Model Averaging and Bootstrap Consensus-based Uncertainty Reduction in Diffusion MRI Tractography

J. Gruen, [iD] G. van der Voort and T. Schultz[*] [iD]

University of Bonn, Bonn, Germany
jgruen@uni-bonn.de, gemmavdvoort@gmail.com, schultz@cs.uni-bonn.de

Abstract

Diffusion magnetic resonance imaging (dMRI) tractography has the unique ability to reconstruct major white matter tracts non-invasively and is, therefore, widely used in neurosurgical planning and neuroscience. In this work, we reduce two sources of uncertainty within the tractography pipeline. The first one is the model uncertainty that arises in crossing fibre tractography, from having to estimate the number of relevant fibre compartments in each voxel. We propose a mathematical framework to estimate model uncertainty, and we reduce this type of uncertainty with a model averaging approach that combines the fibre direction estimates from all candidate models, weighted by the posterior probability of the respective model. The second source of uncertainty is measurement noise. We use bootstrapping to estimate this data uncertainty, and consolidate the fibre direction estimates from all bootstraps into a consensus model. We observe that, in most voxels, a traditional model selection strategy selects different models across bootstraps. In this sense, the bootstrap consensus also reduces model uncertainty. Either approach significantly increases the accuracy of crossing fibre tractography in multiple subjects, and combining them provides an additional benefit. However, model averaging is much more efficient computationally.

Keywords: diffusion MRI, tractography, uncertainty, model averaging, bootstrapping

CCS Concepts: • Applied computing → Life and medical sciences; • Mathematics of computing → Probabilistic algorithms; • Human-centred computing → Visualization techniques

1. Introduction

Diffusion magnetic resonance imaging (dMRI) [LBL*86] is a non-invasive imaging method for the human brain. It allows for unique insights into the geometry and microstructure of major white matter tracts by measuring the Brownian motion of water molecules. Since the fibre tracts impede their movement, molecules move along the fibre tracts more freely than orthogonal to it. This has permitted the reconstruction of many white matter tracts using tractography algorithms, and has made dMRI an important tool for large scientific studies [SJX*13, TSH*18] as well as surgery planning [YYPC21].

The most popular and widely used tractography algorithms recover the local orientation of fibre tracts from dMRI measurements. Earlier approaches like diffusion tensor imaging (DTI) [BML94] were just able to recover a single direction per voxel, which is not sufficient to recover more complex geometries like fibre crossing,

kissing, and bending. Newer approaches rely on high angular resolution imaging (HARDI), and typically lead to ill-conditioned inverse problems. Popular mathematical models for the estimation of multiple local orientations include the ball-and-stick model [BBJ*07], spherical deconvolution [TCC07], and the low-rank approximation of higher-order fODF tensors [SS08], which can be seen as combining aspects of the first two [SWK10].

Tractography is affected by many sources of uncertainty [SVBK14]. One of them is the model uncertainty that arises when having to make an *a priori* choice of the number of fibres in a given voxel. In models such as ball-and-sticks or low-rank tensor approximation, this is a crucial step, since setting the number too low will miss relevant directions, while setting it too high will introduce spurious directions and increase variance in the remaining directions.

We recently introduced the first framework for the quantification of this type of uncertainty, and a method for its reduction, which we refer to as *model averaging* [GvdVS21]. Our current work reports a refined implementation of that idea, and makes the following additional contributions:

[*]Corresponding author

1. We study how the above-described model uncertainty interacts with data uncertainty, which arises due to measurement noise and is commonly estimated via bootstrapping [CLH06].
2. We design a new approach for the joint reduction of data and model uncertainty based on a bootstrap consensus strategy, and compare it to our previous model averaging technique.
3. We evaluate a combination of both ideas and compare it to results from a novel baseline, which simply estimates the maximum number of fibres everywhere.

The remainder of our paper is organized as follows: We first provide the required background by discussing related work (Section 2), and summarizing the crossing fibre model on which our work is based (Section 3). We then introduce our two main approaches, model averaging and bootstrap consensus, in Sections 4 and 5, respectively, and describe an algorithm for dMRI tractography that uses them in Section 6. We report and discuss our experimental results in Section 7, before Section 8 summarizes our main findings and concludes our work.

2. Related Work

There is a substantial body of literature on algorithms for diffusion MRI tractography [JDML19], and many of them were first introduced in visualization venues [WKL99, ZDL03, HS05, SS08]. The visualization and reduction of various sources of uncertainty in the tractography pipeline has been a more recent focus of interest [BVPtH09, BPtHV13, SVBK14, WSSS14, SHV21, GvdVS21]. These sources can broadly be categorized into measurement uncertainty, model uncertainty, parameter uncertainty, and partial voluming [SVBK14, SV19, GSWS21].

The impact of measurement uncertainty on the dMRI pipeline has been widely estimated with probabilistic tractography, based on Bayesian modelling [BBJ*07] or bootstrapping [Jon08]. Instead of a single streamline per seed, this recovers distributions, which can be visualized using hyperstreamlines [JTE*05, JLT*12, WSSS14] or confidence intervals [BPtHV13, SHV21]. Our work explores a different use of bootstrapping, which performs uncertainty reduction by consolidating estimates from all bootstraps into a single consensus that is used for tracking.

We refer to model uncertainty as the uncertainty which arises from the choice between several mathematical models to extract directions from the dMRI data [SVBK14]. There exists a wide range of models to estimate fiber directions [PSS*12], and they might lead to different results. Comparative visualization has been used to investigate such differences [VVL13, SSSS13]. There is no generally preferable model, since the suitability depends on the dMRI acquisition scheme, as well as on the anatomical location [BKN04, FÖK*07]. Our current work significantly extends a recent workshop paper that investigated a special case of model uncertainty, focusing on the aspect of selecting a suitable voxel-specific fibre direction count [GvdVS21].

Another type of uncertainty arises from parameter choices within the tracking algorithm itself, which for example control branching to reproduce fibre spread, or the termination of individual streamlines. Brecheisen *et al.* proposed a visual tool to systematically explore the impact of such parameters [BVPtH09]. Since optimal settings depend on the specific tract, Takemura *et al.* developed an ensembling approach that selects streamlines from candidates that have been generated with different algorithms and parameters [TCWP16].

Finally, the partial volume effect is an important source of uncertainty in dMRI. It arises from the fact that the diameter of individual axons is orders of magnitude smaller than the spatial resolution of dMRI. Even when correctly accounting for cases in which axons cross [AHL*01, BBJ*07] or spread [KKA07] at a voxel level, situations in which two distinct tracts become locally aligned pose a fundamental difficulty for finding their correct continuation. This has been referred to as the bottleneck issue [STR*22], and it contributes to the fact that, even though dMRI tractography quite successfully localizes true tracts in individual subjects, it tends to produce many false positives [MNH*17]. These have to be eliminated using prior anatomical knowledge, which can be represented implicitly using machine learning [WNM18], or explicitly by defining regions of interest to include or exclude streamlines [WCP*07]. Our work employs the latter approach.

