussion of the recall performance if considered alongside a plot of recall versus axial position. However, because FHWM recall depends on the maximal recall, ranking based on this procedure would promote a software that performed poorly everywhere (that is, a flat curve), whereas a software that 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. To analyze the relationship between analysis metrics, we computed the covariance matrix between each metric (Supplementary Fig. 22a) and the principal component analysis on the metrics (Supplementary Fig. 22b–d). Each metric was standardized before application of the covariance and the principal component analysis. For convenience, we took the additive inverse of the metrics for which lower values are best (that is, false positives, false negatives, r.m.s.e., and Fourier ring and Fourier shell correlations).

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.

Baseline localization software. We developed a minimalist Java tool software that carries out localizations of bright emitters on the four modalities of the challenge 2016: 2D, astigmatism, double helix and biplane. This SMLM_BaselineLocalization software was designed only to establish the performance baseline for the SMLM challenge. It has intentionally limited lines of code and relies on only a few threshold parameters to localize particles. It has a 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 by the pairing of two close points.
- Biplane: $z = f(W_{\mathrm{left}} - W_{\mathrm{right}})$, where W_{left} and W_{right} are, respectively, an estimation of the size of the spots in the left and the right planes.

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

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 Supplementary Fig. 6 in ref. [19], and the nuclear pore complex dataset corresponds to raw data for Supplementary Fig. 9 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 using cross-correlation. STORM images were rendered with a constant Gaussian blur with 3 nm s.d. and saturated by 0.1–0.5%. The complete scripts used for assessment and image rendering are available on the competition GitHub page.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

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 to allow the software challenge to remain continuously open to new submissions. However, ground-truth data are available for the simulated training datasets. Source data for Figs. 1–4 and for Supplementary Figs. 4–7, 19, 20 and 22 are available online.

Code availability

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

References

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. USA* **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).

nature research

Corresponding author(s): Seamus Holden

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☒	□	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
□	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Simulated datasets are available on the SMLM challenge website http://bigwww.epfl.ch/smlm/challenge2016/
Data analysis	Software is available on the SMLM challenge github https://github.com/SMLM-Challenge/Challenge2016

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

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 is available for the simulated training datasets.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Study consists of analysis of software results when applied to simulated data. We ensured that simulated datasets were sufficiently large (> 10000s of single molecule localizations) to minimize noise on analysis statistics.
Data exclusions	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. This is described in section 3.2 of the Online Methods.
Replication	Simulated datasets: Each simulated frame is effectively a replicate, and each dataset contained between 3000-20000 simulated frames. Ie replication was successful. Experimental datasetsWe performed qualitative analyses on 2 experimental sets for different biological test structures. Although qualitative, both analyses of experimental datasets gave self consistent results.
Randomization	N/A as analysis was performed on a per-software basis
Blinding	N/A as all analysis was performed automatically performed identically on each dataset.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study	
☒	□	Unique biological materials
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology
☒	□	Animals and other organisms
☒	□	Human research participants

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging