Tractography-based connectomes are dominated by false-positive connections

Klaus H. Maier-Hein*[,a], Peter Neher[a], Jean-Christophe Houde[b], Marc-Alexandre Côté[b], Eleftherios Garyfallidis[b], Jidan Zhong[g], Maxime Chamberland[b], Fang-Cheng Yeh[h], Ying-Chia Lin[i], Qing Ji[j], Wilburn E. Reddick[j], John O. Glass[j], David Qixiang Chen[k], Yuanjing Feng[l], Chengfeng Gao[l], Ye Wu[l], Jieyan Ma[m], H Renjie[m], Qiang Li[m,n], Carl-Fredrik Westin[o], Samuel Deslauriers-Gauthier[b], J. Omar Ocegueda González[p], Michael Paquette[b], Samuel St-Jean[b], Gabriel Girard[b], François Rheault[b], Jasmeen Sidhu[b], Chantal M.W. Tax[r], Fenghua Guo[r], Hamed Y. Mesri[r], Szabolcs Dávid[r], Martijn Froeling[s], Anneriet M. Heemskerk[r], Alexander Leemans[r], Arnaud Boré[q], Basile Pinsard[q,zg], Christophe Bedetti[q,zh], Matthieu Desrosiers[q], Simona Brambati[q], Julien Doyon[q], Alessia Sarica[t], Roberta Vasta[t], Antonio Cerasa[t], Aldo Quattrone[u], Jason Yeatman[v], Ali R. Khan[w], Wes Hodges[x], Simon Alexander[x], David Romascano[d], Muhamed Barakovic[d], Anna Auría[d], Oscar Esteban[zd], Alia Lemkaddem[d], Jean-Philippe Thiran[d,ze], H. Ertan Cetingul[y], Benjamin L. Odry[y], Boris Mailhe[y], Mariappan S. Nadar[y], Fabrizio Pizzagalli[z], Gautam Prasad[z], Julio E. Villalon-Reina[z], Justin Galvis[z], Paul M. Thompson[z], Francisco De Santiago Requejo[za], Pedro Luque Laguna[za], Luis Miguel Lacerda[za], Rachel Barrett[za], Flavio Dell'Acqua[za], Marco Catani[za], Laurent Petit[zb], Emmanuel Caruyer[e], Alessandro Daducci[d], Tim B. Dyrby[f,zf], Tim Holland-Letz[zc], Claus C. Hilgetag[zi], Bram Stieltjes[c], Maxime Descoteaux*[,b]

*indicates corresponding authors.

a. Medical Image Computing Group (MIC), German Cancer Research Center (DKFZ), Heidelberg, Germany
b. Sherbrooke Connectivity Imaging Lab (SCIL), Université de Sherbrooke, Sherbrooke, Quebec, Canada
c. University Hospital Basel, Radiology & Nuclear Medicine Clinic, Basel, Swizterland.
d. Signal Processing Lab (LTS5), Ecole Polytechnique Federale de Lausanne, Switzerland
e. Centre national de la recherche scientifique (CNRS), Institute for Research in IT and Random Systems (IRISA), UMR 6074 VISAGES project-team, Rennes, France
f. Danish Research Centre for Magnetic Resonance, Center for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital Hvidovre, Hvidovre, Denmark
g. Krembil Research Institute, University Health Network, Toronto, Canada
h. Department of Psychology, Carnegie Mellon University, USA
i. IMT - Institute for Advanced Studies, Lucca, Italy
j. Department of Diagnostic Imaging, St. Jude Children's Research Hospital, Memphis, USA
k. University of Toronto Institute of Medical Science, Toronto, Canada
l. Institute of Information Processing and Automation, Zhejiang University of Technology, Hangzhou, Zhejiang, China
m. United Imaging Healthcare Co., Shanghai, China
n. Shanghai Advanced Research Institute, Shanghai, China
o. Laboratory of Mathematics in Imaging, Harvard Medical School, Boston, MA, United States
p. Center for Research in Mathematics, Guanajuato, Mexico
q. Centre de recherche institut universitaire de geriatrie de Montreal (CRIUGM), Université de Montréal, Montreal, Quebec, Canada
r. PROVIDI Lab, Image Sciences Institute, University Medical Center Utrecht, Utrecht, Netherlands
s. Department of Radiology, University Medical Center Utrecht, Utrecht, Netherlands
t. Neuroimaging Unit, Institute of Bioimaging and Molecular Physiology (IBFM), National Research Council (CNR), Policlinico Magna Graecia, Germaneto (CZ), Italy
u. Institute of Neurology, University Magna Graecia, Germaneto (CZ), Italy

v. Institute for Learning & Brain Sciences and Department of Speech & Hearing Sciences, University of Washington, Seattle, WA, USA
w. Departments of Medical Biophysics & Medical Imaging, Schulich School of Medicine and Dentistry, Western University, 1151 Richmond St N, London, Ontario, Canada
x. Synaptive Medical Inc, MaRS Discovery District, 101 College Street, Suite 200, Toronto, Ontario, Canada.
y. Medical Imaging Technologies, Siemens Healthcare, Princeton, NJ, USA
z. Imaging Genetics Center, Stevens Neuro imaging and Informatics Institute, Keck School of Medicine of USC, Marina del Rey, CA, USA
za. NatBrainLab, Institute of Psychiatry, Psychology & Neuroscience, King's College London, UK
zb. Groupe d'imagerie Neurofonctionnelle - Institut des Maladies Neurodégénératives (GIN-IMN), UMR5293 CNRS, CEA, University of Bordeaux, Bordeaux, France
zc. Biostatistics, German Cancer Research Center (DKFZ), Heidelberg, Germany
zd. Biomedical Image Technologies (BIT), ETSI Telecom., U. Politécnica de Madrid and CIBER-BBN, Spain
ze. Department of Radiology, University Hospital Center (CHUV) and University of Lausanne (UNIL), Lausanne, Switzerland
zf. Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kongens Lyngby, Denmark
zg. Sorbonne Universités, UPMC Univ Paris 06, CNRS, INSERM, Laboratoire d'Imagerie Biomédicale (LIB), 75013, Paris, France
zh. Center for Advanced Research in Sleep Medicine, Hôpital du Sacré-Coeur de Montréal, Montreal, Canada
zi. Department of Computational Neuroscience, University Medical Center Eppendorf, Hamburg, Germany

Abstract

Fiber tractography based on non-invasive diffusion imaging is at the heart of connectivity studies of the human brain. To date, the approach has not been systematically validated in ground truth studies. Based on a simulated human brain dataset with ground truth white matter tracts, we organized an open international tractography challenge, which resulted in 96 distinct submissions from 20 research groups. While most state-of-the-art algorithms reconstructed 90% of ground truth bundles to at least some extent, on average they produced four times more invalid than valid bundles. About half of the invalid bundles occurred systematically in the majority of submissions. Our results demonstrate fundamental ambiguities inherent to tract reconstruction methods based on diffusion orientation information, with critical consequences for the approach of diffusion tractography in particular and human connectivity studies in general.

Fiber tractography, a computational reconstruction method based on diffusion-weighted magnetic resonance imaging (DWI), attempts to reveal the white matter pathways of the human brain *in vivo* and to infer the underlying network structure, the structural connectome (1). Numerous algorithms for tractography have been developed and applied to connectome research in the fields of neuroscience (2,3) and psychiatry (4). Given the broad interest in this approach, the advantages and shortcomings of tractography have been widely debated (1,5–9). Particularly, *in vivo* tractography of the human brain was evaluated by subjective assessment of plausibility (10) or qualitative visual agreement with *post-mortem* Klingler-like dissections (11,12). Reproducibility (13) or data prediction errors (14–16) were evaluated in the context of tractographic model verification. However, these evaluations cannot validate the accuracy of reconstructions (17). Moreover, *ex vivo* imaging and tracing (17–25) or physically (26–32) and numerically simulated phantoms (33–37) allow validation to some extent. The complexity of the nervous system, however, and the lack of precise ground truth information on the trajectories of pathways and their origins as well as terminations in the human brain make it hard to quantitatively evaluate tractography results and to determine which discoveries are reliable when regarding brain connectivity in health and disease.

State-of-the-art tractography algorithms are driven by *local orientation fields* estimated from DWI, representing the local tangent direction to the white matter tract of interest (1). Conceptually, the principle of inferring connectivity from local orientation fields can lead to problems as soon as pathways overlap, cross, branch and have complex geometries (Figure 1) (8,38,39). Since the invention of diffusion tractography, these problems have been discussed in schematic representations or theoretical arguments (8,9,40), but have not yet been quantified in brain imaging. To determine the current state of the art in tractography, we organized an international tractography competition (tractometer.org/ismrm_2015_challenge) and employed a novel validation method based on simulated DWI of a brain-like geometry. This ground truth data set represented 25 well-known valid bundles that covered approximately 70% of the human brain white matter.

At the closing of the competition, we evaluated 96 distinct tractography pipelines submitted by 20 different research groups, in order to assess how well the algorithms were able to recover the known connectivity. We also assessed essential processing steps to pinpoint critical flaws that many current pipelines have in common. On average, submissions recovered only a third of the volumetric extent of existing bundles. Also, most algorithms routinely extracted many false positive bundles, even though they were not part of the ground truth. Some of these false-positive bundles resemble previously reported pathways identified by *in vivo* tractography, such as the frontal aslant tract (41) or the vertical occipital fasciculus (42). The average ratio of false-positive to true-positive bundles was approximately four to one. This ratio could not be improved by employing higher quality data or even using the gold standard field of local orientations, highlighting that current tractography approaches are fundamentally ill-posed.