3. Background on Low-Rank fODF Tensor Approximation

Earliest approaches for dMRI tractography used the diffusion tensor model [MCCvZ99], and were, therefore, limited to a single dominant fiber orientation per voxel. However, most white matter voxels contain more than one fibre bundle [JLT*12], and it has been shown that accounting for these more complex fibre geometries greatly improves dMRI tractography [NDH*15].

Our work builds on a previously described low-rank approximation of fODF tensors to infer multiple fibre orientations per voxel [SS08, ALGS17]. It first obtains a symmetric fourth-order tensor representation $\mathcal{T}$ of the fibre orientation distribution function (fODF) via constrained spherical deconvolution, which amounts to solving the linear least squares problem

$$\arg \min_{\mathcal{T}} \|M\mathcal{T} - S\|^2, \tag{1}$$

where M denotes a convolution matrix that is based on the dMRI response from a single fibre compartment, vector S contains all dMRI measurements in a given voxel. A set of r fibre directions is then estimated as unit vectors $\mathbf{v}_i \in \mathbb{R}^3$ with volume fractions λ_i for $i \in \{1, \ldots, r\}$ via a rank-r approximation

$$\mathcal{T}^{(r)} = \sum_{i=1}^{r} \lambda_i \mathbf{v}_i \otimes \mathbf{v}_i \otimes \mathbf{v}_i \otimes \mathbf{v}_i, \tag{2}$$

where $\otimes$ denotes the outer product. The approximation is formalized as

$$\arg \min_{\lambda_1, \ldots, \lambda_r, \mathbf{v}_1, \ldots, \mathbf{v}_r} \|\mathcal{T} - \mathcal{T}^{(r)}\|_F,$$

where $\| \cdot \|_F$ denotes the Frobenius norm.

These steps represent a refinement of a more traditional approach to spherical deconvolution, which is based on spherical harmonics, a soft non-negativity constraint, and taking local fODF maxima as fibre directions [TCC07]. It has been demonstrated that replacing peak extraction with low-rank approximation reduces angular errors, and increases angular resolution, even when replacing the more
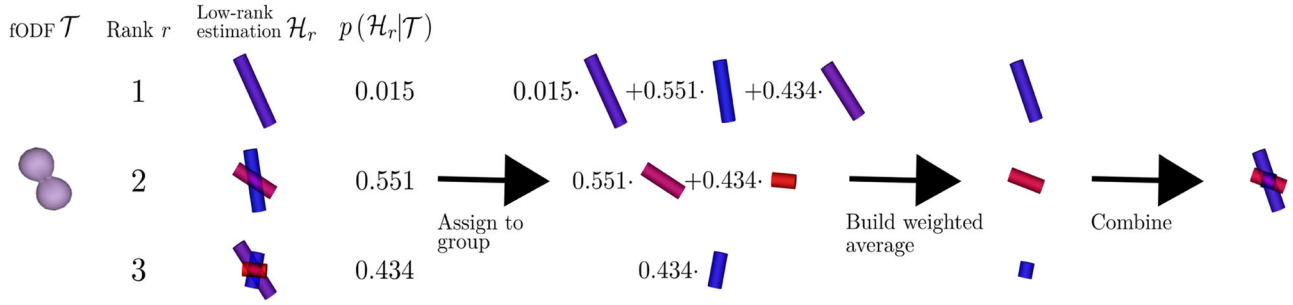
Figure 1: *Illustration of our model averaging strategy. For a given fODF $\mathcal{T}$, low-rank approximations with ranks $r \in \{1, 2, 3\}$ are computed. The resulting directions are clustered, and a weighted average is taken within each group, with weights given by the posterior model probabilities $p(\mathcal{H}_r \mid \mathcal{T})$ from a Bayesian framework. This results in a combined model in which secondary and tertiary fibres fade in and out smoothly in regions of model uncertainty, rather than vanishing abruptly, as when selecting the most likely model.*

traditional spherical harmonics of degree eight with fourth-order tensors, which have much fewer degrees of freedom, and greatly improve the conditioning of the least squares problem [ALGS17].

The non-negativity constraint in spherical deconvolution is justified by the fact that the fODF represents the fibre fraction in any given direction, which must be non-negative. In the tensorial framework, it is implemented by imposing a positive semi-definiteness constraint on $\mathcal{T}$, which also ensures that all λ_i in Equation (2) will be non-negative [ALGS17].

4. Model Averaging and Selection

In many voxels, it is unclear how many fibre directions are supported by the data, i.e., which rank r should be selected in Equation (2). Selecting the most likely number incurs the risk of underestimating r, which could lead to premature termination of a secondary tract. On the other hand, overestimating r confounds the tracking with spurious directions and increases the variance of the true fibre estimates. The idea behind model averaging is that smoothly blending between models, as it is illustrated in Figure 1, should reduce the risk of missing an important fibre, while still relying on the more robust single-fibre estimates in regions that clearly do not support a larger number of directions. We implement this idea based on Bayesian model comparison (Section 4.1), and derive specific model averaging and model selection approaches in Section 4.2.

4.1. Model comparison in a Bayesian framework

In Bayesian model comparison, we are interested in the posterior probability $p(\mathcal{H}_r \mid \mathcal{T})$, where $\mathcal{H}_r$ denotes the hypothesis that extracting r fibres is optimal for a given fODF $\mathcal{T}$. Using Bayes' theorem of conditional probability, up to a common factor that we account for by a final re-normalization, the posterior probability can be rewritten as

$$p(\mathcal{H}_r \mid \mathcal{T}) \propto p(\mathcal{T} \mid \mathcal{H}_r)p(\mathcal{H}_r), \qquad (3)$$

where $p(\mathcal{H}_r)$ is our prior belief that rank r is suitable, without considering the fODF. Since literature values for the prevalence of different r over the white matter vary [BBJ*07, JLT*12, Sch12], we

use a non-informative prior that assigns equal prior probability to the values of $r \in \{1, 2, 3\}$. The case $r = 0$ can be excluded since we limit tracking to a white matter mask.

The factor $p(\mathcal{T} \mid \mathcal{H}_r)$ is the probability of the fODF $\mathcal{T}$ given a rank r. In the context of Bayesian model comparison, it is referred to as model evidence. It is derived from $p(\mathcal{T} \mid \mathcal{H}_r, \Theta_r)$, the posterior probability of T given an r-fibre model with a specific parameter vector Θ_r. In our case, Θ_r contains the variables from Equation (2), that is, $\Theta_r := (\lambda_1, \mathbf{v}_1, \ldots, \lambda_r, \mathbf{v}_r)$.

The overall model evidence is obtained by marginalization over parameter values,

$$p(\mathcal{T} \mid \mathcal{H}_r) = \int p(\mathcal{T} \mid \mathcal{H}_r, \Theta_r)p(\Theta_r \mid \mathcal{H}_r)d\Theta_r. \qquad (4)$$

Since a direct calculation of Equation (4) would require solving a high-dimensional integral, we use an approximation via the Bayesian information criterion

$$\text{BIC} = k \ln(n) - 2 \ln\left(p(\mathcal{T} \mid \mathcal{H}_r, \hat{\Theta}_r)\right),$$

where $p(\mathcal{T} \mid \mathcal{H}_r, \hat{\Theta}_r)$ corresponds to the likelihood of the rank-r with parameters $\hat{\Theta}_r$ that best fit the fODF $\mathcal{T}$, k is the number of parameters in Θ_r, and n denotes the number of data points to which the model was fitted [Sch78]. We note that $k = 3r$ increases with r, which penalizes the choice of multiple fibres, unless it leads to a sufficient increase of $p(\mathcal{T} \mid \mathcal{H}_r, \hat{\Theta}_r)$. Under certain conditions, the BIC is related to the model evidence by Konishi and Kitagawa [KK08]

$$p(\mathcal{T} \mid \mathcal{H}_r) \approx \exp\left(-\frac{\text{BIC}}{2}\right). \qquad (5)$$

This allows us to compute the model evidence in a simple and efficient way. However, we still need to provide an equation for $p(\mathcal{T} \mid \mathcal{H}_r, \hat{\Theta}_r)$. Therefore, we use the relative magnitude of the corresponding low-rank approximation residual

$$\|\tilde{\mathcal{R}}^{(r)}\| = \frac{\|\mathcal{T} - \mathcal{T}^{(r)}\|}{\|\mathcal{T}\|} \in [0, 1], \qquad (6)$$
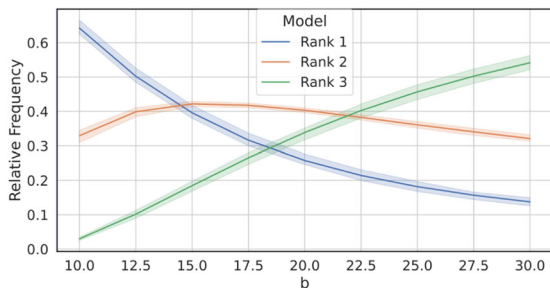
Figure 2: *Relative frequency of selecting ranks* $\{1, 2, 3\}$ *within a white matter mask based on the Kumaraswamy PDF with* $a = 1$ *and different values of* b. *The lines indicate the mean over all subjects from Section 7.1, the tubes the minimum and maximum.*

since a smaller residual from a rank-r approximation should indicate a higher probability of $\mathcal{T}$ being a perturbation of a rank-r tensor. Since many factors contribute to the magnitude of this residual, including measurement noise, fibre spread, and inaccuracies in the convolution kernel, we pragmatically model $p(\mathcal{T} \mid \mathcal{H}_r, \hat{\Theta}_r)$ with the computationally efficient Kumaraswamy Probability Density Function (PDF) [Kum80]

$$f(x, a, b) := abx^{a-1}(1 - x^a)^{b-1} \text{ for } x \in (0, 1) \text{ and } a, b > 0 \quad (7)$$

which is defined on the correct interval (0,1), and can be tuned to achieve the correct qualitative behaviour of monotonically decreasing with increasing $\|\tilde{\mathcal{R}}^{(r)}\|$ by setting $a = 1, b > 2$.

With increasing b, the probability decreases at steeper slopes. As illustrated in Figure 2, this favours modelling larger numbers of fibres, since that permits a reduction of fitting residuals. Thus, the choice of b can be seen as determining how strong the support for a secondary or tertiary compartment needs to be so that it will be used for tracking. Having to make this decision is a source of uncertainty in all methods for crossing fibre tractography, and has contributed to the variability in literature values on the prevalence of two- and three-fibre voxels [BBJ*07, JLT*12, Sch12]. In our experiments, we set $b = 20$, which leads to values within the range of literature values. We expect that the normalization that results from deconvolving with subject-specific response functions [JTD*14] and the normalization in Equation (6) will allow for similar choices of b across datasets. This is supported by the relatively narrow confidence intervals across subjects in Figure 2.

4.2. Computing tracking directions from alternative models

As in our previous work [GvdVS21], we fuse the information from tensor approximations with different ranks $r \in \{1, 2, 3\}$ by taking a weighted sum of the corresponding parameters $\mathbf{v}_i^{(r)}$ and $\lambda_i^{(r)}$ with weights given by the above-defined posterior probabilities. We refer to this strategy as *model averaging*.

Before computing the weighted sum, we have to establish a correspondence between the directions of the different r-fibre models. In other words, we have to re-order the directions from each rank r such that the first directions of the two- and three-fibre models are matched with the direction from the single fibre model, and that the

second directions of the two- and three-fibre models match. This leads to $2! \times 3! = 12$ possible assignments, from which we select the one that minimizes the overall sum of angles between the resulting weighted means $\mathbf{v}_i$ and their corresponding $\mathbf{v}_i^{(r)}$.

A potential objection against the idea of model averaging is that, in case multiple probable models should yield very dissimilar directions, averaging them might produce spurious directions that did not occur in any of the original models. Therefore, we computed the smallest distances between the main direction of the average model and the closest direction of the corresponding low-rank models within the white matter mask of a randomly chosen subject from the experiment in Section 7.1. In 95% of all white matter voxels, that angle was below $1.8°$. Since that is well below the angular resolution of CSD [ALGS17], we conclude that, in practice, corresponding directions are similar enough that it is reasonable to average them.

Previous tractography algorithms that were based on low-rank tensor approximation [ALGS17] used a different strategy, that we refer to as *model selection*: They determined an optimal rank $r \in \{1, 2, 3\}$ in each integration step and used the resulting set of directions $\mathbf{v}_i$ for tracking. In cases where several ranks have non-negligible probabilities, this introduces an uncertainty that model averaging aims to reduce. We include this approach as a baseline in our experiments. To enable a direct comparison, we select the model with the highest probability $p(\mathcal{H}_r \mid \mathcal{T})$ according to the same Bayesian framework.

5. Bootstrapping for Uncertainty Estimation and Reduction

We describe the variability that arises from noise in the dMRI measurement as data uncertainty. Taking measurements repeatedly is a natural way to estimate it, but is usually not feasible in practice. Therefore, different bootstrapping techniques have been established to estimate data uncertainty based on a single DTI or HARDI acquisition [CLH06, Jon08].

We summarize one such bootstrapping approach in Section 5.1, and use it in two ways: In Section 5.2, we investigate the interaction between data and model uncertainty, by studying the variance of the selected model under bootstrapping. In Section 5.3, we define a bootstrap consensus for the joint reduction of data and measurement uncertainty. As a by-product, this allows us to investigate the effect of model selection and averaging on the variability of fibre direction estimates in Section 5.4.

5.1. Wild bootstrapping

We use wild bootstrapping to evaluate the impact of measurement noise. It is a straightforward way to re-sample the original measurements without repeating the measurement process, and has been used widely in the context of dMRI [Jon08, SSSS13, SHV21].

Wild bootstrapping uses model residuals to estimate noise. We combine it with the fODF tensor framework from Section 3 by fitting an initial fODF $\hat{\mathcal{T}}$ via Equation (1), and calculating the residual

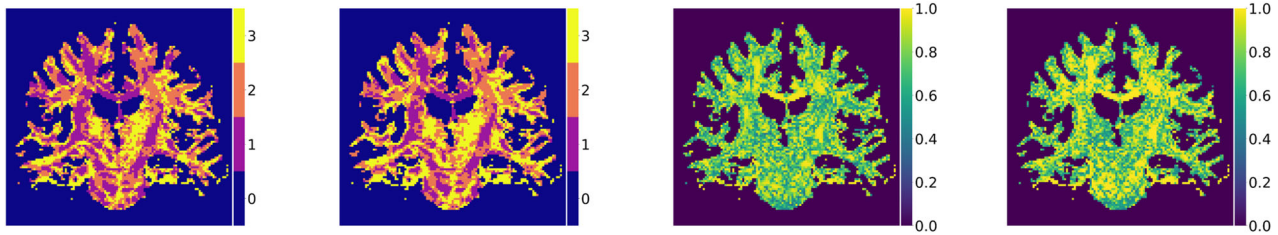$$\hat{\varepsilon} = S - M\hat{\mathcal{T}}.$$

Figure 3: *Small differences between model selection without bootstrapping (a) and the most frequently selected model under bootstrapping (b) indicate that data uncertainty also has a certain effect on the selected model. The stability of model selection under bootstrapping (d) correlates well with the confidence derived from our Bayesian framework (c).*

A new bootstrap realization is calculated by

$$y^* = M\hat{\mathcal{T}} + \hat{\varepsilon} \odot v,$$

where v denotes a random draw from the n dimensional Rademacher distribution

$$f(\mathbf{k}_i) := \begin{cases} {}^1\!/\!_2 \text{ if } \mathbf{k}_i = -1 \\ {}^1\!/\!_2 \text{ if } \mathbf{k}_i = 1 \\ 0 \text{ otherwise} \end{cases} \quad \text{for } i \in \{1 \dots n\}.$$

and $\odot$ denotes component-wise multiplication.

This process is repeated m times to create a sample of m bootstraps. For each bootstrap realization, we fit a new fODF and compute its low-rank approximations.

5.2. Interaction between data and model uncertainty

We hypothesized that data and model uncertainty interact in the sense that, in many voxels, the effect of measurement noise is sufficient to change the selected fibre number. We were also interested in the extent to which the frequencies with which different ranks are selected under bootstrapping agree with the probabilities that are assigned by our Bayesian model.

Therefore, we took 100 bootstraps and applied model selection to all of them. Comparing the most likely number of fibres in the original data (Figure 3a) to the number that is most frequently selected over all bootstraps (Figure 3b) already reveals a certain amount of variability: In 14.1% of the voxels within the white matter, the selected rank differs by 1; in 0.8%, it differs by 2.

The probability of the most likely number of fibres according to Equation (3) is shown in Figure 3c, next to the fraction of bootstraps that agreed on the most frequently selected number in Figure 3d. Visually, we find a remarkable agreement between the regions in which our Bayesian framework has a high confidence in its choice (high values in c), and the regions in which the choice is stable under bootstrapping (high values in d). We believe that this provides further empirical support for the framework proposed in Section 4.1.

5.3. Bootstrap consensus

In analogy to how the model averaging in Section 4.2 reduces model uncertainty by fusing information from alternative models, we derive a bootstrap consensus that fuses information from all bootstraps. We emphasize that this differs from the traditional use of bootstrapping in probabilistic tractography, where bootstrap samples produce a distribution of streamlines from a common seed [Jon08, JLJ*11]. Instead of estimating data uncertainty, our bootstrap consensus aims to reduce its effect. Due to the interaction between data and model uncertainty that was observed in Figure 3, it can also be expected to jointly reduce model uncertainty.

The bootstrap consensus is formed by clustering the directions from all bootstraps into groups, and taking the group means as the final tracking directions. Specifically, we assign the n_i directions from the ith bootstrap to m groups, where we set $m = 3$ as the maximum number of fibres, we are willing to consider. Due to the large number of bootstraps, an enumeration of all possible assignments that guarantees a global optimum, as it could be used in Section 4.2, is no longer feasible.

Instead, we initialize the reference directions for each group with a rank-3 approximation of the original data. Now, the directions from each bootstrap are assigned to the groups such that the sum of distances between the group reference and the directions are minimized over all possible assignments. Formally, we minimize the objective function

$$T : \text{Sym}(n) \mapsto \mathbb{R}_+$$
$$Z \to \sum \| \text{sgn} \left(\langle \mathbf{v}_{Z(i)}, \bar{\mathbf{v}}_i \rangle \right) \mathbf{v}_{Z(i)} - \bar{\mathbf{v}}_i \|, \qquad (8)$$

where $\text{Sym}(n)$ denotes the symmetric group, $\bar{\mathbf{v}}_i$ denotes the reference direction, $\mathbf{v}_i$ denotes the direction from the bootstrap.

5.4. Visualizing the uncertainty in fibre directions

Our main motivation for using model selection or averaging instead of always extracting the maximum number of fibres is that we expect it to reduce variance in the true fibre estimates. The clustering that we perform to compute the bootstrap consensus can be used to quantify and visualize this effect. The Watson distribution is defined as [JM99]

$$f : \mathbb{S}^2 \longrightarrow \mathbb{R}_+$$

(a) Model Selection
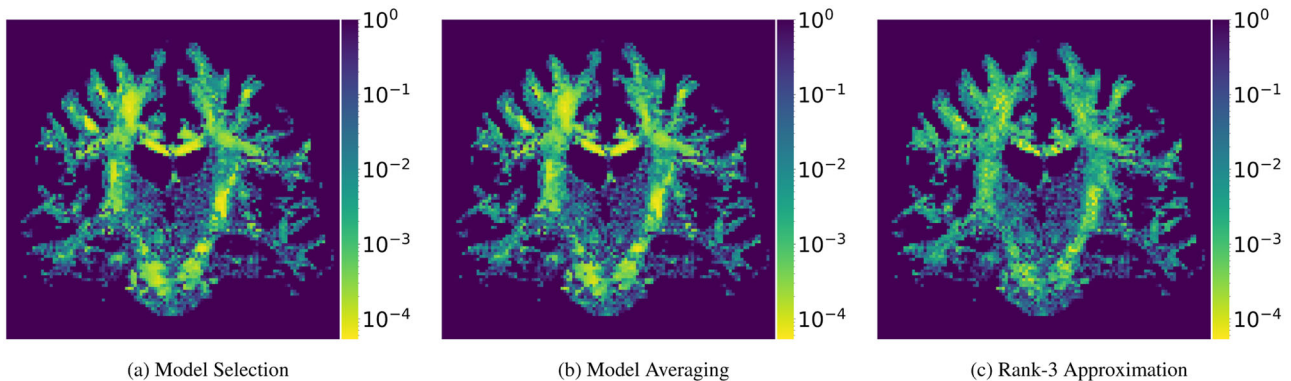
(b) Model Averaging

(c) Rank-3 Approximation

Figure 4: *Mapping orientation dispersion of the principal fibre direction under bootstrapping confirms that the dispersion in the rank-3 result is higher than with model selection or averaging, indicating their ability to decrease susceptibility to noise.*

$$\mathbf{x} \longmapsto \frac{1}{M\left(\frac{1}{2}, \frac{3}{2}, \kappa\right)} \exp\left(\kappa\left(\mu^T\mathbf{x}\right)^2\right),$$

where κ is the dispersion parameter, $\|\mu\| = 1$ the mean direction, and the normalizing factor involves the confluent hypergeometric function M [Kum37]. To estimate the dispersion κ, we used the maximum likelihood estimator [JM99], which allowed us to determine κ via Newton optimization.

