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

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

*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: University of Minnesota Informatics Institute, University of Minnesota Twin Cities, USA
5: Electrical Engineering, University of Texas Dallas, Richardson, Texas, USA
6: MRC Genome Damage and Stability Centre, School of Life Sciences, University of Sussex, Brighton, UK
7 : Double Helix LLC, Boulder, Colorado, USA
8 : Istituto Italiano di Tecnologia, Genova, Italy
9: Edinburgh Super-Resolution Imaging Consortium, University of Edinburgh, UK
10 : Laboratory of Experimental Biophysics, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
11: Department of Imaging Physics, Faculty of Applied Sciences, Delft University of Technology, The Netherlands
12: UCCS center for the Biofrontiers Institute, University of Colorado at Colorado Springs, Colorado, USA
13: Institut Interdisciplinaire de Neurosciences, University of Bordeaux, France
14: European Molecular Biology Laboratory (EMBL), Cell Biology and Biophysics Unit, Heidelberg, Germany
15: Quantitative Imaging and Nanobiophysics Group, MRC Laboratory for Molecular Cell Biology, University College London, UK
16: 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. To guide researchers on the optimal analytical software for their experiments, we have designed, in a large community effort, a competition to extensively characterise and rank these options. 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, provides a holistic view of how the latest 2D and 3D single molecule localization software perform in realistic conditions, and ultimately provides insight into the current limits of the field.

INTRODUCTION

46
47 Image processing software is central to single molecule localization microscopy (SMLM), which
48 delivers an order of magnitude resolution improvement on diffraction limited conventional
49 fluorescence microscopy, from 250 nm to approximately 20 nm resolution, by temporal separation of
50 fluorophores within a sample[1–3]. Efficient and automated image processing is essential to extract the
51 super-resolved positions of individual molecules from thousands of raw microscope images,
52 containing millions of blinking fluorescent spots.

53 Improvements in SMLM image processing algorithms have been crucial in maximizing spatial
54 resolution and in reducing the imaging time of SMLM for compatibly with live cell imaging[4–6]. If SMLM
55 is to achieve a resolving power approaching that of electron microscopy, the analysis software
56 employed needs to be robust, accurate, and performing at current algorithmic limits. This can only be
57 achieved through rigorous quantification of SMLM software performance.

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

65 There are roughly three common approaches for 3D SMLM. First, point spread function engineering,
66 where the axial asymmetry of the microscope point spread function (PSF) is increased by introducing
67 intentional aberrations in the system, ranging from simple astigmatism[10] to more complex PSF
68 manipulation such as the double helix PSF method[11]. Second, biplane or multiplane imaging, where
69 axial position is measured based on simultaneous measurement of PSF shape at two or more focal
70 planes[12]. Third, dual objective based interferometry, where Z-position is calculated from single photon
71 interference between opposing objectives[13]. Multiplane and PSF engineering methods typically obtain
72 axial resolutions on the order of 50 nm[10,11]. Interferometry achieves the best axial resolution, 10-20
73 nm[13], but is not yet widely adopted.

74 Despite the widespread use of 3D localization microscopy, and challenging nature of 3D SMLM image
75 processing, the performance of software for 3D single molecule localization microscopy has previously
76 only been assessed for 2 or 3 software packages at a time, and without standard test data or metrics[14–
77 17]. In the absence of common reference datasets and reliable assessment procedure of 3D software
78 performance, it is not possible to objectively assess how different software affects final image quality,
79 or which algorithmic approaches are most successful. Crucially, end-users cannot determine which 3D
80 SMLM software package and imaging modality is optimal for their application.

81 We therefore ran the first 3D localization microscopy software challenge, to assess the performance
82 of 3D SMLM software. We assessed software performance on synthetic datasets for three popular 3D
83 SMLM modalities: astigmatic imaging, biplane imaging and double helix point spread function
84 microscopy. We also assessed astigmatism software performance on two real STORM datasets. We
85 ran a second 2D localization microscopy software challenge, to reassess the 2D SMLM software state-
86 of-the art on new, tougher, more realistic datasets.

87 Our simulations incorporate experimentally acquired point spread functions for maximal authenticity,
88 used signal and noise levels based closely on common experimental conditions, and incorporated a
89 realistic 4-state model of fluorophore photophysics[18]. Our synthetic data was designed to mimic two
90 common classes of cellular structure: narrow line-like microtubules (MT) and larger tubes similar to
91 the endoplasmic reticulum (ER) or mitochondria. Our simulations also included conditions with low
92 density (LD) of active fluorophores, used experimentally to obtain maximal resolution, and with high
93 density (HD) of active fluorophores, used experimentally for fast or live cell imaging.

94 **RESULTS**

95 **Competition design**

96 We established a large committee from within the SMLM research community, including
97 experimentalists and software developers, to define the scope of the challenge, ensure realism of the
98 datasets and define analysis metrics. We further opened this discussion to the whole community,
99 through an open forum, discussing best practices for the implementation of this contest[19].

100 Thirty-six software packages have been entered in the competition thus far, including four packages
101 used in commercial software (Table S1, Supplementary Note 5). Excitingly, participation in the
102 competition actually led at least 8 teams to their software to support additional 3D SMLM modalities,
103 showing how competition fosters microscopy software development.

104 In 2016, we ran a first round of the 3D SMLM competition with explicit submission deadlines, with 30
105 competitor teams, culmination in a special session at the 6th annual Single Molecule Localization
106 Microscopy Symposium (SMLMS 2016). Since then, the challenge has been opened to continuously
107 accept new entries. We have had 12 new registrations of which 5 have submitted localizations,
108 including a multiple best-in-class performer (SMAP-2018[20], an updated version of previously entered
109 software) demonstrating the utility of the competition as an evolving measure of the state of the field.

110 **Realistic 3D simulations**

111 Testing super-resolution software on experimental data lacks the ground truth information required
112 for rigorous quantification of software performance. Therefore, realistic simulated 3D SMLM datasets
113 are required. After comparison of simulated microscope PSFs with multiple experimental PSFs from
114 SMLM microscopes around the world, we observed that a critical challenge to realistic 3D SMLM
115 simulations was to accurately model the experimental microscope PSF for each 3D modality. Even
116 experimental 2D PSFs showed significant aberrations away from the focal plane (Fig S10).

117 3D SMLM inherently involves addition of aberrations to the microscope PSF to encode the Z-position
118 of the molecule. For the PSF models included in the competition: 2D, astigmatic (AS), double helix
119 (DH), and biplane (BP), we observed that the PSFs showed complex aberrations not well described by
120 simple analytical models (Fig S10). We thus combined experimental 3D PSFs with simulated ground
121 truth by performing simulations using PSFs directly derived from experimental calibration data (Fig 1,
122 *Methods*). The experimental PSFs used to generate the simulated data are available online (*Methods*).
123 As the goal of this study was to compare software obtained on typical SMLM microscopes, we
124 deliberately chose PSFs representative of common implementations of each 3D modality. However,
125 additional PSF engineering should improve results of any specific modality, for example adaptive-
126 optics corrected astigmatism[21], or reduced Z-range, higher SNR DH-PSF designs[22].

127 For the 3D competition, we simulated synthetic 25 nm diameter microtubules (Fig 1). For the 2D
128 competition, in addition to synthetic microtubules (MT), we simulated larger diameter 150 nm
129 cylinders, designed to approximate larger cellular structures such as mitochondria and the
130 endoplasmic reticulum (ER) (Fig 1). We incorporated a 4-state model of fluorophore photophysics,
131 including a transient dark state (dye "blinking") and a bleaching pathway (Fig S1C).

132 As performance at different density of active emitters is a key challenge for SMLM software, we
133 generated 3D competition datasets at both sparse emitter density (0.2 mol. [molecule] μm^{-2}) and high
134 emitter density (2 mol. μm^{-2}). We additionally generated a very high density dataset (5 mol. μm^{-2}) for
135 the 2D competition.

136 We generated data at three different signal-to-noise ratio (SNR) levels, based on real signal to noise
137 levels encountered under common SMLM experimental scenarios: fixed cells antibody labelled with
138 organic dye[10], fluorescent protein labelling[1], and live cell affinity dye labelling[23,24].

139 Together, these simulations closely resemble experimental 3D and 2D data under a range of
140 challenging conditions of SNR, spot density, axial thickness and structure summarized in Table S2. In
141 addition, we provide simulated z-stacks of bright beads for software calibration. The competition
142 datasets are available online (*Methods*).

Quantitative performance assessment of 3D software

144 We assessed software performance by 26 quality metrics (*Supplementary Note 1*). The complete set
145 of summary statistics, axially resolved performance and super-resolved images is available for each
146 competition software on the competition website. We built an interactive ranking and graphing
147 interface that allows easy ranking and graphing of software performance by any metric, including new
148 user defined metrics (Fig S11). Detailed individual software reports can also be accessed, along with
149 a tool for side-by-side comparison of software (Fig S11, S22).

150 We focused our analysis primarily on metrics directly derived from single molecule localizations.
151 Choice of ranking metric is discussed in detail in Supplementary Note 1.6, where several alternative
152 ranking metrics are also presented.

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

157 *2. Ability to successfully detect fluorescent molecules.* In addition to localization precision, SMLM
158 image resolution depends critically on number of localized molecules[25], so it is crucial for SMLM
159 software to accurately detect a large fraction of molecules in a dataset, and minimize false
160 localizations. For every frame, we identified the localizations that are close enough to a ground-truth
161 position as true-positives (TP), the spurious localizations as false-positives (FP) and the undetected
162 molecules as false-negatives (FN). We then computed the *Jaccard index* (JAC, %), which measures the
163 fraction of correctly detected molecules in a dataset:

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

164

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

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

168

169 The trade-off between these two metrics is controlled by a parameter α. In a retrospective analysis,
170 we chose $\alpha = 1$ nm^{-1} for the lateral efficiency E_{lat}, $\alpha = 0.5$ nm^{-1} for the axial efficiency E_{ax}, based on the
171 linear regression slope between the localization errors and Jaccard index (Fig 17J-K). Using this
172 definition, an average software performance has an efficiency in the range 25-75, a perfect software
173 would have the maximum efficiency of 100. Overall 3D efficiency was calculated as the average of
174 lateral and axial efficiencies. Overall software rankings (Fig 2) were calculated as the sum of rankings
175 for high and low SNR datasets.

Performance of 3D software

177 Complete rankings for each imaging modality and spot density are presented (Fig 2), together with
178 summary information on all competition software (Table S1, *Supplementary Note 1*). As these data
179 are continuously updated on the competition website, this resource provides microscopists with a
180 quick reference for the current state of the art, including current best-in-class performers for each
181 category.

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

184 *Astigmatic localization microscopy*

185 Astigmatic localization microscopy is probably the most popular imaging 3D SMLM modality, reflected
186 by the highest number of software submissions in the 3D competition (Fig 2). For astigmatism, we
187 observed a large spread of software performance, even for the most straightforward high SNR, low
188 spot density (LD) conditions (Fig 3, Table S5). The best-in-class software (SMAP-2018) has significantly
189 better localization error and Jaccard index performance than average (lateral RMSE 26 nm best vs 38
190 nm average, axial RMSE 29 nm best vs 66 nm average, Jaccard index 85 % best vs 74 % average).
191 Clearly, the quality of the image reconstruction depends strongly on choice of 3D software.

192 To investigate the reasons for software variation, we inspected plots of software performance as a
193 function of axial position in the low density, high SNR dataset for best-in-class and representative
194 middle-range software (Fig S7A). We observed that the key cause of the spread in software
195 performance is variation in software performance away from the focal plane. Near the focal plane,
196 most software packages perform well. However, the axial and lateral RMSE away from the plane of
197 focus is significantly higher for the best in class software, and the Jaccard index is also slightly improved
198 (Fig S7A). This is also visibly apparent in the super-resolved images (Fig 4A). We observed that best-in-
199 class software had a Z-range (the FWHM range of axially resolved software recall, *Methods*) of 1170
200 nm, greater than two-thirds of the simulated range. Outside this range, the recall and Jaccard index
201 dropped sharply, probably due the large increase in PSF size and decrease in effective SNR at
202 significant defocus (Fig S10).

203 When we examined results for the low SNR, low density dataset (Fig 2B, 3B), we found an expected
204 two-fold degradation in best-in-class RMSE (lateral RMSE 39 nm, axial RMSE 60 nm), due to the
205 decrease in image SNR. However, the best-in-class software (SMolPhot[26]) Jaccard index was effectively
206 constant between the low and high SNR datasets (86 % vs 85 %), although the Z-range did drop at
207 lower SNR (930 nm vs 1120 nm). The best astigmatism software packages were thus remarkably good
208 at finding spots at low SNR, even away from the plane of focus.

209 We analyzed how close software performance was to theoretical limits by calculating the Cramér-Rao
210 lower bound (CRLB) as a function of axial position for each dataset and comparing it to the best-in-
211 class software results (Fig S8, S9, Supplementary Note 4). Close to the focus, best-in-class software
212 was close to CRLB performance (within 25 %), but significant deviations for the CRLB limit occurred >
213 200 nm. This could be due to the difficulty in actually detecting the spots away from focus.

214 When we examined astigmatic software performance for the challenging high spot density datasets
215 (Fig 2B, 3), performance was reduced. For the high SNR high spot density dataset (best software,
216 SMolPhot), localization error increased and Jaccard index decreased significantly compared to the low
217 density condition (lateral RMSE best HD 51 nm vs best LD 27 nm, axial RMSE best HD 66 nm vs best
218 LD 29 nm, Jaccard index best HD 66 % vs best LD 85 %). Inspection of the super-resolved images
219 (Fig S3) nevertheless shows acceptable results for the HD dataset, particularly in the lateral dimension.
220 In many circumstances, the performance reduction at 10x higher spot density should be acceptable
221 for 10x faster, potentially live-cell-compatible, imaging speed. We also observed a large spread of
222 software performance for the high density datasets, probably because a significant fraction of the
223 software packages were primarily designed for low density conditions.

224 We observed poor performance for the most challenging low SNR high spot density astigmatism
225 dataset (Fig 2, 3, S4, best software SMolPhot). Best-in-class localization precision and Jaccard index
226 decreased significantly (lateral RMSE 76 nm, axial RMSE 101 nm, Jaccard index 58 %). These data
227 suggest that low SNR high density 3D astigmatic localization microscopy entails a significant reduction
228 in image resolution.

229 *Double helix point spread function localization microscopy*

230 We next analyzed the performance of the double helix software (Fig 3B, S14A). For the software in the
231 high SNR low spot density condition, double helix software showed more uniform performance than
232 astigmatism. Best-in-class software (SMAP-2018) showed only a limited improvement compared with
233 average software (Fig 3B, lateral RMSE, 27 nm best vs 37 nm average; axial RMSE 21 nm best vs 34 nm
234 average; Jaccard index 77 % best vs 73 % average). In general software localization performance was
235 close to the CRLB (Fig S8, S9). We observed that performance of the software away from the focal
236 plane is relatively uniform (Fig 4A, S7A), and best-in-class Z-range at high SNR was large at 1180 nm
237 (Fig S7, Table S5). Double helix imaging may show less software-to-software variation and large Z-
238 range at low spot density than astigmatic imaging because the PSF shape and intensity are fairly
239 constant as a function of Z – compared to astigmatic imaging, where spot size, shape and intensity
240 vary greatly as a function of Z (Fig S10).

241 Double helix software performance decreased significantly for the low spot density low SNR condition
242 (best software SMAP-2018), particularly in terms of best-in-class Jaccard index (66 % low SNR vs 77 %
243 high SNR, Fig 3B, S4, S14A). DH Jaccard index was also significantly worse than astigmatism results at
244 either high or low SNR (85 % high SNR, 86 % low SNR). This indicates that it was quite hard to
245 successfully find localizations in the low SNR DH dataset, likely because the large size of the DH PSF
246 spreads emitted photons over a large area, lowering effective image SNR. DH PSF designs with reduced
247 Z-range but more compact PSF would likely be less sensitive to this issue[22].

248 Double helix software performed poorly on the high spot density datasets at high SNR (best software
249 CSpline[27]), especially in terms of the Jaccard index (Fig 3B, S14A, best lateral RMSE 67 nm, best axial
250 RMSE 69 nm, best Jaccard index 46 %). The poor performance at high spot density is again probably
251 because the large DH PSF size increases spot density and decreases SNR (Fig S10). DHPSF performance
252 at high spot density and low SNR was also not reliable (Fig. 3B, S14A, best software SMAP-2018).

253 *Biplane localization microscopy*

254 Best-in-class biplane software (SMAP-2018), at low spot density and for both high and low SNR,
255 delivered the best performance in any modality (high SNR: lateral RMSE 12.3 nm, axial RMSE 21.7 nm,
256 Jaccard 87 %), despite a slightly decreased image SNR for the biplane simulations (*Methods*). We
257 observed a significant spread in software performance in terms of lateral RMSE and Jaccard index,
258 with the best-in-class software significantly outperforming the other competitors (Fig S14B, 2D). At
259 low spot density, best-in-class biplane software (SMAP-2018) showed good performance as a function
260 of Z, with high Jaccard index over almost the entire Z-range of the simulations, and with a Z-range of
261 1200 nm at high SNR (Fig S7, Table S5). The axial RMSE was relatively uniform as a function of Z and
262 close to the CRLB limit (Fig S7). As axial and lateral RMSE are both averaged over the entire Z-range,
263 the strong biplane results arise from good performance across a large Z-range (Fig S7).

264 At high spot density and high SNR, best-in-class biplane software (SMAP-2018) showed acceptable
265 super-resolved performance (Fig 3B, S3, S14B, best lateral RMSE 43 nm, best axial RMSE 49 nm, best
266 Jaccard index 61 %). Uniquely among the 3D modalities, best-in-class biplane software also gave
267 acceptable performance at high spot density and low SNR (Fig 3B, S3, S14B, best lateral RMSE 55 nm,
268 best axial RMSE 72 nm, best Jaccard index 61 %, best software SMAP-2018).

269 **Performance of 2D software**
270 Alongside the 3D challenge, we ran a second edition of the 2D localization microscopy software
271 challenge[7] to assess how the latest 2D software performed on more challenging, more realistic
272 datasets, and to provide an assessment of how the field had progressed since the last challenge. We
273 used the new simulation software, including an experimentally derived PSF and a realistic blinking
274 model, and also simulated a very high spot density condition (5 molecules/ μm^2). We created a more
275 spatially extended test structure, "pseudo-endoplasmic reticulum" (pseudo-ER), composed of 150 nm

276   diameter hollow tubes, to avoid artefacts due to 1D simulated structures[28]. We generated two
277   different imaging conditions with overall similar SNR but different brightness properties; one with low
278   fluorophore brightness and low autofluorescence (the low SNR condition for the 3D challenge,
279   designed to simulate fluorescent protein based SMLM, Fig S5) and one with high fluorophore
280   brightness and high autofluorescence (to simulate affinity-dye-based live cell SMLM, Fig S6). We used
281   lateral RMSE, Jaccard index and overall lateral efficiency to rank the 2D software (Fig 2, S2, Table S1).