Results

Datasets and submissions

Prior investigations of tractography methodology have chosen artificial fiber geometries to construct synthetic ground truth models (28,43). Here, we defined our ground truth based on the fiber bundle geometry of a high-quality Human Connectome Project (HCP) dataset that was constructed from multiple whole-brain global tractography maps (44) (Supplementary Figure 1). Following the concepts introduced in (45), an expert radiologist (B.S.) extracted 25 major tracts (i.e., bundles of streamlines) from the tractogram. These association, projection and commissural fibers covered more than 70% of the white matter across the whole brain. The dataset features a brain-like macro-

structure of long-range connections, mimicking *in vivo* DWI clinical-like acquisitions based on a simulated diffusion signal. An additional anatomical image with T1-like contrast was simulated as a reference. The final datasets and all files needed to perform the simulation are available online (Supplementary Data 1).

20 research groups across 12 countries (Supplementary Figure 2a) participated in the competition and submitted a total of 96 processing results comprising a large variety of tractography pipelines with different pre-processing, local reconstruction, tractography and post-processing algorithms (Supplementary Figure 2b, Supplementary Notes 2).

Performance metrics and evaluation

The Tractometer connectivity metrics (43) were used for evaluating submissions. Based on the known ground truth bundles, we calculated true positives, corresponding to the valid connection ratio (VC), that is, the proportion of streamlines connecting valid end points and the associated number of valid bundles (VB), where a bundle is a group of streamlines. We also computed false positives, corresponding to the invalid connection ratio (IC) and the associated number of invalid bundles (IB), as well as reconstructed volumes, based on the bundle volumetric overlap (OL) and volumetric overreach (OR) in percent (Supplementary Figure 3).

Tractography detects major bundles, but not to the full extent

While the tractography algorithms detected most existing bundles, the extent of their volumetric reconstruction varied. Figure 2a groups identified valid bundles into three clusters of “very hard”, “hard” and “medium” bundles. Figure 2b shows corresponding examples that were reconstructed by different tractography techniques. Most submissions had difficulties identifying the smallest tracts, that is, the anterior (CA) and posterior commissures (CP) that have a cross-sectional diameter of no more than two millimeters, or one or two voxels (“very hard”). A family of “hard” bundles was recovered by almost all algorithms, with a low overlap score of approximately 30%. Bundles of "medium" difficulty were the corpus callosum (CC), inferior longitudinal fasciculus (ILF), superior longitudinal fasciculus (SLF) and uncinate fasciculus (UF). These tracts span different shapes, lengths and sizes, but were recovered most frequently, with above 45% overlap and around 40% overreach.

As shown in Figure 3, the submissions identified an average of 21 out of 25 valid bundles (median 23). No team detected all valid bundles, but three teams (10 submissions) recovered 24 valid bundles, and 15 out of 20 teams (69 submissions) detected 23 or more valid bundles (Figure 4a, red entries in connectivity matrix). However, tractography pipelines clearly need to improve their recovery of the full spatial extent of bundles (Figure 3c). The mean value of bundle volume overlap (OL) across all submissions was 31% ± 17%, with an average overreach (OR) of 23% ± 21%. At the level of individual streamlines, an average of 54% ± 23% connections were valid (Figure 3a).

Tractography identifies more invalid than valid bundles

Across submissions, 36% ± 17% of the reconstructed individual streamlines connected regions that were not actually connected. The fraction of streamlines not connecting any endpoints was 10% ± 15%. Even though not part of the ground truth, these streamlines often occur in dense, structured and coherent bundles. On average, submissions identified more than four times as many invalid bundles as they identified valid bundles (Figure 3b). Submissions with at least 23 valid bundles showed no fewer than 37 invalid bundles (mean 88 ± 39, n = 69). Submissions with 23 or more valid bundles *and* a volumetric bundle overlap of > 50% identified between 99 and 204 invalid bundles (corresponding to more than four invalid bundles per valid bundle). The submissions produced an average of 88 ± 58 invalid bundles, demonstrating the inability of current state-of-the-art

tractography algorithms to control for false positives (Figure 4a, blue entries in connectivity matrix). 41 of these invalid bundles occurred in the majority of submissions (Supplementary Figure 4).

The bundles illustrated in Figures 4b and 4c were systematically found by more than 80% of submissions without being part of the ground truth. Interestingly, several of these invalid streamline clusters exhibited similarities to bundles known or previously debated in tractography literature. They anatomically resemble the following pathways: 1) the frontal aslant tract (FAT) (41), 2) the arcuate fasciculus (AF) (46,47), 3) bundles passing through the temporal stem (48), such as the inferior-frontal occipital fasciculus (IFOF) (49,50), middle longitudinal fasciculus (MdLF) (51,52) and the extreme capsule fasciculus (EmC) (53), 4) the superior fronto-occipital fasciculus (SFOF) (49,54) and 5) the vertical occipital fasciculus (VOF) (42). The existence of the FAT, SFOF and VOF is controversial (41,42,49,54).

Higher image quality does not solve the problem

To confirm that our findings revealed fundamental properties of tractography itself and are not related to effects of our specific phantom simulation process, we ran additional tractography experiments directly on the *ground truth field of fiber orientations* (Supplementary Figure 5 and Supplementary Notes 3), that is, without using the diffusion-weighted data at all. This setup was, thus, independent of image quality, artifacts and many other influences from specific pipeline configurations and the phantom generation process. Based on the ground truth orientations, the tractography pipelines achieved much higher overlap (71% ± 2%) and lower overreach (20% ± 0.2%) scores, while achieving valid connection ratios between 75% and 81%. However, they still generated more than three times as many invalid bundles than valid bundles (at least 82 invalid bundles).

Tractography is fundamentally ill-posed

Our results show that the geometry of many junctions in the brain is too complex to be resolved by current tractography algorithms, even when given a perfect ground truth field of orientations. In the temporal lobe, for example, multiple bundles overlap and clearly outnumber the count of fiber orientations in most of the voxels. As illustrated in Figure 5, *single* fiber directions in the diffusion signal regularly represent multiple bundles (see also Supplementary Video 1). Such funnels embody hard bottlenecks for tractography, leading to massive combinatorial possibilities of plausible configurations for connecting the associated bundle endpoints as sketched in Figure 5c. Consequently, for the real dataset as well as the synthetic phantom, dozens of structured and coherent bundles pass through this bottleneck, exhibiting a wide range of anatomically reasonable geometries as illustrated in Supplementary Video 2. A tractogram based on real HCP data exhibits whole families of theoretically plausible bundles going through the temporal lobe bottleneck even though, locally, the diffusion signal often shows only one fiber direction (cf. Figure 5d).

Statistical analysis of processing steps

Effects of the methodological setup of the different submissions on the results were investigated in a multivariable linear mixed model and revealed the influence of the individual processing steps on the tractography outcome (Supplementary Table 1, Supplementary Notes 4). The choice of tractography algorithm, as well as the post-tracking filtering strategy and the underlying diffusion modeling had a strong effect on overall scores, revealing a clear tradeoff between sensitivity and specificity.

Discussion

We assessed current state-of-the-art fiber tractography approaches using a ground truth dataset of white matter tracts and connectivity that is representative of the challenges that occur in human

brain imaging *in vivo*. Advanced tractography strategies in combination with current diffusion modeling techniques successfully recovered most valid bundles, covering up to 76% of their volumetric extent. This sensitivity comes at a high cost. Tractography systematically also produced thick and dense bundles of plausible looking streamlines in locations where such streamlines did not actually exist. When focusing on the 64 bundles that were systematically recovered by the majority of submissions, 64% of them were in fact absent from the ground truth. Currently even the best tractography pipelines, or *tracking of the ground truth fiber orientations,* produce more false-positive than true-positive bundles. These findings expose the degree of ambiguity associated with whole-brain tractography and show how the computational problem of tractography goes far beyond the local reconstruction of fiber directions (1,8) and issues of data quality. Our findings, therefore, present a core challenge for the field of tractography and connectivity mapping.

High invalid-connection ratios were previously reported under simplified conditions (28,43) (www.tractometer.org), and some of the underlying ambiguities in tractography have been discussed using schematic representations and theoretical arguments (1,8,9,40). Regions of white matter bottlenecks have been discussed in the past (38) and have been highlighted as critical with respect to tractographic findings (39). The present results reveal and quantify the consequences of such limitations under conditions found in human brain studies *in vivo*, addressing important questions that previously remained speculative. The findings were derived from a brain-like geometry that encompasses some of the major known long-range connections and covers more than 70% of the white matter. Future versions of the phantom are planned to include additional bundles such as the middle and inferior temporal projections of the AF, the MdLF and the IFOF as well as smaller U-fibers, medial forebrain fibers, deep nuclei and connections between them. In addition, more advanced diffusion modeling methods will allow generating even more realistic DWI signals, potentially simulated at increased spatial and q-space resolutions. These developments, however, will not resolve the fundamental ambiguities which tractography faces and thus will only have a limited effect on the main findings of our study. We showed that false-positive bundles occur at similar rates even when using the maximal angular precision of the signal, that is, using ground truth orientations. Increasing the anatomic complexity of the phantom by adding more bundles will most likely lead to even increased false-positive rates. The construction process of the current phantom resembles a potential limitation, since it involves tractography itself and thus raises self-validation issues. This should be considered in direct method comparisons as there may exist a possible bias towards algorithms that equal the algorithm used for phantom generation. This caveat, however, has only a very limited effect on our general findings. It can be expected that the identified limitations of tractography will become even more pronounced in phantoms of higher anatomic complexity that might be achievable by involving independent methods such as polarized light imaging (PLI) (55). In summary, our observations confirm the fundamental ill-posedness of the computational problem that current tractography approaches strive to solve.

The large number of invalid bundles that were systematically identified by most state-of-the-art algorithms can only be resolved by substantial methodological innovation. Several directions of current research might improve the specificity of tractography. Streamline filtering techniques can optimize the signal prediction error in order to reduce tractography biases (14,16,56). Such techniques can increase the reproducibility and quality of tractograms. Teams 13, 16 and 17 applied such filtering techniques, showing a positive effect (albeit not significant) on the valid connection ratio, the invalid bundle count and the overreach score, at the expense of decreasing valid bundle counts and decreasing overlap. The mentioned techniques are part of the more general trend to integrate non-local as well as advanced diffusion microstructure modeling information that goes beyond the raw directional vectors (57–62). Recent advances in machine-learning-driven

tractography also show high potential in improving the specificity of tractograms (63). Future versions of our phantom will be generated with multiple b-values, better signal-to-noise ratio (SNR) and fewer artifacts to further encourage research in these directions.

In addition, tractography should employ reliable anatomical priors, such as from animal experiments, for optimal guidance. While manual or automated clean-up of streamlines may help (see our results), the real challenge is our limited knowledge of the anatomy to be reconstructed. Currently, post-mortem dissection with Klingler's method reveals the macroscopic organization of the human brain white matter (11,64–66). In the future, the community will have to gain further insights into the underlying principles of white matter organization and increasingly learn how to leverage such information for tractography (1,67,68).

Another limitation of tractography as emphasized by our results is highly relevant for the field of connectomics: The traditional metrics that require streamlines to exactly end in head or tail regions of a bundle are far too restrictive for bundle dissection and connectivity assessment. None of the submissions generated exact overlaps of streamlines with ground truth bundles and dilated endpoint masks. This finding raises an important warning for connectomics and structural connectivity studies, where a voxel-wise definition of parcellations on the T1 image is the state-of-the-art method for selecting relevant streamlines. This finding is in line with previous reports which found termination of tracts in the grey matter to be inaccurate (6). Future versions of our phantom will include ground-truth parcellations of the white matter/gray matter cortical band to encourage further developments for tackling these problems and extend the evaluation method to apply to graph theory metrics.

DWI is the only tool to map long-range structural brain connectivity *in vivo* and is essential for comparing brains, detecting differences and simulating functions (69). However, our findings should foster the development of novel tractography methods that are carefully evaluated using the presented approach. The most important goal for the next generation of tractography algorithms will no longer be to find existing valid connections, but to reconstruct the full spatial extent of tracts while controlling for the many false-positive connections polluting the tractograms. A tighter integration of advanced diffusion microstructure modeling and multi-modality imaging in tractography should help resolve ambiguities in the signal and overcome current limitations of tractography (62,70). Fundamentally, tractography will require severe methodological innovation to become tractable (1,8).

Online Methods

Generation of ground truth fiber bundles

The set of ground truth long-range fiber bundles was designed to cover the whole human brain and feature many of the relevant spatial configurations, such as crossing, kissing, twisting and fanning fibers, thus representing the morphology of the major known *in vivo* fiber bundles. The process to obtain these bundles consisted of three steps. First, a whole-brain global tractography was performed on a high quality *in vivo* diffusion-weighted image. Then, 25 major long-range bundles were manually extracted from the resulting tractogram. In the third step, these bundles were refined to obtain smooth and well defined bundles.

We chose one of the diffusion-weighted datasets included in the Q3 data release of the Human Connectome Project (71), subject 100307, to perform whole-brain global fiber tractography (57,72). Amongst other customizations, the HCP scanners are equipped with a set of high-end gradient coils, enabling diffusion encoding gradient strengths of 100 mT m^{-1}. By comparison, most standard magnetic resonance scanners feature gradient strengths of about 30 to 40 mT m^{-1}. This hardware setup allows the acquisition of datasets featuring exceptionally high resolutions (1.25 mm isotropic, 270 gradient directions) while maintaining an excellent SNR. All datasets were corrected for head motion, eddy currents and susceptibility distortions and are, in general, of very high quality (73–77). Detailed information regarding the employed imaging protocols as well as the datasets themselves may be found on http://humanconnectome.org.

Global fiber tractography was performed using *MITK Diffusion* (78) with the following parameters: 900,000,000 iterations, a particle length of 1 mm, a particle width of 0.1 mm and a particle weight of 0.002. Furthermore, we repeated the tractography six times and combined the resulting whole-brain tractograms into one large dataset consisting of over five million streamlines. The selected parameters provided for a very high sensitivity of the tractography method. The specificity of the resulting tractogram was of lesser concern since the tracts of interest were extracted manually in the second step.

Bundle segmentation was performed by an expert radiologist using manually placed inclusion and exclusion regions of interest (ROI). We followed the concepts introduced in (45) for the ROI placement and fiber extraction. 25 bundles were extracted, covering association, projection and commissural fibers across the whole brain (Figure 1): corpus callosum (CC), left and right cingulum (Cg), Fornix (Fx), anterior commissure (CA), left and right optic radiation (OR), posterior commissure (CP), left and right inferior cerebellar peduncle (ICP), middle cerebellar peduncle (MCP), left and right superior cerebellar peduncle (SCP), left and right parieto-occipital pontine tract (POPT), left and right cortico-spinal tract (CST), left and right frontopontine tracts (FPT), left and right inferior longitudinal fasciculus (ILF), left and right uncinate fasciculus (UF) and left and right superior longitudinal fasciculus (SLF). As mentioned in the Discussion section, the inferior fronto-occipital fasciculus (IFOF), the middle longitudinal fasciculus (MdLF) as well as the middle and inferior temporal projections of the arcuate fasciculus (AF) were not included.

After manual extraction, the individual long-range bundles were further refined to serve as ground truth for the image simulation as also shown in Figure 1. The original extracted tracts featured a large number of prematurely ending fibers and the individual streamlines were not smooth. To obtain smooth tracts without prematurely ending fibers, we simulated a diffusion-weighted image from each original tract individually using *Fiberfox* (www.mitk.org (35)). Since no complex fiber configurations, such as crossings, were present in the individual tract images and no artifacts were simulated, it was possible to obtain very smooth and complete tracts from these images with a

simple tensor-based streamline tractography. Supplementary Figure 6 illustrates the result of this refining procedure on the left CST.

Simulation of phantom images with brain-like geometry

The phantom diffusion-weighted images (Supplementary Video 3) were simulated using Fiberfox (www.mitk.org (35)), which is available as open-source software. We employed a four-compartment model of brain tissue (intra and inter-axonal), grey matter (GM) and cerebrospinal fluid (CSF) (35). The parameters for simulation of the four-compartment diffusion-weighted signal were chosen to obtain representative diffusion properties and image contrasts (compare (79) for details on the models). The *intra-axonal compartment* was simulated using the stick model with a T2 relaxation time of 110 ms and a diffusivity of 1.2×10^{-9} m^2/s. The *inter-axonal compartment* was simulated using the zeppelin model with a T2 relaxation time of 110 ms, an axial diffusivity of 1.2×10^{-9} m^2/s and a radial diffusivity of 0.3×10^{-9} m^2/s. The *grey matter compartment* was simulated using the ball model with a T2 relaxation time of 80 ms and a diffusivity of 1.0×10^{-9} m^2/s. The *CSF compartment* was also simulated using the ball model with a T2 relaxation time of 2500 ms and a diffusivity of 2.0×10^{-9} m^2/s.

Using Fiberfox, one set of diffusion-weighted images and one T1-weighted image were simulated. The final datasets as well as all files needed to perform the simulation are available online (Supplementary Data 1).

The acquisition parameters that we report below were chosen to simulate images that are representative for a practical (e.g., clinical) setting, specifically a 5-to-10-minute single shot EPI scan with 2 mm isotropic voxels, 32 gradient directions and a b-value of 1000 s/mm^2. The chosen acquisition setup represents a typical scenario for an applied tractography study and embodies a common denominator supported by the large majority of methods. Since acquisitions with higher b-values, more gradient directions and fewer artifacts are beneficial for tractography, we additionally report a least upper bound tractography performance under perfect image quality conditions using a dataset that directly contains ground truth fiber orientation distribution functions and no artifacts (Supplementary Figure 5 and Supplementary Notes 3).

The parameters are a matrix size of 90×108×90, echo time (TE) 108 ms, dwell time 1 ms; T2' relaxation time 50 ms. The simulation corresponded to a single-coil acquisition with constant coil sensitivity, no partial Fourier and no parallel imaging. Phase encoding was posterior-anterior. Two unweighted images with posterior-anterior/anterior-posterior phase encoding were also generated.

Since *Fiberfox* simulates the actual k-space acquisition, it was possible to introduce a number of common artifacts into the final image. Complex *Gaussian noise* was simulated yielding a final SNR relative to the mean white matter baseline signal of about 20. 10 *spikes* were distributed randomly throughout the image volumes (Supplementary Figure 7a). *N/2 ghosts* were simulated (Supplementary Figure 7b). Distortions caused by *B_0 field inhomogeneities* are introduced using an existing field map measured in a real acquisition and registered to the employed reference HCP dataset (Supplementary Figure 7c). *Head motion* was introduced as random rotation (+-4° around z-axis) and translation (+-2mm along x-axis) in three randomly chosen volumes. Volume 6 was rotated by 3.36° and translated by -1.74 mm, volume 12 was rotated by 1.23° and translated by -0.72 mm, and volume 24 was rotated by -3.12° and translated by -1.55 mm.

The image with the T1-like contrast was generated at an isotropic resolution of 1 mm, an SNR of about 40 and no further artifacts as an anatomical reference.

Performance metrics and evaluation

The Tractometer definition of a valid connection is extremely restrictive for current tractography algorithms, as it requires streamlines 1) not to exit the area of the ground truth bundle at any point and 2) to terminate exactly within the endpoint region that is defined by the dilated ground truth fiber endpoints (Supplementary Figure 8 and Supplementary Figure 9) (43). Hence, we propose an alternative definition that relaxes these strict criteria based on robust shape distance measures (80) and clustering between streamlines (81), as detailed in Supplementary Notes 1. The bundle-specific thresholds were manually configured to account for bundle shape and proximity to other bundles. The following distances were used, with identical distances on both sides for lateralized bundles: 2 mm for CA and CP; 3 mm for CST and SCP; 5 mm for Cingulum; 6 mm for Fornix, ICP, OR and UF; 7 mm for FPT, ILF and POPT; 10 mm for CC, MCP and SLF. The full script used to run this bundle recognition implementation was based on the *DIPY library* (82) (www.dipy.org) and is available online (Supplementary Software 1).

Once valid connections are identified, the remaining streamlines can be classified into *invalid connections* and *non-connecting streamlines*. The details of this procedure are described in Supplementary Notes 1. We clustered the remaining invalid streamlines using a QuickBundles-based clustering algorithm (81). The best matching endpoint regions for each resulting cluster were identified by majority voting of the contained streamlines. If multiple clusters were assigned to the same pair of regions, they were merged. Streamlines that were not assigned to any cluster or that fell below a length threshold were labelled as non-connecting.

On the basis of this classification of streamlines, the following metrics were calculated:

- Valid connection ratio (VC): number of valid connections / total number of streamlines (percentage between 0 and 100).
- Valid bundles (VB): For each bundle that has a valid streamline associated with it, this counter is incremented by one (integer number between 0 and 25).
- Invalid bundles (IB): With 25 bundles in the ground truth, each having two endpoint regions, there are 1,275 possible combinations of endpoint regions. Taking the 25 valid bundles out of the equation, 1,250 potential invalid bundles remain (integer number between 0 and 1,250).
- Overlap: Proportion of the voxels within the volume of a ground truth bundle that is traversed by at least one valid streamline associated with the bundle. This value shows how well the tractography result recovers the original volume of the bundle (percentage between 0 and 100).
- Overreach: Fraction of voxels outside the volume of a ground truth bundle that is traversed by at least one valid streamline associated with the bundle over the total number of voxels within the ground truth bundle. This value shows how much the valid connections extend beyond the ground truth bundle volume (percentage between 0 and 100). This value is always zero for the traditional definition of a valid connection but can be non-zero for the relaxed evaluation.

Statistical multi-variable analysis

Effects of the experimental settings were investigated in a multivariable linear mixed model. The experimental variables describing the methods used for the different submissions were included as fixed effects (Figure 2b). The valid connection ratio, the valid bundle count, the invalid bundle count,

the bundle overlap percentage and the bundle overreach percentage were modeled as dependent variables, each of which is used for the calculation of a separate model. The submitting group was modeled as a random effect. The software *SAS 9.2*, Proc Mixed, SAS Institute Inc., Cary, NC, USA was used for the analysis.

References

1. Jbabdi S, Sotiropoulos SN, Haber SN, Van Essen DC, Behrens TE. Measuring macroscopic brain connections in vivo. Nat Neurosci. 2015 Nov;18(11):1546–55.

2. Bullmore E, Sporns O. Complex brain networks: graph theoretical analysis of structural and functional systems. Nat Rev Neurosci. 2009;10(3):186–198.

3. Sporns O. Contributions and challenges for network models in cognitive neuroscience. Nat Neurosci. 2014 May;17(5):652–60.

4. Deco G, Kringelbach ML. Great Expectations: Using Whole-Brain Computational Connectomics for Understanding Neuropsychiatric Disorders. Neuron. 2014 Mar 12;84(5):892–905.

5. Craddock RC, Jbabdi S, Yan C-G, Vogelstein JT, Castellanos FX, Di Martino A, et al. Imaging human connectomes at the macroscale. Nat Methods. 2013 Jun;10(6):524–39.

6. Thomas C, Ye FQ, Irfanoglu MO, Modi P, Saleem KS, Leopold DA, et al. Anatomical accuracy of brain connections derived from diffusion MRI tractography is inherently limited. Proc Natl Acad Sci U S A. 2014 Nov 18;111(46):16574–9.

7. Reveley C, Seth AK, Pierpaoli C, Silva AC, Yu D, Saunders RC, et al. Superficial white matter fiber systems impede detection of long-range cortical connections in diffusion MR tractography. Proc Natl Acad Sci U S A. 2015 May 26;112(21):E2820-2828.

8. Jbabdi S, Johansen-Berg H. Tractography: where do we go from here? Brain Connect. 2011;1(3):169–83.

9. Jones DK. Challenges and limitations of quantifying brain connectivity in vivo with diffusion MRI. Imaging Med. 2010 May 24;2(3):341–55.

10. Pujol C, Neher P, Maier-Hein KH, Golby A, Kikinis R. The DTI Challenge: Towards Standardized Evaluation of Diffusion Tensor Imaging Tractography for Neurosurgery. J Neuroimaging. 2015;JON-15-4431 (in press).

11. Martino J, De Witt Hamer PC, Vergani F, Brogna C, de Lucas EM, Vázquez-Barquero A, et al. Cortex-sparing fiber dissection: an improved method for the study of white matter anatomy in the human brain. J Anat. 2011 Oct;219(4):531–41.

12. Wang X, Pathak S, Stefaneanu L, Yeh F-C, Li S, Fernandez-Miranda JC. Subcomponents and connectivity of the superior longitudinal fasciculus in the human brain. Brain Struct Funct. 2015 Mar 18;

13. Wakana S, Caprihan A, Panzenboeck MM, Fallon JH, Perry M, Gollub RL, et al. Reproducibility of quantitative tractography methods applied to cerebral white matter. Neuroimage. 2007;36:630–644.

14. Pestilli F, Yeatman JD, Rokem A, Kay KN, Wandell BA. Evaluation and statistical inference for human connectomes. Nat Methods. 2014 Oct;11(10):1058–63.

15. Neher PF, Descoteaux M, Houde J-C, Stieltjes B, Maier-Hein KH. Strengths and weaknesses of state of the art fiber tractography pipelines – A comprehensive in-vivo and phantom evaluation study using Tractometer. Med Image Anal. 2015 Dec;26(1):287–305.

16. Daducci A, Dal Palù A, Lemkaddem A, Thiran J-P. COMMIT: Convex optimization modeling for microstructure informed tractography. IEEE Trans Med Imaging. 2015 Jan;34(1):246–57.

17. Dyrby TB, Søgaard LV, Parker GJ, Alexander DC, Lind NM, Baaré WFC, et al. Validation of in vitro probabilistic tractography. Neuroimage. 2007;37:1267–1277.

18. Campbell JS, Siddiqi K, Rymar VV, Sadikot AF, Pike GB. Flow-based fiber tracking with diffusion tensor and q-ball data: validation and comparison to principal diffusion direction techniques. Neuroimage. 2005;27:725–736.

19. Dauguet J, Peled S, Berezovskii V, Delzescaux T, Warfield SK, Born R, et al. Comparison of fiber tracts derived from in-vivo DTI tractography with 3D histological neural tract tracer reconstruction on a macaque brain. Neuroimage. 2007;37:530–538.

20. Jbabdi S, Lehman J, Haber S, Behrens T. Human and monkey ventral prefrontal fibers use the same organizational principles to reach their targets: tracing versus tractography. J Neurosci. 2013;33:3190–3201.

21. Lawes INC, Barrick TR, Murugam V, Spierings N, Evans DR, Song M, et al. Atlas-based segmentation of white matter tracts of the human brain using diffusion tensor tractography and comparison with classical dissection. Neuroimage. 2008;39:62–79.

22. Schmahmann JD, Pandya DN, Wang R, Dai G, D'Arceuil HE, de Crespigny AJ, et al. Association fibre pathways of the brain: parallel observations from diffusion spectrum imaging and autoradiography. Brain. 2007;130:630–653.

23. Seehaus AK, Roebroeck A, Chiry O, Kim D-S, Ronen I, Bratzke H, et al. Histological Validation of DW-MRI Tractography in Human Postmortem Tissue. Cereb Cortex. 2013;23:442–450.

24. Knösche TR, Anwander A, Liptrot M, Dyrby TB. Validation of tractography: Comparison with manganese tracing. Hum Brain Mapp. 2015 Oct;36(10):4116–34.

25. Donahue CJ, Sotiropoulos SN, Jbabdi S, Hernandez-Fernandez M, Behrens TE, Dyrby TB, et al. Using Diffusion Tractography to Predict Cortical Connection Strength and Distance: A Quantitative Comparison with Tracers in the Monkey. J Neurosci Off J Soc Neurosci. 2016 Jun 22;36(25):6758–70.

26. Bach M, Maier-Hein (ne Fritzsche) KH, Stieltjes B, Laun FB. Investigation of resolution effects using a specialized diffusion tensor phantom. Magn Reson Med. 2013 May 8;

27. Fieremans E, De Deene Y, Delputte S, Ö MS zdemir, Achten E, Lemahieu I. The design of anisotropic diffusion phantoms for the validation of diffusion weighted magnetic resonance imaging. Phys Med Biol. 2008;53:5405–5421.

28. Fillard P, Descoteaux M, Goh A, Gouttard S, Jeurissen B, Malcolm J, et al. Quantitative evaluation of 10 tractography algorithms on a realistic diffusion MR phantom. NeuroImage. 2011 May 1;56(1):220–34.

29. Maier-Hein (ne Fritzsche) KH, Laun FB, Meinzer H-P, Stieltjes B. Opportunities and pitfalls in the quantification of fiber integrity: What can we gain from Q-ball imaging? Neuroimage. 2010;51(1):242–251.

30. Moussavi-Biugui A, Stieltjes B, Fritzsche K, Semmler W, Laun FB. Novel spherical phantoms for Q-ball imaging under in vivo conditions. Magn Reson Med. 2011;65:190–194.

31. Poupon C, Rieul B, Kezele I, Perrin M, Poupon F, Mangin ois, J-F. New diffusion phantoms dedicated to the study and validation of high-angular-resolution diffusion imaging (HARDI) models. Magn Reson Med. 2008;60:1276–1283.

32. Pullens P, Roebroeck A, Goebel R. Ground truth hardware phantoms for validation of diffusion-weighted MRI applications. J Magn Reson Imaging. 2010;32:482–488.

33. Close TG, Tournier J-D, Calamante F, Johnston LA, Mareels I, Connelly A. A software tool to generate simulated white matter structures for the assessment of fibre-tracking algorithms. Neuroimage. 2009;47:1288–1300.

34. Leemans A, Sijbers J, Verhoye M, Van der Linden A, Van Dyck D. Mathematical Framework for Simulating Diffusion Tensor MR Neural Fiber Bundles. Magn Reson Med. 2005;53:944–953.

35. Neher PF, Laun FB, Stieltjes B, Maier-Hein KH. Fiberfox: facilitating the creation of realistic white matter software phantoms. Magn Reson Med. 2014 Nov;72(5):1460–1470.

36. Daducci A, Canales-Rodríguez EJ, Descoteaux M, Garyfallidis E, Gur Y, Lin Y-C, et al. Quantitative comparison of reconstruction methods for intra-voxel fiber recovery from diffusion MRI. IEEE Trans Med Imaging. 2014 Feb;33(2):384–99.

37. Perrone D, Jeurissen B, Aelterman J, Roine T, Sijbers J, Pizurica A, et al. D-BRAIN: Anatomically Accurate Simulated Diffusion MRI Brain Data. PloS One. 2016;11(3):e0149778.

38. Mangin J-F, Regis J, Frouin V. Shape Bottlenecks and Conservative Flow Systems. In: Proceedings of the 1996 Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA '96) [Internet]. Washington, DC, USA: IEEE Computer Society; 1996 [cited 2016 Oct 26]. p. 319–. (MMBIA '96). Available from: http://dl.acm.org/citation.cfm?id=882463.882767

39. Guevara P, Poupon C, Rivière D, Cointepas Y, Descoteaux M, Thirion B, et al. Robust clustering of massive tractography datasets. NeuroImage. 2011 Feb 1;54(3):1975–93.

40. Basser PJ. Fiber-tractography via diffusion tensor MRI. In: Proc International Society for Magnetic Resonance in Medicine. 1998.

41. Catani M, Mesulam MM, Jakobsen E, Malik F, Martersteck A, Wieneke C, et al. A novel frontal pathway underlies verbal fluency in primary progressive aphasia. Brain J Neurol. 2013 Aug;136(Pt 8):2619–28.

42. Yeatman JD, Weiner KS, Pestilli F, Rokem A, Mezer A, Wandell BA. The vertical occipital fasciculus: a century of controversy resolved by in vivo measurements. Proc Natl Acad Sci U S A. 2014 Dec 2;111(48):E5214-5223.

43. Cote MA, Girard G, Bore A, Garyfallidis E, Houde JC, Descoteaux M. Tractometer: Towards validation of tractography pipelines. Med Image Anal. 2013;17:844–857.

44. Van Essen DC, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K, et al. The WU-Minn Human Connectome Project: an overview. NeuroImage. 2013 Oct 15;80:62–79.

45. Stieltjes B, Brunner RM, Maier-Hein (ne Fritzsche) KH, Laun FB. Diffusion Tensor Imaging: Introduction and Atlas. Springer Berlin Heidelberg; 2013.

46. Forkel SJ, Thiebaut de Schotten M, Dell'Acqua F, Kalra L, Murphy DGM, Williams SCR, et al. Anatomical predictors of aphasia recovery: a tractography study of bilateral perisylvian language networks. Brain J Neurol. 2014 Jul;137(Pt 7):2027–39.

47. de Schotten MT, Dell'Acqua F, Forkel SJ, Simmons A, Vergani F, Murphy DGM, et al. A lateralized brain network for visuospatial attention. Nat Neurosci. 2011 Oct;14(10):1245–6.

48. Hau J, Sarubbo S, Perchey G, Crivello F, Zago L, Mellet E, et al. Cortical Terminations of the Inferior Fronto-Occipital and Uncinate Fasciculi: Anatomical Stem-Based Virtual Dissection. Front Neuroanat [Internet]. 2016 May 24 [cited 2016 Oct 28];10. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4877506/

49. Forkel SJ, Thiebaut de Schotten M, Kawadler JM, Dell'Acqua F, Danek A, Catani M. The anatomy of fronto-occipital connections from early blunt dissections to contemporary tractography. Cortex J Devoted Study Nerv Syst Behav. 2014 Jul;56:73–84.

50. Caverzasi E, Papinutto N, Amirbekian B, Berger MS, Henry RG. Q-Ball of Inferior Fronto-Occipital Fasciculus and Beyond. PLOS ONE. 2014 Jun 19;9(6):e100274.

51. Makris N, Preti MG, Asami T, Pelavin P, Campbell B, Papadimitriou GM, et al. Human middle longitudinal fascicle: variations in patterns of anatomical connections. Brain Struct Funct. 2013 Jul;218(4):951–68.

52. Menjot de Champfleur N, Lima Maldonado I, Moritz-Gasser S, Machi P, Le Bars E, Bonafé A, et al. Middle longitudinal fasciculus delineation within language pathways: a diffusion tensor imaging study in human. Eur J Radiol. 2013 Jan;82(1):151–7.

53. Mars RB, Foxley S, Verhagen L, Jbabdi S, Sallet J, Noonan MP, et al. The extreme capsule fiber complex in humans and macaque monkeys: a comparative diffusion MRI tractography study. Brain Struct Funct. 2015 Dec 1;

54. Meola A, Comert A, Yeh F-C, Stefaneanu L, Fernandez-Miranda JC. The controversial existence of the human superior fronto-occipital fasciculus: Connectome-based tractographic study with microdissection validation. Hum Brain Mapp. 2015 Dec 1;36(12):4964–71.

55. Larsen L, Griffin LD, Graessel D, Witte OW, Axer H. Polarized light imaging of white matter architecture. Microsc Res Tech. 2007 Oct;70(10):851–63.

56. Smith RE, Tournier J-D, Calamante F, Connelly A. The effects of SIFT on the reproducibility and biological accuracy of the structural connectome. NeuroImage. 2015 Jan 1;104:253–65.

57. Neher PF, Stieltjes B, Reisert M, Reicht I, Meinzer HP, K. Maier-Hein. MITK global tractography. In: SPIE Medical Imaging: Image Processing. 2012.

58. Mangin J-F, Fillard P, Cointepas Y, Bihan DL, Frouin V, Poupon C. Toward global tractography. Neuroimage. 2013;80:290–296.

59. Jbabdi S, Woolrich MW, Andersson JLR, Behrens TEJ. A Bayesian framework for global tractography. NeuroImage. 2007 Aug 1;37(1):116–29.

60. Christiaens D, Reisert M, Dhollander T, Sunaert S, Suetens P, Maes F. Global tractography of multi-shell diffusion-weighted imaging data using a multi-tissue model. NeuroImage. 2015 Dec;123:89–101.

61. Reisert M, Kiselev VG, Dihtal B, Kellner E, Novikov DS. MesoFT: unifying diffusion modelling and fiber tracking. Med Image Comput Comput-Assist Interv MICCAI Int Conf Med Image Comput Comput-Assist Interv. 2014;17(Pt 3):201–8.

62. Girard G, Fick R, Descoteaux M, Deriche R, Wassermann D. AxTract: Microstructure-Driven Tractography Based on the Ensemble Average Propagator. Inf Process Med Imaging Proc Conf. 2015;24:675–86.

63. Neher PF, Maier-Hein* KH. A machine learning based approach to fiber tractography using classifier voting. In: In Proceedings MICCAI. 2015.

64. Zemmoura I, Serres B, Andersson F, Barantin L, Tauber C, Filipiak I, et al. FIBRASCAN: a novel method for 3D white matter tract reconstruction in MR space from cadaveric dissection. NeuroImage. 2014 Dec;103:106–18.

65. De Benedictis A, Petit L, Descoteaux M, Marras CE, Barbareschi M, Corsini F, et al. New insights in the homotopic and heterotopic connectivity of the frontal portion of the human corpus callosum revealed by microdissection and diffusion tractography. Hum Brain Mapp. 2016 Aug 8;

66. Hau J, Sarubbo S, Houde JC, Corsini F, Girard G, Deledalle C, et al. Revisiting the human uncinate fasciculus, its subcomponents and asymmetries with stem-based tractography and microdissection validation. Brain Struct Funct. 2016 Aug 31;1–18.

67. Wedeen VJ, Rosene DL, Wang R, Dai G, Mortazavi F, Hagmann P, et al. The Geometric Structure of the Brain Fiber Pathways. Science. 2012 Mar 30;335(6076):1628–34.

68. Galinsky VL, Frank LR. The Lamellar Structure of the Brain Fiber Pathways. Neural Comput. 2016 Sep 14;28(11):2533–56.

69. Glasser MF, Smith SM, Marcus DS, Andersson JLR, Auerbach EJ, Behrens TEJ, et al. The Human Connectome Project's neuroimaging approach. Nat Neurosci. 2016 Sep;19(9):1175–87.

70. Daducci A, Dal Palú A, Descoteaux M, Thiran J-P. Microstructure Informed Tractography: Pitfalls and Open Challenges. Front Neurosci. 2016;10:247.

71. Van Essen DC, Ugurbil K, Auerbach E, Barch D, Behrens TEJ, Bucholz R, et al. The Human Connectome Project: A data acquisition perspective. Neuroimage. 2012;62:2222–2231.

72. Reisert M, Mader I, Anastasopoulos C, Weigel M, Schnell S, Kiselev V. Global fiber reconstruction becomes practical. Neuroimage. 2011;54:955–962.

73. Andersson J, Xu J, Yacoub E, Auerbach E, Moeller S, Ugurbil K. A comprehensive gaussian process framework for correcting distortions and movements in diffusion images. In: Proceedings of International Society of Magnetic Resonance in Medicine. 2012. p. 2426.

74. Andersson JLR, Skare S, Ashburner J. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. Neuroimage. 2003;20:870–888.

75. Fischl B. FreeSurfer. Neuroimage. 2012;62:774–781.

76. Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. FSL. Neuroimage. 2012;62:782–790.

77. Jenkinson M, Bannister P, Brady M, Smith S. Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. Neuroimage. 2002;17:825–841.

78. Maier-Hein (ne Fritzsche) KH, Neher PF, Reicht I, van Bruggen T, Goch C, Reisert M, et al. MITK diffusion imaging. Methods Inf Med. 2012;51(5):441–8.

79. Panagiotaki E, Schneider T, Siow B, Hall MG, Lythgoe MF, Alexander DC. Compartment models of the diffusion MR signal in brain white matter: a taxonomy and comparison. NeuroImage. 2012 Feb 1;59(3):2241–54.

80. Garyfallidis E, Ocegueda O, Wassermann D, Descoteaux M. Robust and efficient linear registration of white-matter fascicles in the space of streamlines. NeuroImage. 2015 Aug 15;117:124–40.

81. Garyfallidis E, Brett M, Correia MM, Williams GB, Nimmo-Smith I. QuickBundles, a method for tractography simplification. Front Neurosci. 2012;6.

82. Garyfallidis E, Brett M, Amirbekian B, Rokem A, van der Walt S, Descoteaux M, et al. Dipy, a library for the analysis of diffusion MRI data. Front Neuroinformatics. 2014;8:8.

Acknowledgements

This work was supported by the German Research Foundation (DFG), grants MA 6340/10-1, MA 6340/12-1 and the NSERC Discovery Grant program as well as the institutional Université de Sherbrooke Research Chair in Neuroinformatics. C.M.W.T. is supported by a grant (No. 612.001.104) from the Physical Sciences division of the Netherlands Organization for Scientific Research (NWO). The research of H.Y.M., S.D., S.S., A.M.H. and A.L. is supported by VIDI Grant 639.072.411 from NWO. The research of F.G. was funded by the Chinese Scholarship Council (CSC). M.C. is supported by the Alexander Graham Bell Canada Graduate Scholarships-Doctoral Program (CGS-D3) from the Natural Sciences and Engineering Research Council of Canada (NSERC). C.C.H. is supported by DFG SFB grants 936/A1, Z3 and TRR 169/A2.

Author contributions

K.M.H., M.D. and J-C.H. performed the data analysis and wrote the paper with input from all authors. P.N. and B.S. designed the phantom. P.N. and J-C.H. supported the data analysis and J-C.H handled the Tractometer scoring and evaluation metrics proposed. M-A.C and E.G. developed the clustering and bundle recognition algorithm for the relaxed scoring system. K.M.H., P.N., J-C.H., E.C., A.D., T.D., B.S. and M.D. coordinated the tractography challenge at the International Society for Magnetic Resonance in Medicine (ISMRM) 2015 Diffusion Study Group meeting. T.H-L. set up the

multivariable statistical model. P.N. wrote parts of the Online Methods. L.P. and C.C.H. were mentors in the discussion of the paper and neuroanatomical as well as neuroscientific context. Submissions were made by the following teams: J.Z. team 1; M.C. and C.M.W.T. team 2; F-C.Y. team 3; Y-C.L. team 4; Q.J. team 5; D.Q.C. team 6; Y.F., C.G., Y.W., J.M., H.R., Q.L. and C-F.W. team 7; S.D-G., J.O.O.G., M.P., S.S-J. and G.G. team 8; S.S-J., F.R. and J.S. team 9; C.M.W.T., F.G., H.Y.M., S.D., M.F., A.M.H. and A.L. team 10; S.S-J., G.G. and F.R. team 11; J.O.O.G., M.P., G.G. and F.R. team 12; A.B., B.P., C.B., M.D., S.B. and J.D. team 13; A.S., R.V., A.C., A.Q. and J.Y. team 14; A.R.K., W.H. and S.A. team 15; D.R., M.B., A.A., O.E., A.L. and J-P.T. team 16; D.R., M.B., A.A., O.E., A.L. and J-P.T. team 17; H.E.C., B.L.O., B.M. and M.S.N. team 18; F.P., G.P., J.E.V-R., J.G. and P.M.T. team 19; F.D.S.R., P.L.L., L.M.L., R.B. and F.D'A team 20.

Additional information

Supplementary information is available in the online version of the paper. Correspondence and requests for materials should be addressed to K.M.H. or M.D.

Competing financial interests

The authors declare no competing financial interests.

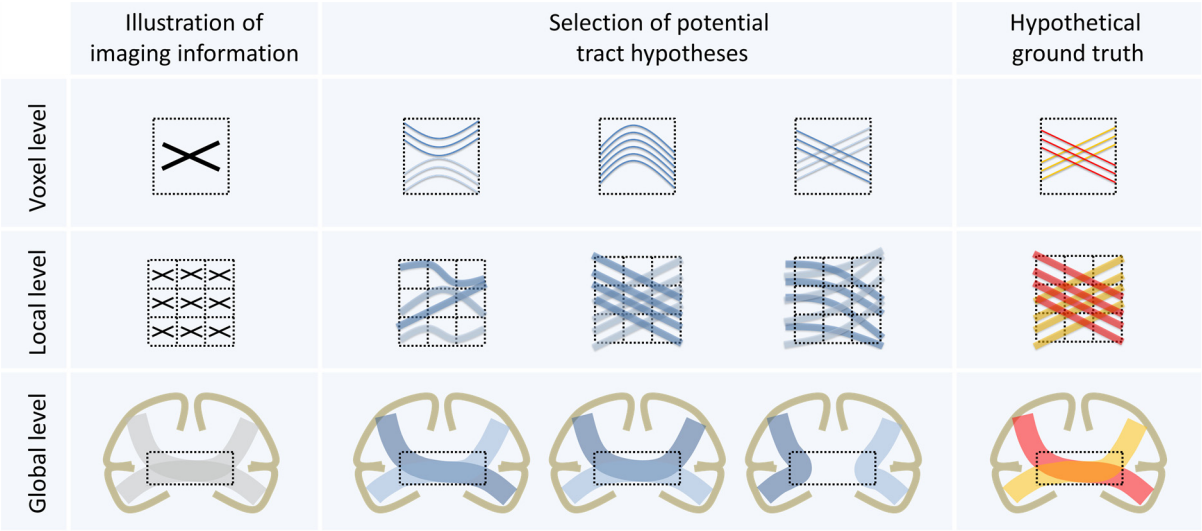

Figure 1. Ambiguous correspondences between diffusion directions and fiber geometry. The three illustrations at voxel, local and global level exemplarily illustrate ambiguities in apparent diffusion imaging information, leading to several potential tract reconstructions. The intra-voxel crossing of fibers in the hypothetical ground truth (first row) leads to ambiguous imaging information at voxel level **(8)**. Similarly, the imaging representation of local fiber crossings (second row) can be explained by several other configurations **(8)**. At a global level (third row), white matter regions that are shared by multiple bundles ("bottlenecks", dotted rectangle) **(38)** can lead to many spurious tractographic reconstructions **(39)**. With only two bundles in the hypothetical ground truth (red and yellow bundle), four potential false-positive bundles emerge.