We use the Watson distribution for two reasons: First, it is antipodally symmetric, which fits the fact that in Equation (2), directions $\pm\mathbf{v}$ are indistinguishable. Second, its dispersion parameter κ allows us to quantify the variability in the fibre directions within each group. κ can be re-parameterized to obtain an orientation dispersion index

$$\text{OD} = \frac{2}{\pi} \arctan\left(\frac{1}{\kappa}\right)$$

that is normalized to $(0,1)$ and indicates higher variability with higher values [ZSGA12].

Figure 4 visualizes orientation dispersion of the primary fibre that was estimated with model selection, model averaging, or a rank-3 approximation. In particular in the regions in which model selection picks a single fibre (cf. Figure 3; this includes parts of the corpus callosum (CC) and CST), this results in lower dispersion compared to the rank-3 model, confirming our expectation that estimating a single fibre direction is less susceptible to noise than estimating multiple directions from the same data. The same regions still have reduced dispersion after model averaging, confirming its effectiveness for uncertainty reduction.

6. Crossing Fibre Tractography with Uncertainty Reduction

6.1. Probabilistic streamline-based tractography

Our work reduces the data and model uncertainty that affects local fibre directions. Despite this, the fundamental uncertainty that arises from partial volume effects remains, as discussed in Section 2: Even if multiple fibre directions were estimated with perfect accuracy, it

remains uncertain which of them should be followed in each step of the tracking. In regions where tracts fan out, there might even be multiple valid ways to continue them. Previous work used deterministic tracking with a branching mechanism to handle these situations [ALGS17]. In our work, we follow a probabilistic approach instead. However, its probabilistic nature is limited to deciding which of multiple fibre compartments should be followed next. Unlike most other probabilistic tractography methods [BBJ*07, Jon08], it does not sample a distribution of fibre directions *within* each fibre compartment, since this type of uncertainty is reduced by our bootstrap consensus.

Our tracking acts on multi-vector fields that are pre-computed using any of the methods proposed above. For a given seed point, the streamline is grown iteratively in both directions using Euler integration, which was sufficient to reconstruct even the curved part of the cingulum (CG) bundle at a step size of 0.9 mm. The multi-vector fields are interpolated to the current position. To keep computational effort manageable, we use trilinear interpolation. It requires solving another matching problem, since we have to decide which directions belong together during interpolation. Assuming smoothness between the voxels, we use the r directions from the last interpolation step as initial group means $\bar{\mathbf{v}}_i$, $i \in \{1 \ldots r\}$ and assign the vectors from all grid points that are involved in the current interpolation by minimizing the same cost function as in Equation (8). If fewer than three vectors are available due to model selection, zero vectors replace the missing ones for the purpose of interpolation. After all assignments have been made, we re-orient all vectors to have a non-negative dot product with their current group mean, compute new group means from them, and iterate the assignment once. To reduce the computational expense, we cache the final assignments, so that they only need to be computed once for each voxel through which we are tracking.

To investigate how this matching step affects the stability of interpolation results, we rotated all vectors at all voxels independently and uniformly at random by up to 10°, and computed the resulting change in tracking direction at 1000 randomly selected streamline vertices, drawn from the experiment in Section 7.3. When interpolating the perturbed directions, but keeping the original correspondences, 95% of the deviations were below 5.5°. Re-matching the

perturbed direction increased this only slightly, to 6.3°. We conclude that ambiguous correspondences contribute relatively little to the overall uncertainty in tracking directions.

Given the interpolated directions $\mathbf{v}_i$ at the current point, we reorient them to have a non-negative inner product with the current tracking direction $\mathbf{w}$. We select one of the r possible directions by assigning each unit direction $\mathbf{v}_i$ with volume fraction λ_i for $i \in \{1, \dots, r\}$ a probability following the probability scheme

$$p(\mathbf{v}_i) := \frac{\mathbb{1}_{\{\theta_i < \frac{1}{4}\pi\}}\lambda_i \cos\left(\left(\frac{9}{2\sqrt{2\pi}}\theta_i\right)^2\right)^2}{\sum_j \mathbb{1}_{\{\theta_j < \frac{1}{4}\pi\}}\lambda_j \cos\left(\left(\frac{9}{2\sqrt{2\pi}}\theta_j\right)^2\right)^2}, \qquad (9)$$

where θ_i denotes the angle between the possible direction $\mathbf{v}_i$ and the current direction $\mathbf{w}$. The indicator function $\mathbb{1}_{\{\theta_i < \frac{1}{4}\pi\}}$ restricts the maximum angle to 45°, to limit diversions into neighbouring tracts. We note that the use of trilinear interpolation allowed us to make this threshold slightly stricter compared to our prior work [GvdVS21]. Equation (9) assigns almost the same probability to directions with angles below 20°, which coincides with the limited angular resolution of spherical deconvolution [TCC07].

This iterative algorithm proceeds until we either reach a region with an overall white matter volume fraction below 0.3, or the summed angle over the last 30 mm is greater than 130°. This prevents streamlines from going back and forth. In the latter case, the entire streamline gets removed.

6.2. Postprocessing

While diffusion MRI is able to reconstruct most major white matter tracts, it is known to yield many false positives, which have to be removed according to anatomical knowledge [MNH*17].

Therefore, we define inclusion and exclusion regions for each tract. If a streamline intersects with an exclusion region, or if it does not intersect with all inclusion regions, the whole streamline is discarded. All regions are set carefully for a reference subject according to the protocols in Wakana *et al.* [WCP*07]. The remaining subjects are linearly registered to the reference using FSL's `flirt` [JBBS02], and the transformation is used to transfer the regions automatically.

Moreover, we remove obvious outliers by creating a density map for each streamline bundle. Therefore, we count the number of streamlines intersecting each voxel. All streamlines are cut off at the first intersection with a low density area starting from the seed. The density threshold is defined for the reference subject and then mapped to all other subjects according to the ratio of seed points.

7. Results

7.1. Data

We evaluate the proposed methods on data from the Human Connectome Project (HCP) [VSB*13]. The diffusion MR images have a resolution of 1.25 mm isotropic with $145 \times 174 \times 145$ voxels and have been acquired with 90 gradient directions each at $b \approx$

$\{1000, 2000, 3000\}\,\text{s/mm}^2$. We compare our results to reference tractographies which were published within the scope of the TractSeg paper [WNM18]. This reference data have been created by manually refining a segmented full brain fibre tractography. Similar to TractSeg, we apply a variant of multi-shell multi-tissue deconvolution [JTD*14, ALGS17] that uses all available diffusion weighted images. It also yields per-voxel estimates of white matter volume fractions, which we threshold at 0.3 to obtain the white matter masks used in our analyses.