282   For the pseudo-ER dataset, at low density, best-in-class software (ADCG) performed well (Fig. S2, S5),
283   with a Jaccard index of 90 % and lateral RMSE of 31 nm, substantially better than the class average
284   (Jaccard index 72 %, lateral RMSE 36 nm). Low density results for the dimmer fluorophore
285   microtubules dataset were similar to the brighter pseudo-ER dataset (Fig S2, best software SMolPhot).
286   For the very high density 2D dataset, which had 25x higher spot density than the LD dataset, best-in-
287   class software (ADCG) showed excellent performance, with Jaccard index of 75% and lateral RMSE of
288   45.5 nm (Fig S2). Best-in-class performance (ADCG) on the dimmer fluorophore data at high spot
289   density was also strong (Fig S2, best Jaccard index 70 %, best lateral RMSE 51 nm).

Algorithms
291   We identified several classes of algorithm participant software (Table S1):

292   1) *Non-iterative* software tends to regroup the pixels in the local neighborhood of the candidates, like
293   interpolation, center of mass (QuickPALM[29]) or template matching (WTM[30]). These (often older)
294   algorithms are fast but tend to achieve poor performance (Table S1).

295   2) S*ingle emitter fitting* software is usually built on a multi-step strategy of detection, spot localization,
296   and optional spot rejection. The detection step finds bright spots in noisy images on the pixel grid. The
297   selection of candidates is usually performed by local maximum search after a denoising filter. Others
298   rely on more complex algorithms like the wavelet transform (*e.g.,* WaveTracer[31]). We did not observe
299   software ranking to depend significantly on the choice of optimization scheme, least-square, weighted
300   least-square or maximum-likelihood estimator (Table S1).

301   3) *Multi-emitter fitting* software groups clusters of overlapping spots, and simultaneously fits multiple
302   model PSFs to the data. Typically, fitted spots are added to the cluster until a stopping condition is
303   met[4,5]. This leads to improved localization performance at high spot density, at the cost of reduced
304   speed. This class of software (*e.g.,* 3D-DAOSTORM[14], CSpline, PeakFit[32], ThunderSTORM[33]) was
305   amongst the top performers in each 2D and 3D competition category (Table S1).

306   As expected, single- and multiple-emitter fitting methods both performed well on low density data
307   (Table S1); apparently at the densities studied, exclusion of occasionally overlapping spots by single-
308   emitter software is sufficient for strong performance; explicit multi-emitter fitting is not required. For
309   the 2D challenge, multi-emitter fitting showed a clear advantage over single emitter fitting at high
310   density (Table S1). Surprisingly however, well-tuned single-emitter fitting algorithms (SMolPhot,
311   SMAP-2018) outperformed multi-emitter algorithms for the 3D high density conditions.

312   4) *Compressed sensing algorithms.* One subset of these algorithms utilize deconvolution with sparsity
313   constraints to reconstruct super-resolved images[34–36]. Although deconvolution approaches can give
314   good results, they are limited by the necessary use of a sub-pixel grid; increased localization precision
315   requires smaller grid resolution, which must be balanced against increased computational time.
316   Recent approaches address this issue by localizing the point sources in a grid-less manner using an
317   alternating descent conditional gradient scheme under some sparsity constraint (ADCG[37], SMfit,
318   SOLAR_STORM, TVSTORM[38]). This software class consistently gave the overall best performance for
319   2D high-density (ADCG[37] 1st, FALCON[36] 2nd, SMfit 3rd).

320   5) *Other approaches.* Of the alternative algorithmic approaches used (Table S1), the annihilating filter-
321   based method LEAP[39] gave good performance for biplane imaging (the only modality for which it was

322 entered). Recently, we received the first challenge submission from a deep learning SMLM software
323 (DECODE); these promising preliminary results are available on the competition website.

Post-hoc temporal grouping

325 Because molecule on-time is stochastically distributed across multiple frames, a common post-
326 processing approach to improve localization precision is to group molecules detected multiple times
327 in adjacent frames, and average their position[40] (Supplementary Note 3). Temporal grouping was used
328 by the top performers (including SMolPhot, MIATool[41] and SMAP-2018), and is visibly apparent as a
329 more punctate super-resolved image (Fig 4A).

Choice of PSF model

331 Most software used a variant of Gaussian PSF model. A few participants designed more accurate PSF
332 models (Table S1). Either diffraction theory was used (MIATool[41], LEAP[39]) or spline fitting of an
333 analytical function to the experimental PSF was adopted (CSpline, SMAP-2018). Although simple
334 Gaussian model PSFs were sufficient to obtain best-in-class performance for the 2D and astigmatic
335 modalities (ADCG[37], PeakFit, SMolPhot), top results for the more optically complex biplane and double
336 helix modalities were exclusively PSF-modelling algorithms (SMAP-2018, CSpline, MIATool, LEAP).

Multi-algorithm packages

338 Several software packages take a Swiss army knife approach of integrating multiple optional
339 localization algorithms into one program, to be flexible enough to suit various experimental
340 conditions[20,33]. SMAP-2018 and ThunderSTORM achieved strong across-the-board performance
341 supporting this rationale.

Software run time

343 Software run time is an important parameter for ease of use, and to facilitate real time analysis. We
344 did not see any correlation between software localization performance (Efficiency) and software run
345 time (Fig S24A). We thus created an alternative ranking metric, *Efficiency-Runtime*, which gave 25 %
346 weighting to run time (Supplementary Note 1.7, Fig S24B). Many good performers in the efficiency-
347 only ranking were relatively fast and thus retained good ranking (SMAP-2018, SMolPhot, 3D-
348 DAOSTORM). Interestingly, two software packages highly optimized for speed gained top ranking in
349 this analysis: pSMLM-3D[42] and QC-STORM.

Diagnostic tools for software and algorithm performance

351 During our analysis, we frequently noticed common types of deviation between software results and
352 ground truth which were easily diagnosed by visual inspection of the super-resolved comparison
353 overlay of ground truth and observed localizations (Fig S19-20). This included not only obvious issues
354 of poor localization precision or spot averaging at high density, but also other problems such as a
355 common error of structural warping which significantly reduced software performance. On the
356 competition website, we provide detailed diagnostic software reports including multiple examples of
357 software performance on individual frames which should help developers to identify algorithm and
358 software limitations and maximize software performance (Fig S21-22).

Assessment of software performance on real 3D STORM data

360 We investigated the performance of a representative subset of astigmatism software on real STORM
361 datasets of well-characterized test structures, microtubules and nuclear pore complex, NPC (Fig 4B,
362 S15). This qualitative assessment was consistent with findings for simulated data. No performance
363 difference between single and multi-emitter fitters was observed, which is not surprising since spot
364 density in these datasets was low. Relatively poor software performance was immediately obvious
365 from visual inspection (QuickPALM). Temporal grouping noticeably improved resolution (3D
366 DAOSTORM, CSpline, MIAtool, SMAP-2018). Gaussian fitting software (3D DAOSTORM, MIATool,

367 ThunderSTORM) gave robust nanoscale resolution images. Interestingly, experimental PSF fitting
368 software (CSpline, SMAP-2018) gave noticeably improved resolution of fine structural features such
369 as the top and bottom of the NPC (Fig 4B) or the hollow core of antibody-labelled microtubules (Fig
370 S15).

DISCUSSION

371

372 We performed the first broad evaluation of software for 3D single molecule localization microscopy,
373 to assess the state of the field and to allow non-specialists to determine the optimal software for their
374 experiments.

375 In order to provide a realistic assessment of 3D software performance we tested software on
376 simulations incorporating experimentally acquired microscope point spread functions. Our
377 experimental-PSF-derived simulation approach is readily adaptable to novel engineered 3D SMLM
378 PSFs[43] or to the PSF of individual microscopes. For instance, it would be possible to combine our
379 derived-PSF approach with the SMLM sample simulation tool SuReSim[44] in order to generate ultra-
380 realistic synthetic data, which could then be personalized to each experimentalists sample and
381 microscope, to easily determine the blocker factors to maximal resolution, for a given experiment.

382 The strongest conclusion we draw from the 3D localization microscopy challenge is that choice of
383 localization software greatly affects the quality of final super-resolution data, even at "easy" high SNR,
384 low spot density conditions. Biplane performance was particularly dependent on software choice, with
385 only one software (SMAP-2018) achieving near-Cramér-Rao lower bound performance. Double helix
386 SMLM showed less sensitivity to choice of software than biplane, with astigmatic SMLM intermediate
387 between the two. The best software in each modality performed close to the Cramér-Rao lower
388 bounds over a wide focal range and successfully detected most molecules, even at low signal to noise.
389 Average software in all three modalities was significantly worse, with the obtained axial resolution
390 being particularly sensitive to software choice.

391 The second major conclusion of the 3D challenge is that localization software that explicitly includes
392 the experimental PSF in the fitting model gives a significant performance increase for 3D SMLM. For
393 the more optically complex biplane and double helix modalities in particular, the best results were
394 exclusively from software using PSF modelling approaches (SMAP-2018, CSpline, MIATool). This result
395 also highlights the need for experimental PSF modelling not only in SMLM software, but also
396 emphasizes the high degree of experimental realism required of SMLM simulations. The clear
397 performance advantage of experimental PSF modelling software in the 3D software challenge would
398 have been unobservable had it been run with a simple Gaussian PSF.