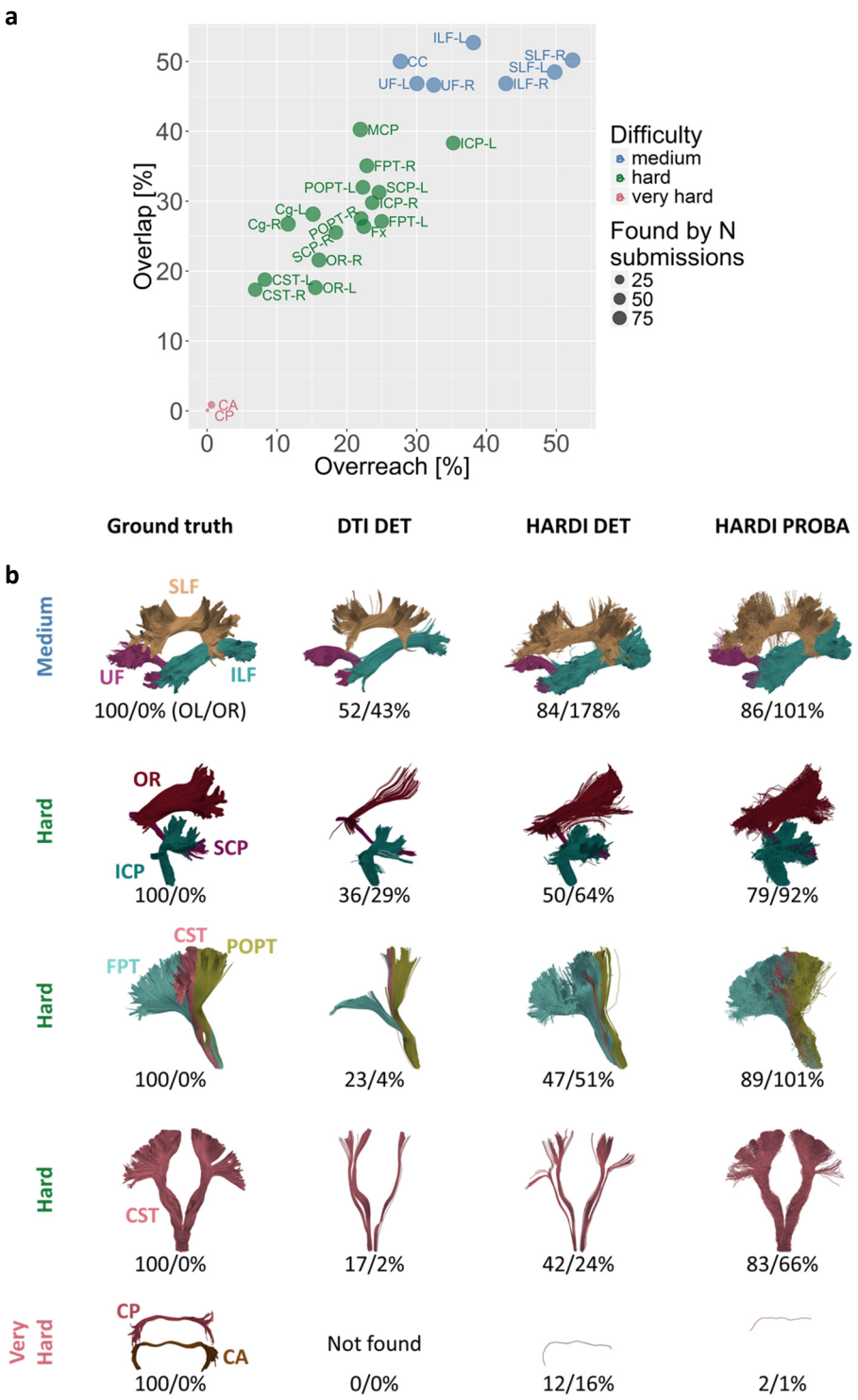

Figure 2. Tractography detects major bundles, but not to the full extent. (a) Overview of scores reached for different bundles in ground truth. Average overlap (OL) and overreach (OR) scores for the submissions (red: "very hard", green: "hard", blue: "medium", for abbreviations see Supplementary Figure 1). (b) Representative bundles for diffusion tensor imaging (DTI) deterministic (DET) tracking come from submission 6 / team 20, high angular resolution diffusion imaging (HARDI) deterministic tracking from submission 0 / team 9 and HARDI probabilistic (PROBA) tracking from submission 2 / team 12 (see Supplementary Notes 5 for a discussion of these submissions). The first column shows ground truth valid bundles for reference. The reported OL and OR scores correspond to the highest OL score reached within the respective class of algorithms.

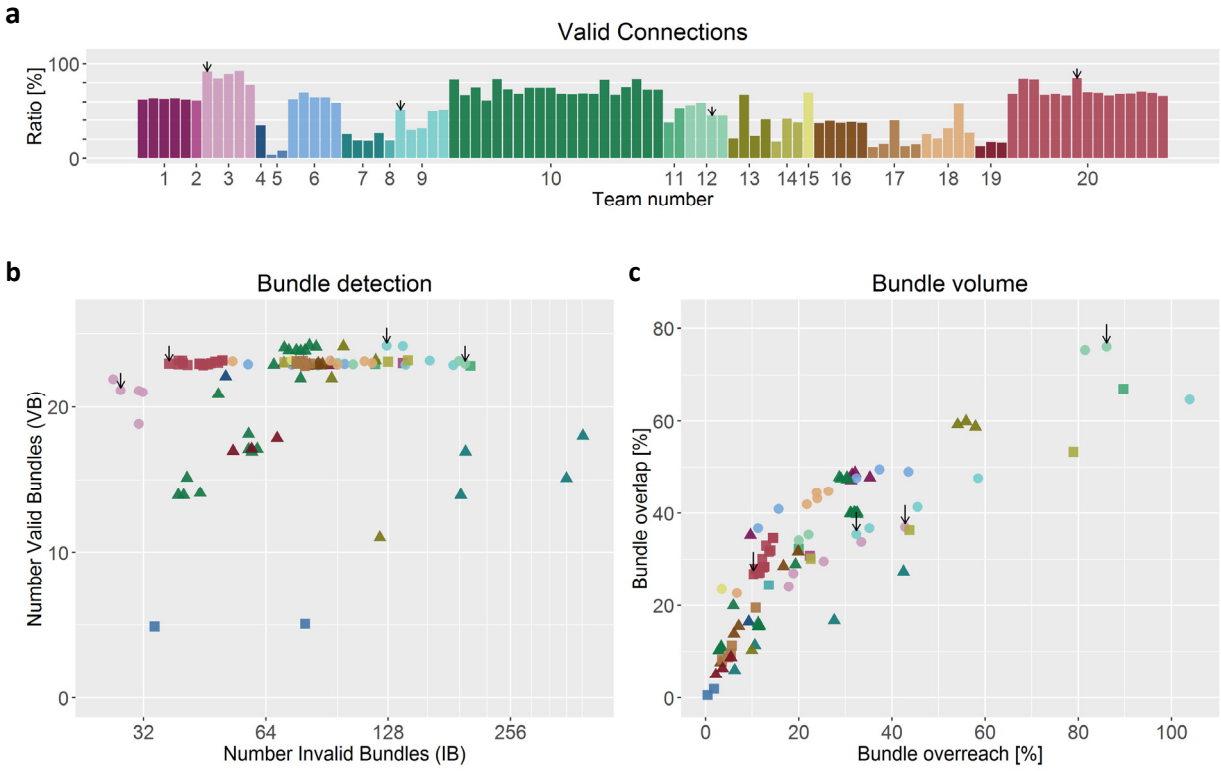

Figure 3. Tractography identifies more invalid than valid bundles. Overview of scores reached by the different teams. (a) Percentage of streamlines connecting valid regions. (b) Number of detected valid and invalid bundles (data points are jittered to improve legibility). (c) Volume overlap (OL) and overreach (OR) scores averaged over bundles. Black arrows mark submissions used in the following figures (see Supplementary Notes 5 for discussion).