All tests were performed on 12 randomly chosen subjects for which such reference tractographies exist. For each tract, we created seed points by intersecting the reference fibre bundle with a plane and initialize the tracking process with the direction of the fibre bundle at the seed point. This should mimic a directional region of interest as it might be defined by an expert on brain anatomy [GRNM16]. We then apply the tracking process until we have as many streamlines as the reference tractography. This should guarantee a fair comparison between the different models.

Our comparison is between generic fibre tracking algorithms that can track from any given seed region. In contrast, bundle-specific approaches such as TractSeg are limited to reconstructing a predefined set of bundles, but have the benefit of tract-specific anatomical priors, which improve their accuracy.

7.2. Qualitative comparison

We compare results from model selection and model averaging, with and without the additional use of the bootstrap consensus. Our previous work [GvdVS21] included results from standard constrained spherical deconvolution as another baseline. However, this did not achieve competitive results and its computational expense made it difficult to combine it with bootstrapping. Therefore, we replaced it by a new baseline, which uses low-rank approximation with rank 3 throughout the brain.

As a first experiment, we track the right corticospinal tract (CST). We attempt to reconstruct the reference result, shown in Figure 5a, from a seed region which is indicated with a dashed black line. To provide further guidance for visual comparison, the outline of the reference is overlaid on the remaining reconstructions as a black contour.

The main risk of a model selection strategy (f) is premature tract termination due to a local underestimation of the true fibre number. It is obvious that this prevented full recovery of the lateral spread of the CST. Model averaging (d) and bootstrap consensus (g) both lead to more complete reconstructions. The bootstrap consensus leads to an ever denser sampling of the spreading region even when combining it with model averaging or the rank-3 model. An important advantage of model averaging compared to the rank-3 model is the reduction of false positives. This is less obvious in the image, because most of them have been filtered out successfully by the mechanisms in Section 6.2. However, it is clear from running times, which will be reported in Section 7.4.

Inspecting the multi-vector fields in Figure 6 provides further insight. The most obvious difference between the average and selection models is that, in many voxels, model averaging (top left) leads
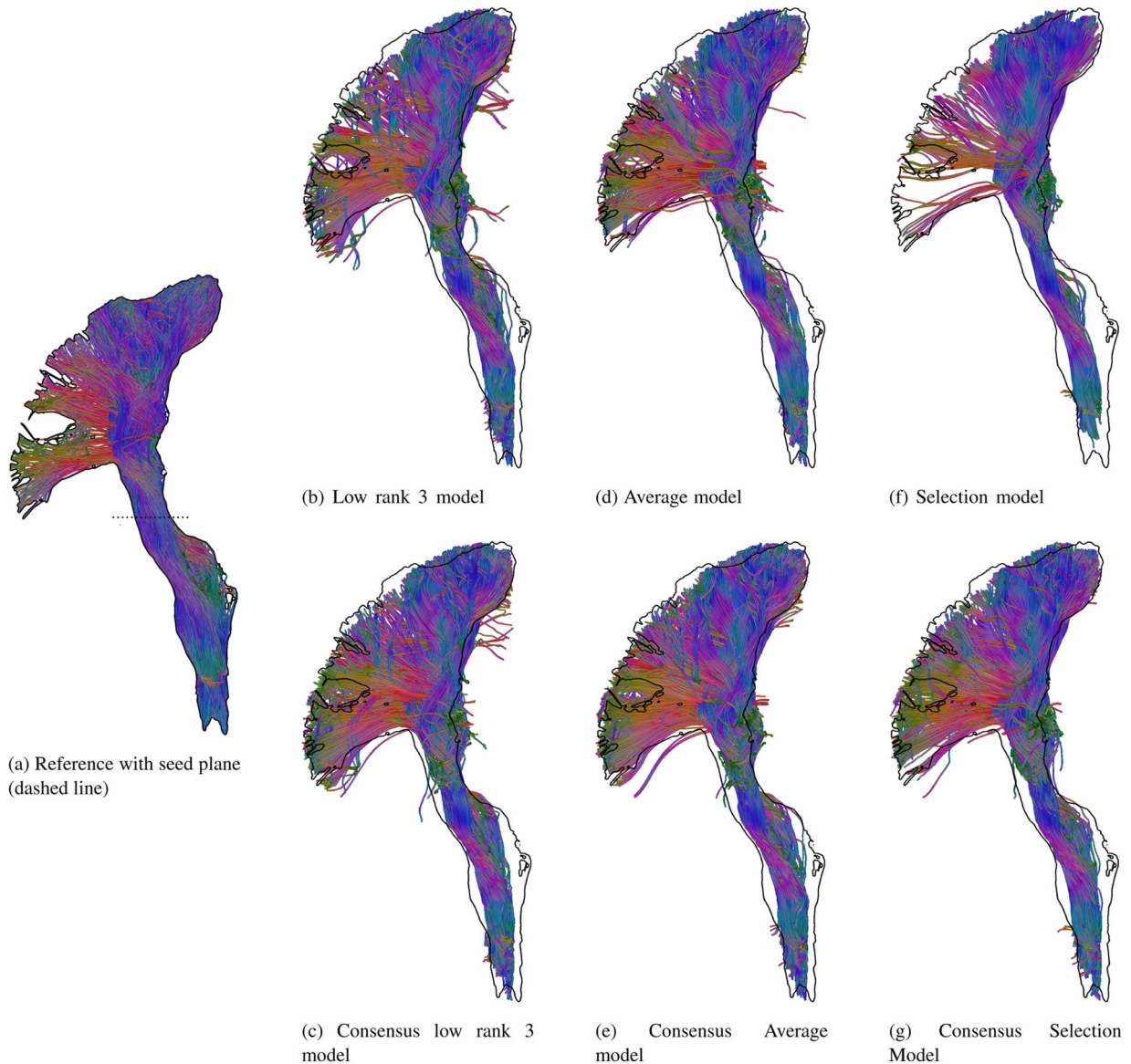
(b) Low rank 3 model

(d) Average model

(f) Selection model

(a) Reference with seed plane (dashed line)

(c) Consensus low rank 3 model

(e) Consensus Average model

(g) Consensus Selection Model

Figure 5: *Reconstruction of the right corticospinal tract. For a more direct comparison between the reference and the tractographies, we overlaid the contour of the reference on each reconstruction with a black curve. The bootstrap consensus (bottom) leads to a much higher ability to reconstruct the lateral spread compared to the base models (top). This is especially true for the selection model (right).*

to a larger number of fibres compared to model selection (top right). This explains the more complete reconstruction that was observed in Figure 5. On the other hand, applying the bootstrap consensus makes the results from averaging (bottom right) and selection (bottom left) rather similar. The red circle highlights a voxel in which results from model averaging and selection differ, but agree after forming a bootstrap consensus. This illustrates that the bootstrap consensus reduces not just data uncertainty, but also model uncertainty when it is combined with model selection.