399 Of the different algorithm classes, well-tuned single-emitter and multi-emitter fitting algorithms (each
400 capable of dealing well with occasional molecule overlap) gave good results for low density 3D SMLM.
401 We also found that several software packages for astigmatic or biplane imaging gave adequate
402 performance for the challenging case of high molecule densities, as long as the image SNR was high.
403 Current software packages gave poor performance when molecule density was high and image SNR
404 was low. These results suggest that, at least with current algorithms, high density 3D SMLM
405 performance is mediocre at high SNR, and poor at low SNR. Surprisingly, multi-emitter fitting did not
406 show significant improvement over well-tuned single emitter fitting for the 3D high-density datasets;
407 this may indicate that significant potential for improvement remains in this category.

408 Many software packages did not apply temporal grouping[40], resulting in reduced software
409 performance. Since temporal grouping is a simple step for maximum precision, we urge all software
410 developers to integrate this approach into their software as an optional final step in the localization
411 process.

412 The second 2D localization microscopy challenge provided the opportunity to reassess the state of the
413 field. The performance of best-in-class 2D software over a range of conditions, at both high and low

414 spot density, is excellent. The performance of the best-in-class software at high spot density (ADCG[37])
415 was only moderately decreased compared with the low spot density results, with nearly identical
416 molecule detection performance, and a 30 % increase in localization error. Interestingly, the top three
417 performers in the 2D high density condition were all compressed sensing algorithms (ADCG[37],
418 FALCON[36], SMfit). In low density 2D conditions, the best single-emitter, multi-emitter and compressed
419 sensing algorithms all gave comparable, excellent, performance. We speculate that performance in
420 this category may now be near optimal levels.

421 In future we plan to extend the SMLM challenge website and software into an open platform where
422 the assessment process is fully automated, and where new competition simulations and assessment
423 metrics can easily be created and contributed by the community. Scientific CMOS cameras are rapidly
424 becoming a major platform for single molecule localization microscopy[6] and it will be important to
425 include sCMOS simulations in future SMLM software assessments. Furthermore, there remain two
426 important classes of super-resolution microscopy for which software performance is crucial, but no
427 broad software assessment has yet been performed: fluorescence-fluctuation-based super-resolution
428 microscopies (*e.g.*, 3B[45], SOFI[46], SRRF[47]) and structured illumination microscopy[48].

429 The results of this competition clearly demonstrate the formidable algorithmic performance of the
430 best 2D and 3D localization microscopy software. However, a key outstanding challenge that often
431 hinders adoption of new algorithms is that only a small subset of algorithms are packaged in, or
432 compatible with fast, well-maintained, user-friendly software packages, which include all stages of the
433 SMLM data analysis pipeline – analysis, visualization and quantification. One solution would be for the
434 SMLM software community to collectively adopt both a standard data format and a single software
435 platform for future software development, such as FIJI/ ImageJ[49]. Any new algorithm released in this
436 environment could be immediately and widely adopted by users, and easily integrated into existing
437 packages for SMLM analysis, visualization and quantification.

438 Both the 3D and 2D localization challenges remain open and continuously updated on the competition
439 website. This continuously evolving analysis of state of the art super-resolution software performance
440 provides a valuable resource to super-resolution microscopists, helping to ensure they use software
441 that gets the best out of hard-won data. It also provides SMLM software developers with a robust
442 means of benchmarking new algorithms against current state of the art.

ACKNOWLEDGEMENTS

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.

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).

485   4. Holden, S. J., Uphoff, S. & Kapanidis, A. N. DAOSTORM: an algorithm for high- density super-

486   resolution microscopy. *Nat Meth* **8**, 279–280 (2011).

487   5. Huang, F., Schwartz, S. L., Byars, J. M. & Lidke, K. A. Simultaneous multiple-emitter fitting for single

488   molecule super-resolution imaging. *Biomed. Opt. Express* **2**, 1377–1393 (2011).

489   6. Huang, F. *et al.* Video-rate nanoscopy using sCMOS camera-specific single-molecule localization

490   algorithms. *Nat. Methods* **10**, 653–658 (2013).

491   7. Sage, D. *et al.* Quantitative evaluation of software packages for single-molecule localization

492   microscopy. *Nat. Methods* **12**, 717–724 (2015).

493   8. Huang, B., Jones, S. A., Brandenburg, B. & Zhuang, X. Whole-cell 3D STORM reveals interactions

494   between cellular structures with nanometer-scale resolution. *Nat Meth* **5**, 1047–1052 (2008).

495   9. Shtengel, G. *et al.* Interferometric fluorescent super-resolution microscopy resolves 3D cellular

496   ultrastructure. *Proc. Natl. Acad. Sci.* **106**, 3125–3130 (2009).

497   10.   Huang, B., Wang, W., Bates, M. & Zhuang, X. Three-Dimensional Super-Resolution Imaging by

498   Stochastic Optical Reconstruction Microscopy. *Science* **319**, 810–813 (2008).

499   11.   Pavani, S. R. P. *et al.* Three-dimensional, single-molecule fluorescence imaging beyond the

500   diffraction limit by using a double-helix point spread function. *Proc. Natl. Acad. Sci.* **106**, 2995–2999

501   (2009).

502   12.   Juette, M. F. *et al.* Three-dimensional sub–100 nm resolution fluorescence microscopy of thick

503   samples. *Nat. Methods* **5**, 527–529 (2008).

504   13.   Shtengel, G. *et al.* Interferometric fluorescent super-resolution microscopy resolves 3D

505   cellular ultrastructure. *Proc. Natl. Acad. Sci.* **106**, 3125–3130 (2009).

506   14.   Babcock, H., Sigal, Y. M. & Zhuang, X. A high-density 3D localization algorithm for stochastic

507   optical reconstruction microscopy. *Opt. Nanoscopy* **1**, 1–10 (2012).

508   15.   Ovesný, M., Křížek, P., Švindrych, Z. & Hagen, G. M. High density 3D localization microscopy

509   using sparse support recovery. *Opt. Express* **22**, 31263–31276 (2014).

510   16.   Min, J. *et al.* 3D high-density localization microscopy using hybrid astigmatic/ biplane imaging

511   and sparse image reconstruction. *Biomed. Opt. Express* **5**, 3935–3948 (2014).

512   17.   Zhang, S., Chen, D. & Niu, H. 3D localization of high particle density images using sparse

513   recovery. *Appl. Opt.* **54**, 7859–7864 (2015).

514   18.   Annibale, P., Vanni, S., Scarselli, M., Rothlisberger, U. & Radenovic, A. Quantitative Photo

515   Activated Localization Microscopy: Unraveling the Effects of Photoblinking. *PLOS ONE* **6**, e22678

516   (2011).

517   19.   Collaboration through competition. *Nat. Methods* **11**, 695 (2014).

518   20.   Li, Y. *et al.* Real-time 3D single-molecule localization using experimental point spread

519   functions. *Nat. Methods* (2018). doi:10.1038/nmeth.4661

520   21.   Izeddin, I. *et al.* PSF shaping using adaptive optics for three-dimensional single-molecule

521   super-resolution imaging and tracking. *Opt. Express* **20**, 4957–4967 (2012).

522   22.   Grover, G., DeLuca, K., Quirin, S., DeLuca, J. & Piestun, R. Super-resolution photon-efficient

523   imaging by nanometric double-helix point spread function localization of emitters (SPINDLE). *Opt.*

524   *Express* **20**, 26681–26695 (2012).

525   23.   Carlini, L. & Manley, S. Live Intracellular Super-Resolution Imaging Using Site-Specific Stains.

526   *ACS Chem. Biol.* **8**, 2643–2648 (2013).

527   24.   Shim, S.-H. *et al.* Super-resolution fluorescence imaging of organelles in live cells with

528   photoswitchable membrane probes. *Proc. Natl. Acad. Sci.* **109**, 13978–13983 (2012).

529   25.   Baddeley, D. & Bewersdorf, J. Biological Insight from Super-Resolution Microscopy: What We

530   Can Learn from Localization-Based Images. *Annu. Rev. Biochem.* **87**, 965–989 (2018).

531   26.   Loot A. , Valdmann A., Eltermann M., Kree M., Pärs M. SMolPhot Software. Available at:

532   https://bitbucket.org/ardiloot/smolphot-software/wiki/Home. (Accessed: 29th November 2018)

533   27.   Babcock, H. P. & Zhuang, X. Analyzing Single Molecule Localization Microscopy Data Using

534   Cubic Splines. *Sci. Rep.* **7**, 552 (2017).

535   28.   Fox-Roberts, P. *et al.* Local dimensionality determines imaging speed in localization

536   microscopy. *Nat. Commun.* **8**, 13558 (2017).

537   29.   Henriques, R. *et al.* QuickPALM: 3D real-time photoactivation nanoscopy image processing in

538   ImageJ. *Nat Meth* **7**, 339–340 (2010).

539   30.   Takeshima, T., Takahashi, T., Yamashita, J., Okada, Y. & Watanabe, S. A multi-emitter fitting

540   algorithm for potential live cell super-resolution imaging over a wide range of molecular densities.

541   *J. Microsc.* **271**, 266–281 (2018).