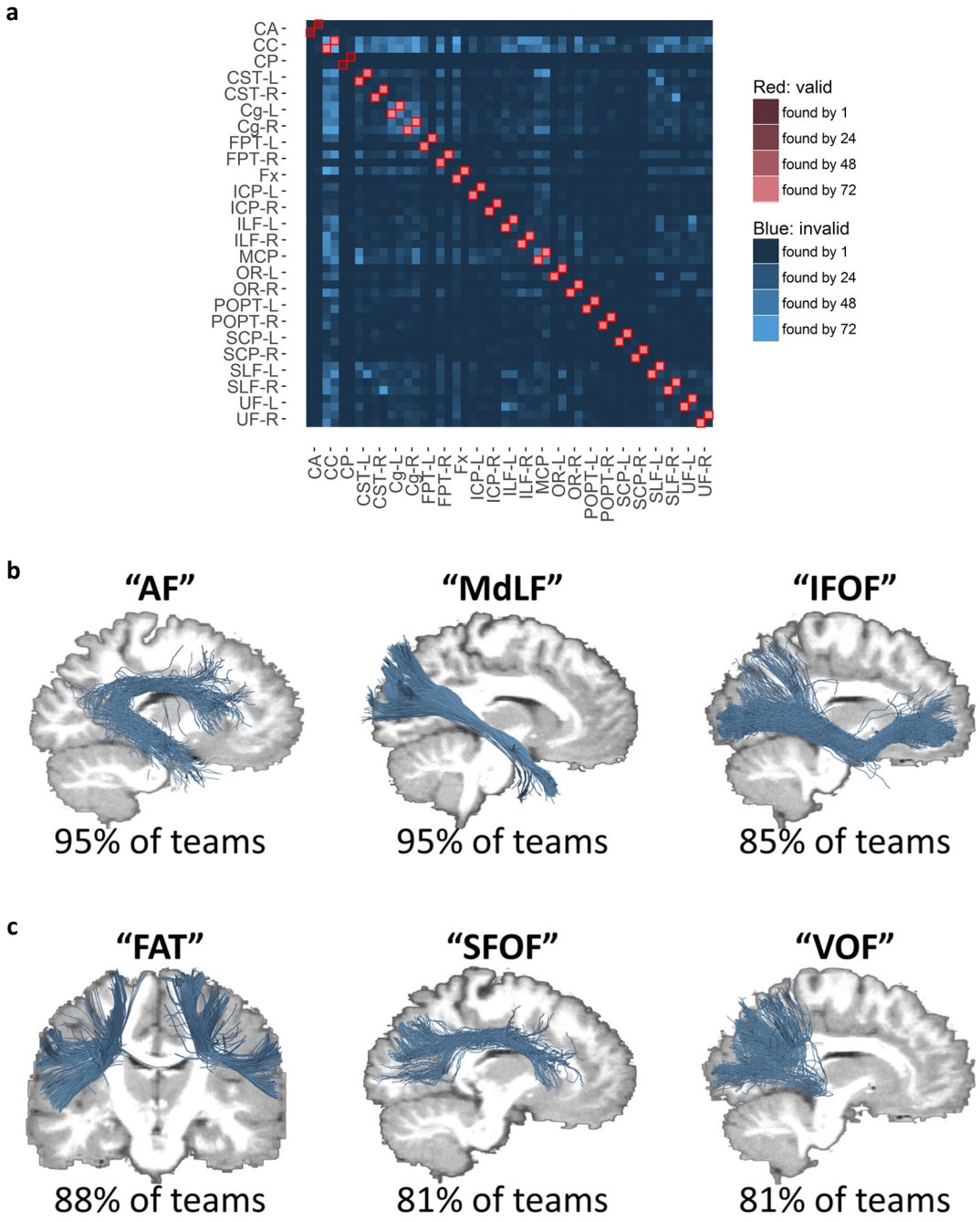

Figure 4. Overview of valid (red) and invalid (blue) bundles. Clusters of invalid streamlines exhibit similarities to previously reported, yet unconfirmed bundles from the literature. (a) Each entry in the connectivity matrix indicates the number of submissions that have identified the respective bundle. The two rows/columns per bundle represent the head-endpoint and tail-endpoint regions of a bundle. (b, c) The shown invalid bundles have been consistently identified by more than 80% of the teams, but do not exist in the ground truth. As discussed in the main text, some of these bundles exhibit a high similarity to bundles that are known to exist, as shown in (b), including the arcuate fasciculus (AF) and bundles traversing the temporal stem such as the middle longitudinal fasciculus (MdLF) and the inferior-frontal occipital fasciculus (IFOF). Other bundles – shown in (c) – are not known to exist and are debated in literature. These include the frontal aslant tract (FAT), the superior fronto-occipital fasciculus (SFOF) and the vertical occipital fasciculus (VOF).

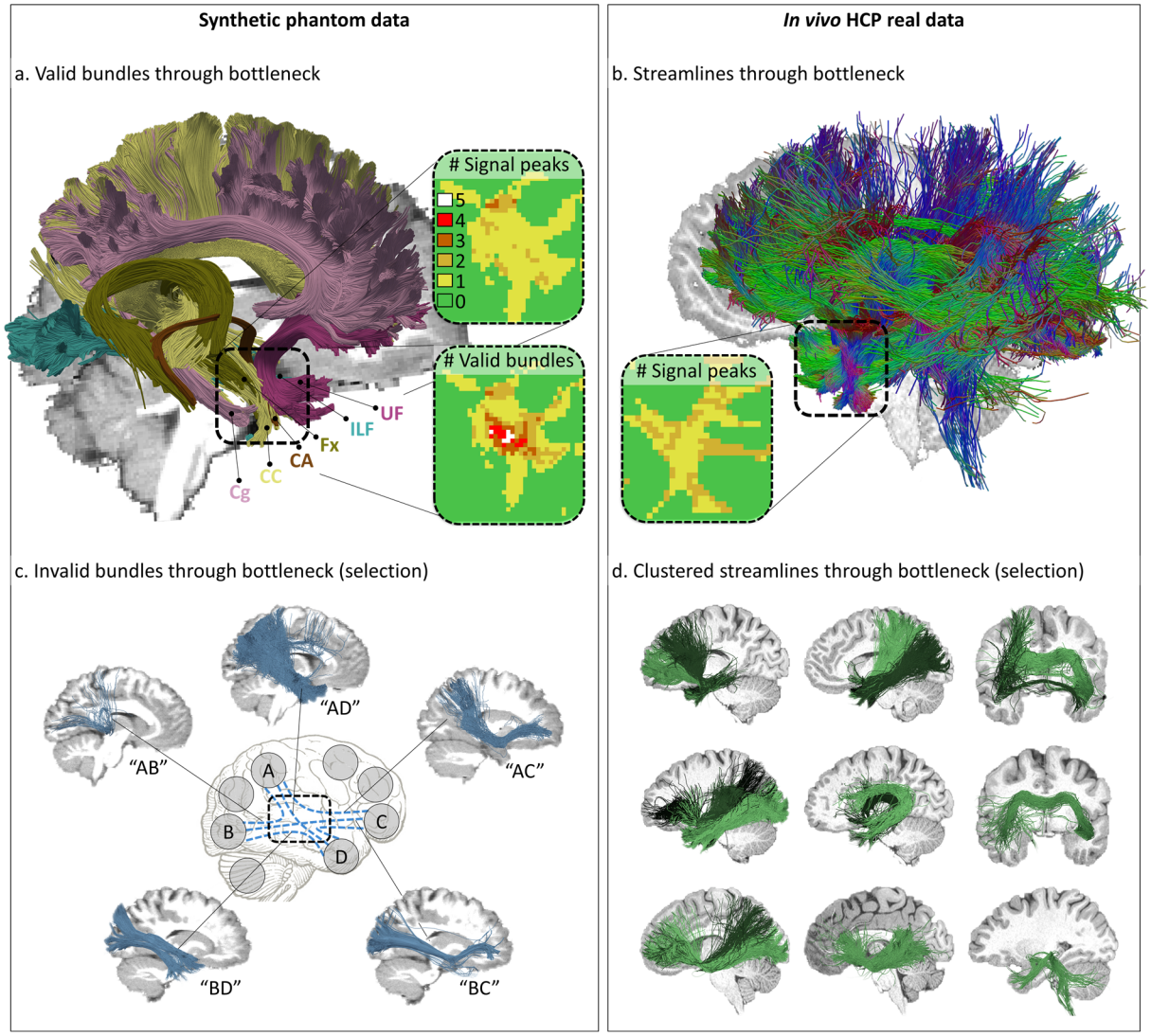

Figure 5. Fundamental ill-posedness of tractography. (a) Visualization of six ground truth bundles converging into a nearly parallel funnel in the bottleneck region of the left temporal lobe (indicated by square region). The bundles per voxel (box "# Valid bundles") clearly outnumber the peak directions in the diffusion signal (box "# Signal peaks"). (b) Visualization of streamlines from a Human Connectome Project (HCP) *in vivo* tractogram passing through the same region. (c) Exemplary invalid bundles that have been identified by more than 50% of the submissions, showing that tractography cannot differentiate between the massive amount of plausible combinatorial possibilities connecting different endpoint regions (see Supplementary Video 1). (d) Automatically Quickbundle-clustered streamlines from the *in vivo* tractogram going through the temporal region of interest. The clustered bundles are illustrated in different shades of green. These clusters represent a mixture of true-positive and false-positive bundles going through that bottleneck area of the HCP data set (see Supplementary Video 2).