Figure 7 provides qualitative results on another bundle, the optic radiation. Model selection (f) leads to an incomplete reconstruction of the spread in the posterior part, which is again more fully sampled

with model averaging. The bootstrap consensus further improves both models.

7.3. Quantitative comparison

For a quantitative evaluation, we reconstructed the corpus callosum (CC), the cingulum (CG), the CST, the inferior fronto-occipital (IFO) and the inferior longitudinal fasciculus (ILF), the optic radiation (OR), and the superior longitudinal fasciculus (SLF). To keep the analysis more manageable, we combined the left and right tracts, and merged the subtracts that were defined for CC and SLF in the reference [WNM18].
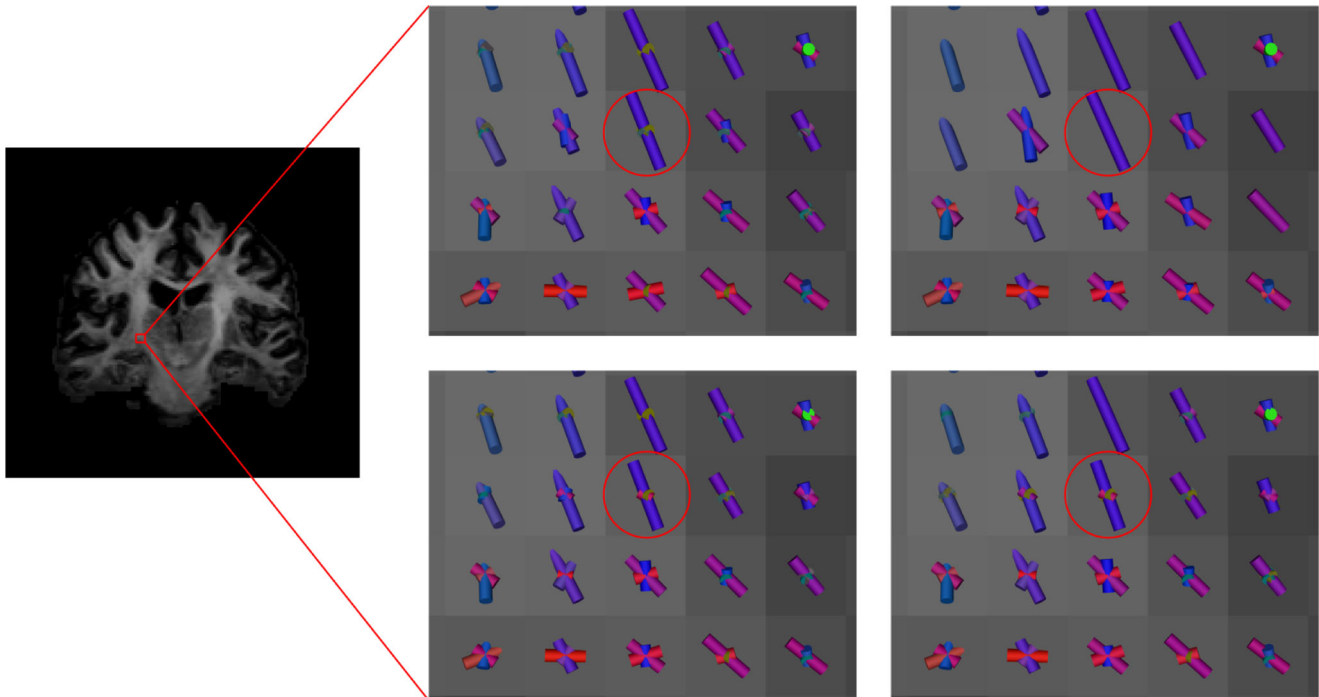
Figure 6: *Reconstructed fibre orientations from the different models. The red box in the left image denotes the position within the brain. Both consensus models (bottom row) agree quite well in most voxels, although results from model averaging (left) and model selection (right) differ, for example, in the voxel highlighted by the red circle.*

Table 1: *Comparison of all models by a Nemenyi post hoc test. All p values below 0.05 are marked in bold if the top model is significantly better, else in italics. Our model averaging strategy is significantly better than model selection and the rank 3 model. The bootstrap consensus further improves model selection and averaging..*

	Consensus average model	Selection model	Consensus selection model	Rank-3 model	Consensus rank-3 model
Average model	**0.001**	*0.001*	**0.002**	*0.001*	0.063
Consensus average model		*0.001*	0.900	*0.001*	*0.001*
Selection model			**0.001**	0.900	0.494
Consensus selection model				*0.001*	*0.001*
Rank-3 model					0.494

We evaluated the reconstruction quality by the Dice score which is defined as

$$\text{DICE} = \frac{2|RD \cap TR|}{|TR| + |RD|},$$

where RD denotes a binary mask that is derived from the reference data, and TR a mask from the tracking results [SDM*19]. A high Dice score is desired, as it denotes a high overlap and a low over-reach. In Figure 8, results from the average, selection and low-rank model with and without use of the bootstrap consensus are visualized as box plots to show the tract-wise distribution (over subjects) of the Dice scores. In most cases, we observe that model averaging yields a more accurate reconstruction compared to model selection, and that the bootstrap consensus increases Dice compared to the corresponding strategy without bootstrapping.

To further quantify these findings, we compared all six models with a Friedman test, to investigate whether differences between models are statistically significant [Fri37]. Since that was the case ($p \ll 0.001$), we applied a Nemenyi post hoc test to identify which models differ significantly [Nem63].

According to the results of that post hoc test, which are shown in Table 1, when comparing strategies without the bootstrap consensus, model averaging is significantly better than either model selection or the rank-3 model, while neither of the latter two are preferable to the other overall. From this, we conclude that model averaging as it was first described in our previous work [GvdVS21] provides a significant advantage.