542   31.   Kechkar, A., Nair, D., Heilemann, M., Choquet, D. & Sibarita, J.-B. Real-Time Analysis and

543   Visualization for Single-Molecule Based Super-Resolution Microscopy. *PLOS ONE* **8**, e62918 (2013).

544   32.   Alex   Herbert.   PeakFit   Software.   Available   at:

545   http://www.sussex.ac.uk/gdsc/intranet/microscopy/UserSupport/AnalysisProtocol/imagej/smlm

546   _plugins/. (Accessed: 29th November 2018)

547   33.   Ovesný, M., Křížek, P., Borkovec, J., Švindrych, Z. & Hagen, G. M. ThunderSTORM: a

548   comprehensive ImageJ plug-in for PALM and STORM data analysis and super-resolution imaging.

549   *Bioinformatics* **30**, 2389–2390 (2014).

550   34.   Soubies, E., Blanc-Féraud, L. & Aubert, G. A Continuous Exact ℓ_0 Penalty (CEL0) for Least

551   Squares Regularized Problem. *SIAM J. Imaging Sci.* **8**, 1607–1639 (2015).

552   35.   Babcock, H. P., Moffitt, J. R., Cao, Y. & Zhuang, X. Fast compressed sensing analysis for super-

553   resolution imaging using L1-homotopy. *Opt. Express* **21**, 28583–28596 (2013).

554   36.   Min, J. *et al.* FALCON: fast and unbiased reconstruction of high-density super-resolution

555   microscopy data. *Sci. Rep.* **4**, 4577 (2014).

556   37.   Boyd, N., Schiebinger, G. & Recht, B. The Alternating Descent Conditional Gradient Method

557   for Sparse Inverse Problems. *SIAM J. Optim.* **27**, 616–639 (2017).

558   38.   Huang, J., Sun, M. & Chi, Y. Super-resolution image reconstruction for high-density 3D single-

559   molecule microscopy. in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*

560   241–244 (2016). doi:10.1109/ISBI.2016.7493254

561   39.   Pan, H., Simeoni, M., Hurley, P., Blu, T. & Vetterli, M. LEAP: Looking beyond pixels with

562   continuous-space EstimAtion of Point sources. *Astron. Astrophys.* **608**, A136 (2017).

563   40.   Durisic, N., Laparra-Cuervo, L., Sandoval-Álvarez, Á., Borbely, J. S. & Lakadamyali, M. Single-

564   molecule evaluation of fluorescent protein photoactivation efficiency using an in vivo

565   nanotemplate. *Nat. Methods* **11**, 156–162 (2014).

566   41.   Chao, J., Ward, E. S. & Ober, R. J. A software framework for the analysis of complex microscopy

567   image data. *IEEE Trans. Inf. Technol. Biomed. Publ. IEEE Eng. Med. Biol. Soc.* **14**, 1075–1087 (2010).

568   42.   Martens, K. J. A., Bader, A. N., Baas, S., Rieger, B. & Hohlbein, J. Phasor based single-molecule

569   localization microscopy in 3D (pSMLM-3D): An algorithm for MHz localization rates using standard

570   CPUs. *J. Chem. Phys.* **148**, 123311 (2017).

571   43.   Shechtman, Y., Weiss, L. E., Backer, A. S., Sahl, S. J. & Moerner, W. E. Precise Three-

572   Dimensional Scan-Free Multiple-Particle Tracking over Large Axial Ranges with Tetrapod Point

573   Spread Functions. *Nano Lett.* **15**, 4194–4199 (2015).

574   44.   Venkataramani, V., Herrmannsdörfer, F., Heilemann, M. & Kuner, T. SuReSim: simulating

575   localization microscopy experiments from ground truth models. *Nat. Methods* **13**, 319–321 (2016).

576   45.   Cox, S. *et al.* Bayesian localization microscopy reveals nanoscale podosome dynamics. *Nat.*

577   *Methods* **9**, 195–200 (2012).

578   46.   Dertinger, T., Colyer, R., Iyer, G., Weiss, S. & Enderlein, J. Fast, background-free, 3D super-

579   resolution optical fluctuation imaging (SOFI). *Proc. Natl. Acad. Sci.* **106**, 22287–22292 (2009).

580   47.   Gustafsson, N. *et al.* Fast live-cell conventional fluorophore nanoscopy with ImageJ through

581   super-resolution radial fluctuations. *Nat. Commun.* **7**, (2016).

582   48.   Gustafsson, M. G. L. Surpassing the lateral resolution limit by a factor of two using structured

583   illumination microscopy. SHORT COMMUNICATION. *J. Microsc.* **198**, 82–87 (2000).

584   49.   Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods*

585   **9**, 676–682 (2012).

586   50.   Hanser B. M., Gustafsson M. G. L., Agard D. A. & Sedat J. W. Phase-retrieved pupil functions in

587   wide-field fluorescence microscopy. *J. Microsc.* **216**, 32–48 (2004).

588   51.   Hirsch, M., Wareham, R. J., Martin-Fernandez, M. L., Hobson, M. P. & Rolfe, D. J. A Stochastic

589   Model for Electron Multiplication Charge-Coupled Devices – From Theory to Practice. *PLOS ONE* **8**,

590   e53671 (2013).

591   52.   Basden, A. G., Haniff, C. A. & Mackay, C. D. Photon counting strategies with low-light-level

592   CCDs. *Mon. Not. R. Astron. Soc.* **345**, 985–991 (2003).

593   53.   Carlini, L., Holden, S. J., Douglass, K. M. & Manley, S. Correction of a Depth-Dependent Lateral

594   Distortion in 3D Super-Resolution Imaging. *PLoS ONE* **10**, e0142949 (2015).

595
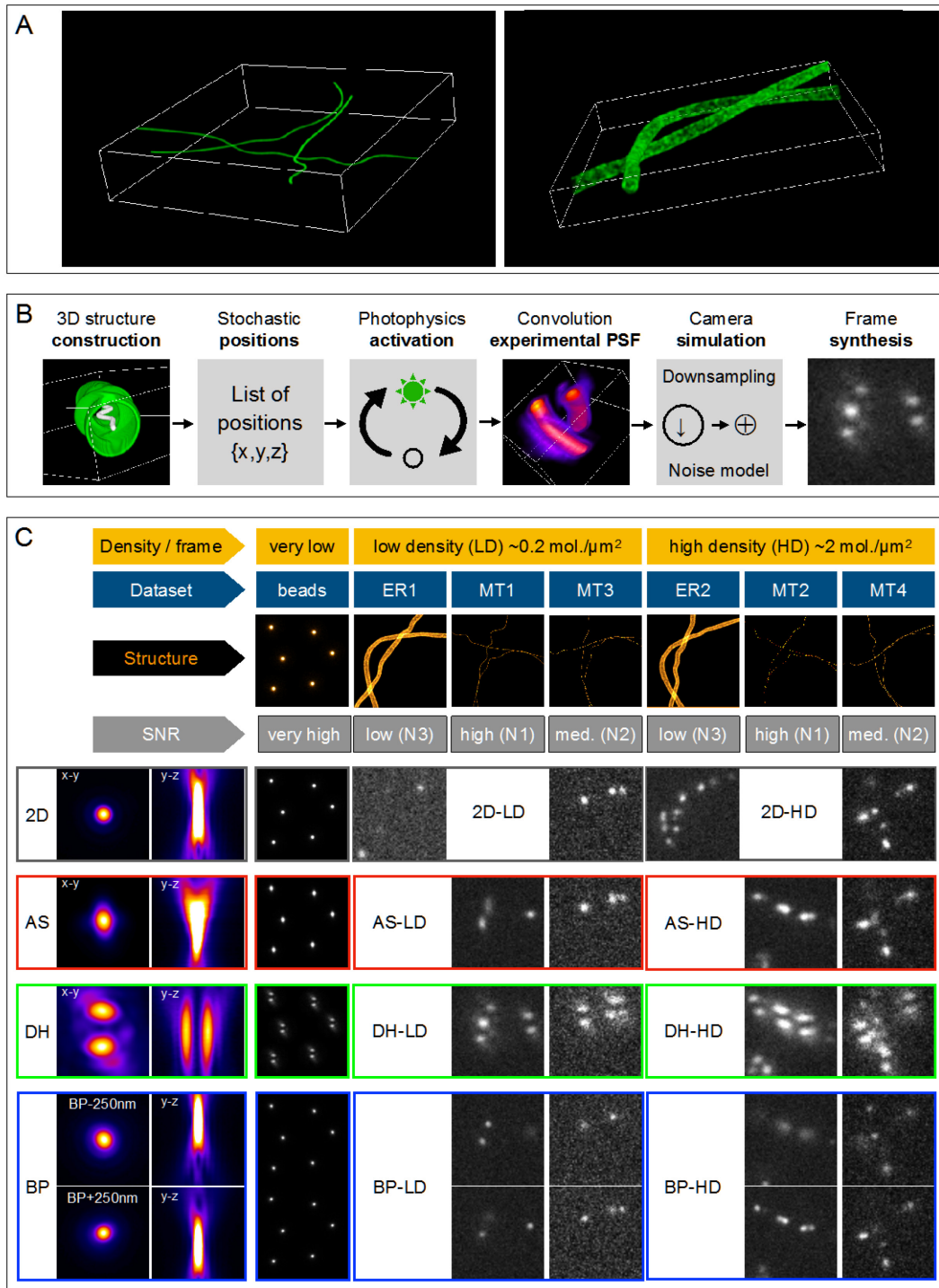
596
597 **Figure 1**: *Summary of SMLM challenge simulations*. **A**. 3D rendering of microtubules and endoplasmic
598 reticulum samples in a 6.4 μm x 6.4 μm x 1.5 μm volume. **B**. *Key simulation steps.* The structure is
599 constructed from 3D tubes continuously defined by three B-spline functions in the volume of interest.
600 Membranes of the tubes are densely populated with possible positions. Fluorophores follow a 4-state
601 photophysics model. Activations of a given frame are convolved with the experimental PSF and shot
602 & camera noise is added. **C.** Summary of all 16 challenge datasets, calibration data and experimental
603 PSFs. Each dataset is characterized by its structure (endoplasmic reticulum (ER) or microtubules (MT)),
604 by it modality (2D, AS, DH, BP), its density (LD or HD) and by its SNR determined by the level of noise
605 N1, N2, and N3. Left column: orthogonal projections of the experimentally-derived PSF. Eight
606 categories were proposed for the challenge containing two datasets each, 2D-LD and 2D-HD, grey; AS-
607 LD and AS-HD, red; DH-LD and DH-HD, green; BP-LD and BP-HD, blue.
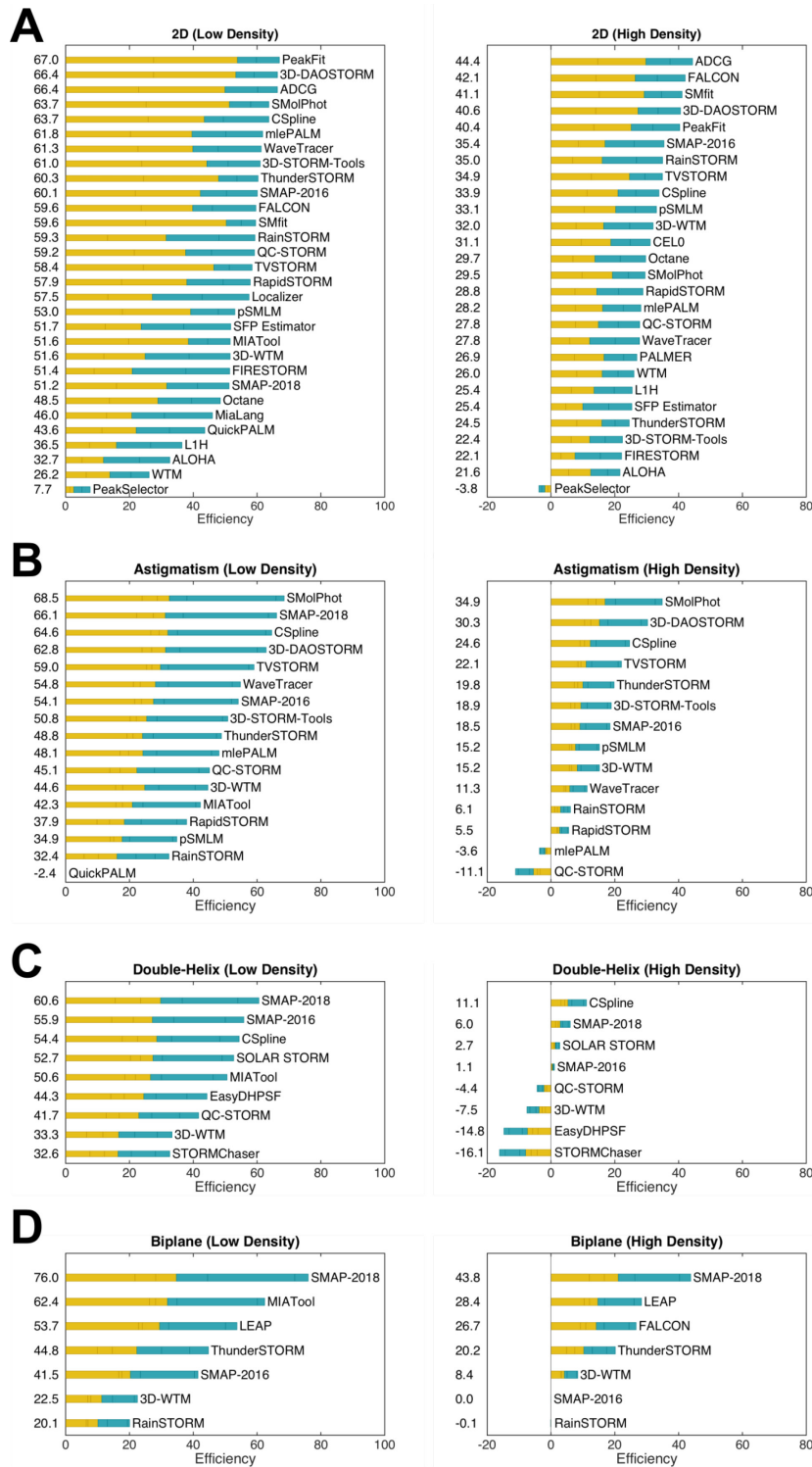
Figure 2: *Leaderboards for each competition modality, at low and high spot density*. Ranking is based on the efficiency of software based on fraction of successfully detected molecules (Jaccard index) and precision of localization (RMSE, root mean square error, lateral & axial). The contribution of the high SNR dataset is plotted in orange and the contribution of the low SNR dataset to the efficiency is plotted in blue.
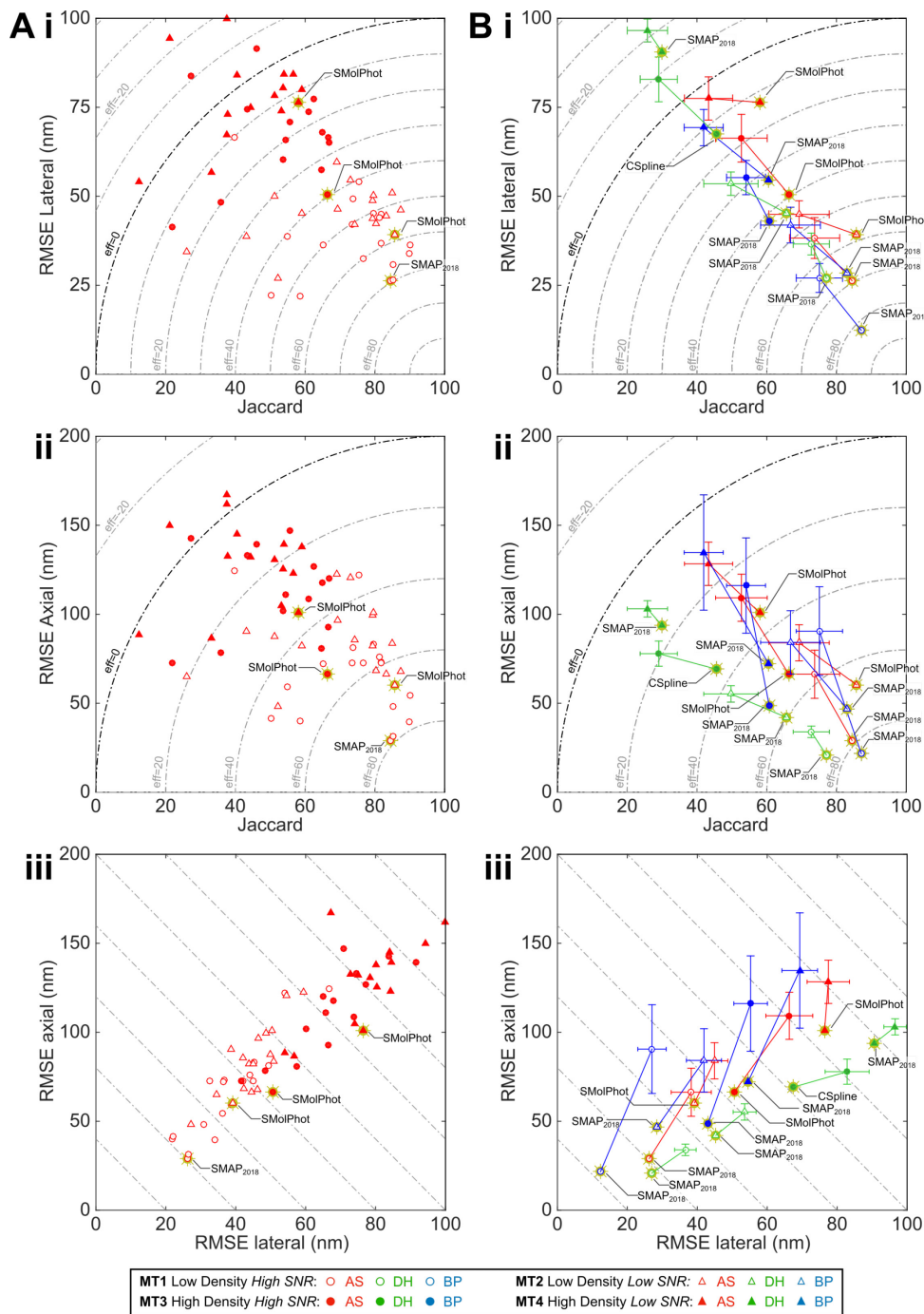
Figure 3: *Comparison of 3D software performance*. Gold stars indicate top performers for each dataset. Dashed lines in top, middle panels indicate overall efficiency (higher is better). **A.** Localization error and spot detection performance of all astigmatic SMLM software. **B.** Average (colored marker with *s.d.* error bars) and best-in-class (colored marker with gold star) software performance for all competition modalities. *AS, astigmatism; DH, double helix; BP, biplane.*
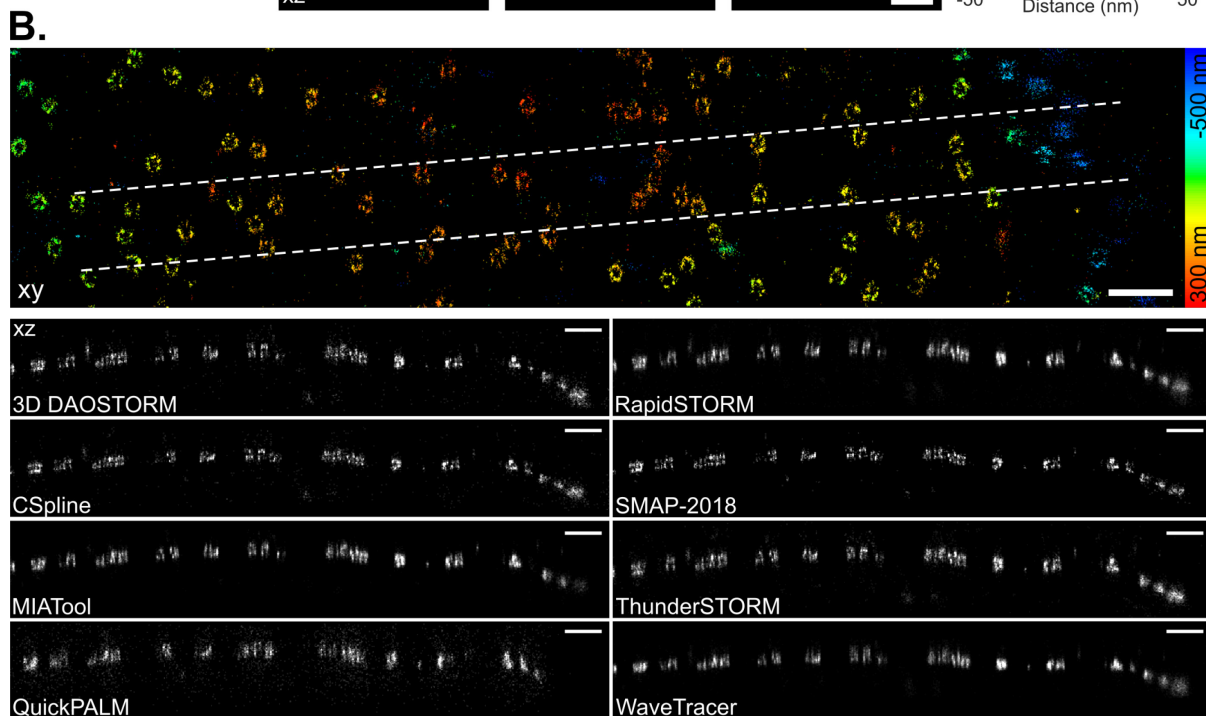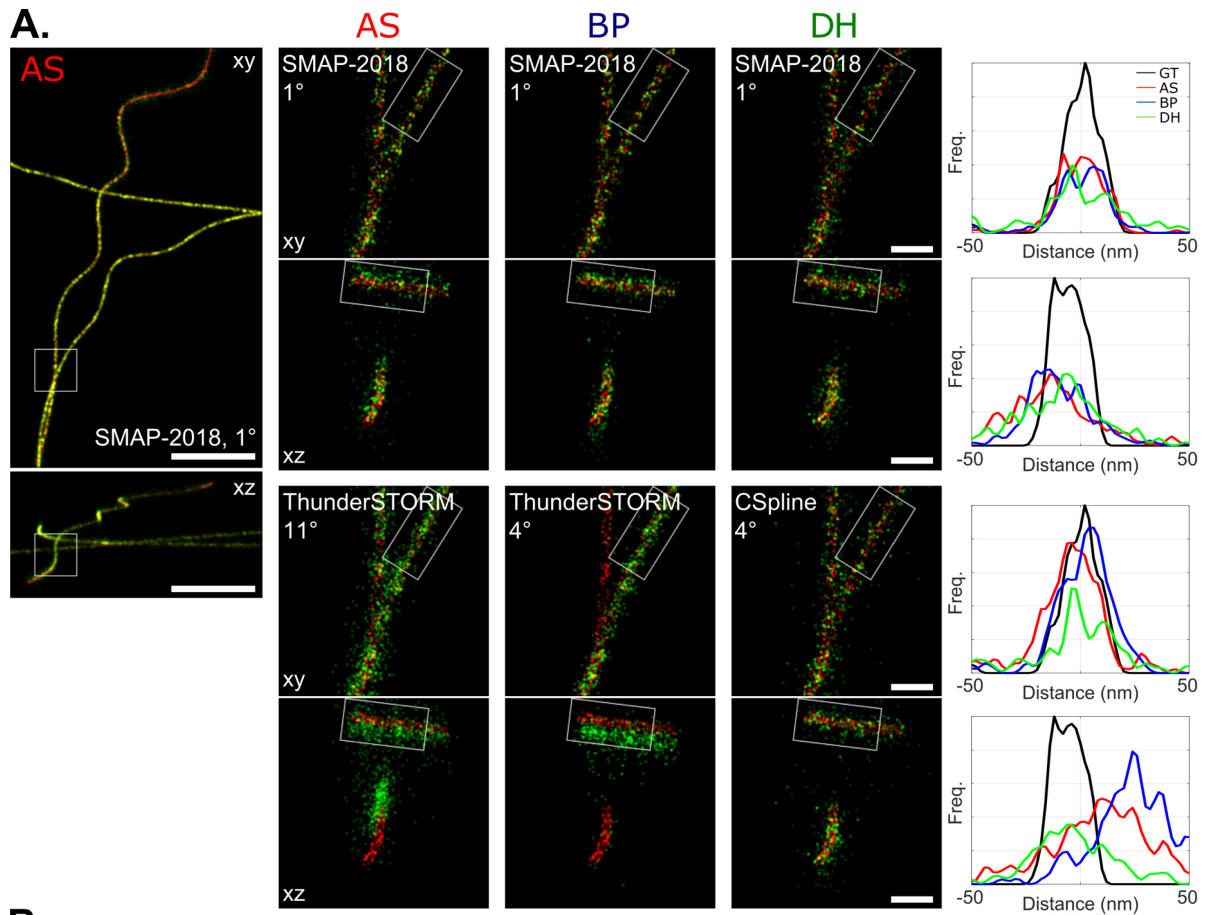
Figure 4: *A. Super-resolved xy and xz projection images of 3D competition datasets for best-in-class (top) and representative average (bottom) software in each modality, for high SNR low density dataset. Left*: *xy* and *xz* overview images for winning AS software. *Middle*: *xy* and *xz* zoom images 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;

627   BP, biplane, blue; DH, double helix, green. *Panel key:* Software-name Dataset-Ranking°. *Scale bar*: full

628   image, 1 μm, magnified regions, 100 nm. **B. Performance of astigmatism software on real 3D STORM**

629   *images of the nuclear pore complex. Top:* Super-resolved overview image in *Xy* for 3D-DAOSTORM

630   software, color coded for depth. *Bottom: Xz* orthoslices along 600 nm wide dashed region indicated

631   in top panel for 8 astigmatism software packages. The best performing software clearly resolves the

632   top and bottom of the NPCs. *Scale bar*s, 500 nm.

633

634   *METHODS*

635   ## 1. CHALLENGE ORGANIZATION

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

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

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

645   ## 2. LOCALIZATION MICROSCOPY SIMULATIONS

646   ### 2.1. Structure

647   The synthetic datasets were designed to be similar to images derived from cellular structures in real
648   experimental conditions. We defined mathematical models for cellular structures that imitate
649   cytoskeletal filaments such as microtubules and larger tubular structures such as the endoplasmic
650   reticulum or mitochondria (Fig S1A). These structures have a tubular shape in the 3D space. Psuedo-
651   microtubules are defined with their central axis elongating in a 3D space having an average outer
652   diameter of 25 nm with an inner, hollow tube of 15 nm diameter. Pseudo-endoplasmic reticulum is
653   defined as having a diameter of approximately 150nm.

654   The underlying sample structure is formalized in a continuous space which allows rendering of digital
655   images at any scale, from very high resolution (up to 1 nm/pixel) to low resolution (camera resolution:
656   100 nm/ pixel). The continuous-domain 3D curve is represented by means of a polynomial spline. The
657   sample is imaged in a $6.4 \times 6.4 \ \mu m^2$ field of view, and the center lines of the microtubules have limited
658   variation along the z (vertical) axis, *i.e.*, less than 1.5 μm. The fluorescent markers are uniform
659   randomly distributed over the structure according to the required density. The photon emission rate
660   of each fluorophore is controlled by a photo-activation model (see below).

661   The exact locations of all fluorophores are stored at high precision floating-point numbers expressed
662   in nanometers. This ground-truth file is useful for conducting objective evaluations without human
663   bias.

664   ### 2.2. Photophysics activation model

665   Given a list of source locations from the structure simulator, fluorophore blinking was simulated by a
666   4-states Markov chain model. The states are ON, OFF, BLEACH, DARK and the transition are Poisson
667   distributed (Fig S1C), except for the OFF to ON transitions which follow a uniform random distribution,
668   to reflect that in typical experimental conditions, constant imaging density is maintained by tuning the
669   photoactivation rate during the experiment. All switching is calculated at sub-frame resolution and
670   then total fluorophore on-time was integrated over each frame.

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

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

672

673   Switching rates were chosen to approximate photoactivatable fluorescent proteins T_{ON} = 3 frame, T_{DARK}
674   = 2.5 frames, and T_{BLEACH}=1.5 frames.

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

677 $$\mathbf{F} = \frac{\phi \cdot \mathrm{P} \cdot \sigma}{e}$$

678 Φ is the quantum yield of the dye, P is power of the laser in W/cm², e = h c / λ is the energy of one
679 photon, σ = 1000 ln(10) ε / N_A is the absorption cross section in cm² and ε is the molar extinction
680 coefficient (EC) or absorptivity in cm²/mol which is a characteristic of a given fluorophore. The laser
681 power was Gaussian distributed over the field of view. At the end of this process a list of XY positions,
682 on-frames and (noise-free) intensities for all activated fluorophores was obtained.

683 Analysis of the resulting simulated photon counting distribution is presented in Supplementary
684 Note 1.

2.3. Experimental Point-Spread Function

686 Model PSFs, stored as high resolution look up tables, were derived from experimentally measured
687 PSFs. Although the algorithmic approach is distinct, this concept of accurately modelling the
688 experimental PSF based on calibration data bears relation to the PSF phase retrieval approach
689 previously employed by Hanser and coworkers[50].

690 Images of fluorescent beads were recorded for each modality (Table S4). Signal to noise ratio of
691 recorded PSFs was maximized in all cases by maximizing exposure time and averaging over several
692 frames to increase dynamic range.

693 To acquire experimental PSFs, we took 100 nm Tetraspek beads (Invitrogen) adsorbed to #1.5 (170 µm
694 thick) coverglass, imaged in water. The excitation wavelength was between 640 nm and 647 nm, and
695 a Cy5 emission filter was used. Data acquisition parameters for each modality are listed in Table S4.

2.4. Simulation PSF construction

697 For each modality, 3-6 beads were selected within a small (< 32 µm) region, to minimize PSF variation
698 due to spherical aberration. Images for each selected bead were interpolated in XY to a pixel size of
699 10 nm. Beads were then coaligned by cross-correlation on the in-focus frame. Coaligned beads were
700 averaged in XY to minimize pixel quantization artefacts and to increase SNR. Where necessary, Z-stacks
701 were interpolated to a Z-step size of 10 nm. A central Z-range of 1.5 µm was selected that represents
702 151 optical planes with a Z-step of 10 nm. The Z-range covers -750 nm to +750 nm. The plane of best
703 focus was chosen as the simulation 0 nm plane. Each model PSF was normalized such that the total
704 intensity of the PSF in the in-focus frame within a diameter of 3 FWHM from the PSF center was equal
705 to 1.

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

708 In biplane super-resolution microscopy, emitted fluorescence is split into two simultaneously imaged
709 channels, with a small (500-1000 nm) defocus introduced between the two channels[12]. As the small
710 defocus should introduce minimal additional aberration into an optical system, we semi-synthetically
711 constructed a realistic biplane PSF from the experimental 2D PSF. The two defocused PSFs were
712 constructed by duplicating the 2D PSF and offsetting it by -250 nm and 250 nm for each Z-plane.

713 This yielded five high SNR model PSFs with an isotropic voxel size of 10x10x10 nm³.

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

2.5. Noise model

718
719 A constant mean autofluorescent background was added to the noise-free simulated images, and
720 these images were then fed through the noise model representing Poisson distributed fluorescence
721 emission recorded on a high quantum efficiency back-illuminated EMCCD[51,52].

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

723 • σ_S , the shot noise produced by the fluorescence background and signal and the spurious
724 charge. Shot noise can be derived from the second moment of the Poisson distribution
725 • σ_R, the read noise of EMCCD camera, which is described by second moment of the Gaussian
726 distribution
727 • σ_{EM}, the electron multiplication noise introduced by the gain process, which is described by
728 the second moment of the Gamma distribution[52].
729

730 We assumed as camera parameters the ones specified for the Photometrics Evolve Delta 512 EMCCD
731 camera:

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

740 The final simulated photon electrons will thus be given by:

741
$$n_{ie} \; = \; \mathcal{P}(QE \cdot n_{photIn} + c)$$

742
$$n_{oe} \; = \; \Gamma\big(n_{ie}, EM_{gain}\big) + \mathcal{G}(0, \sigma_R)$$

743 which leads to the final pixel counts:

744
$$ADU_{out} \; = \; min\left(\frac{n_{oe} - n_{oe} \, mod \; e_{ADU}}{e_{per_{adu}}} + BL, 65535\right)$$

2.6. Depth-dependent lateral distortion: Wobble

745
746 As the PSF models are experimentally derived, the 3D estimated localizations exhibit a depth-
747 dependent lateral distortion, here called *wobble*. This optical distortion is due to a combination of a
748 systematic offset (arbitrary definition of PSF center) and optical aberrations[53]. In order to compare
749 estimated and true localizations, we correct this effect during the assessment (Section 3.1).

2.7 Comparison of software results between different modalities.

750
751 The intensities of the PSF in each imaging modality were normalized to facilitate comparison of results
752 between different modalities. Software results between 2D, 3D AS and 3D DH modalities are expected
753 to be directly comparable.

754 For the biplane model PSF, as the emitted fluorescence is split into two channels, the intensity in each
755 of the two simulated biplane channels was additionally reduced by 50 %. We note that the
756 fluorescence background was not reduced by 50 % as intended, leading to artificially high background
757 for the biplane simulation. *I.e.*, the background in each of the two biplane channels is the same as in
758 the single channel of the other modalities. However, due to the low background level in the 3D
759 simulations, the effect on image SNR and thus localization error is small (see Fig S7), less than 5 nm

760 near the plane of focus. Therefore, as long as the small drop in image SNR is taken into account,
761 approximate comparisons of the biplane data to the other modalities can still be made.

3. SOFTWARE ASSESSMENT

3.1 Protocol

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

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

777 The pairing between true and estimated localizations was performed frame by frame. The procedure
778 matches two sets of localizations. We deployed the presorted nearest-neighbor search for its
779 efficiency, with a linking threshold of 250 nm. The results are effectively similar to the computationally
780 intensive Hungarian algorithm[7].

Photon thresholding

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

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

Wobble correction

793 The centroid of experimental point spread functions shifts laterally by as much as 50 nm, as a function
794 of axial position[10,53]. This is most often ignored by localization software, and instead corrected post-
795 hoc by reference to a calibration curve[37]. Since our simulated PSF is experimentally derived, it was
796 necessary to correct for these artefactual shifts between the observed localizations and ground truth,
797 as part of the assessment process. This correction was performed using calibration data uploaded by
798 competitors, similar to the correction typically performed on experimental data[53].

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

804 In certain specific cases (identified on the competition website), at the request of authors, we did not
805 apply this correction, for example because the software explicitly considered the whole 3D PSF during
806 fitting and was thus immune to this lateral shift artefact. For accurate results, application of lateral
807 shift correction is critical for analysis of localization microscopy simulations using experimentally
808 derived PSFs, as can be seen by comparison of typical software results with and without wobble
809 correction (Fig S11).

3.2 Metrics

811 We calculated a large number of analysis metrics to quantify the performance of software relative to
812 ground truth. These are discussed in detail in Supplementary Note 1. The metrics are split into two
813 categories: localization based and image based metrics.

814 The former directly relies on the localizations positions and notably includes the Recall, the Precision,
815 the Jaccard Index, the RMSE (axial and lateral) and the consolidated Z-range. For the calculation of
816 average software performance (Fig 3B) outlier software with an efficiency less than *eff=-30* were
817 excluded from the measurement.

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

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

827 In order to summarize software axial performance, we analyzed how the recall varied as a function of
828 Z. A typical recall versus axial position curve (Fig S9) will drop at positions far from the focal plane,
829 *i.e.,* where software can no longer detect spots to defocus. We first smoothed the curve using a sliding
830 window. Then we computed the software Z-range, defined as the full width half maximal Recall of the
831 smoothed curve (Fig S12). This quantity is visually intuitive and useful for discussion of the recall
832 performance if considered alongside a plot of recall vs axial position. However, because FHWM recall
833 depends on the maximal recall, ranking based on this procedure would promote a software which
834 poorly performed everywhere (*i.e.,* flat curve), whereas a software which performed well in the focal
835 plane but less well outside would obtain a worse FWHM recall. This observation leads us to produce
836 a so-called consolidated Z-range, by multiplying the Z-range value by the maximal Recall, which should
837 provide a robust metric that avoids the previous case scenario.

838 *Principal component analysis.* In order to analyse the relationship between analysis metrics we
839 computed the covariance matrix between each metric and the principal component analysis (PCA) on
840 the metrics (Fig S14B). Each metric was standardized before applying the covariance and the PCA. For
841 convenience, we took the additive inverse of the metrics for which lower values are best (*i.e.,* FP, FN,
842 RMSE, FRC, FSC).

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

3.3 Baseline Localization Software

847 We developed a minimalist Java tool software that performs localizations of bright emitters on the 4
848 modalities of the challenge 2016: 2D, Astigmatism, Double-Helix, and Biplane. This

849 SMLM_BaselineLocalization software is only designed to establish the performance baseline for the
850 SMLM challenge. It has intentionally limited lines of code and relies only on few threshold parameters
851 to localize particles. It has basic calibration tool that has to run on a z-stack of beads to find the linear
852 f(x) relation between the axial position Z and the shape of the bead.

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

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

4 REAL DATA ASSESSMENT

860 Astigmatism software was tested on previously published real 3D STORM datasets of microtubules
861 and nuclear pore complex[20]. The tubulin dataset corresponds to the raw data for Fig S6 in Ref 20, and
862 the nuclear pore complex dataset corresponds to raw data for Fig S9 in Ref 20. Key acquisition
863 parameters for data analysis are summarized on the competition website.

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

5 DATA AVAILABILITY

5.1 Data availability statement

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

5.2 Code availability statement

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