Concerning the bootstrap consensus that is introduced in our current work, we observe in Figure 8 that, in six out of seven tracts,
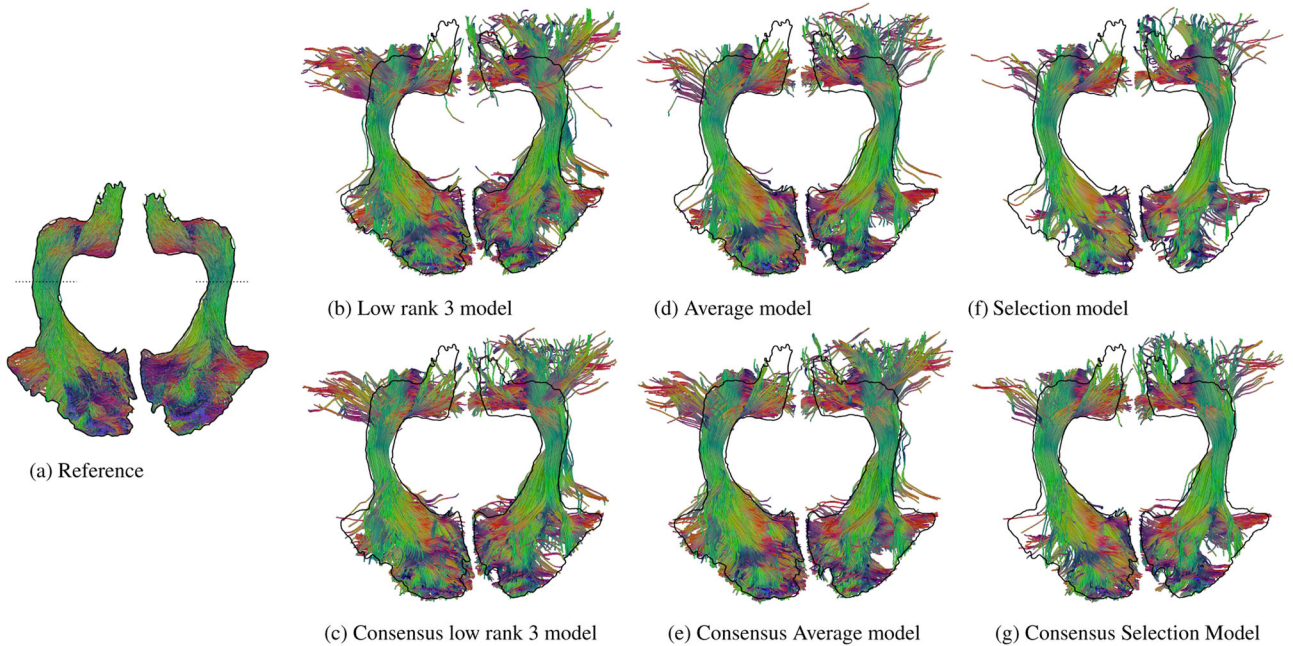
(a) Reference

(b) Low rank 3 model

(c) Consensus low rank 3 model

(d) Average model

(e) Consensus Average model

(f) Selection model

(g) Consensus Selection Model

Figure 7: *Reconstruction of the optic radiation (OR). Again, we overlaid the contour of the reference on each reconstruction with a black curve. The bootstrap consensus increases the overlap with the reference both for model averaging (d) and selection (f). For the rank-3 model (b), it mostly reduces the false positives, even though that effect is less visible due to the streamline filtering during post-processing.*

the consensus average model achieves a higher median Dice score than model averaging by itself, that the consensus selection model always beats model selection alone, and that the consensus rank-3 model has a higher median Dice in four out of seven tracts. According to Table 1, these improvements are significant for model averaging and selection, but not for the rank-3 model. We believe that this is due to the fact that the main advantage there is to reduce false positive streamlines, most of which are, however, successfully filtered out before computing the Dice.

Comparing the consensus models with each other, Figure 8 suggests that the consensus average and selection models are quite similar, and both are preferable compared to the consensus rank-3 model. This is confirmed by the significance tests. In terms of practical utility, it is relevant to put these results in context with the computational effort of the different approaches.

7.4. Computational effort

All experiments were computed on an Intel i9 with 3.3 GHz and 64 GB RAM. The following durations are denoted in (h:)min:s.

The computation of a single bootstrap data sample took 1:10 on a single core, multi-threaded fourth-order fODF tensor estimation took 4:33, the computation of the multi-vector fields took 0:40 with the rank-3 model, 1:10 with model selection, 1:30 with model averaging. Therefore, the calculation of 100 bootstraps took 8:08:00, using all CPU cores to create the bootstrap data as well as the models. This also includes the computation of the consensus model, which on its own took 11:00.

Table 2: *Comparison of the average time needed to reconstruct 1000 valid streamlines of the right CST in Figure 5. Differences mostly result from the number of successfully filtered false positive streamlines, which do not count towards the valid ones..*

	Time per 1000 valid streamlines in seconds
Average model	27.51
Consensus average model	28.29
Selection model	23.70
Consensus average model	24.59
Rank-3 model	43.27
Consensus rank-3 model	35.90

Typical times for the tracking itself are reported in Table 2. Times are averages for reconstructing 1000 valid streamlines in the CST. Differences mostly reflect the time spent on tracking false positive streamlines that are discarded by our filters, since those do not contribute towards the valid ones. Here, the fact that the rank-3 baseline generates considerably more false positives becomes even more apparent than in our previous qualitative and quantitative results. Tracking with model selection is fastest, but it is clear from the results above that this comes at the expense of the least complete reconstructions.

8. Conclusion

In this work, we presented and evaluated two approaches to reduce different types of uncertainty in diffusion MRI tractography. The
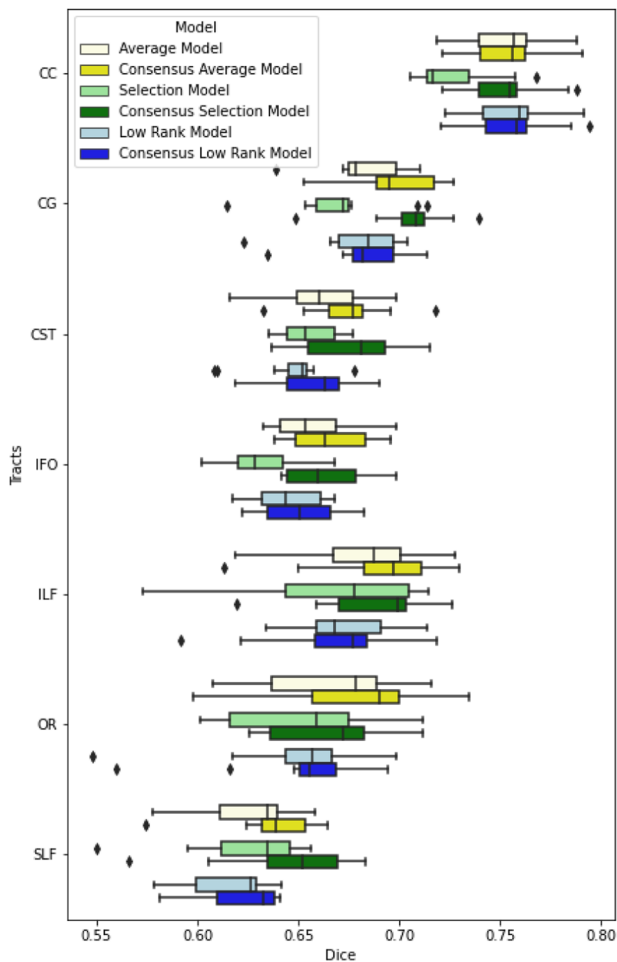
Figure 8: *Boxplots of the Dice scores of all subjects for all models. In almost all cases, model averaging yields a higher median Dice than model selection, and the bootstrap consensus improves upon the results of the corresponding base model.*

first one, model averaging, relies on Bayesian model comparison, and reduces the model uncertainty in crossing fibre tractography that results from having to select the most suitable number of fibres in each tracking step. The second one, bootstrap consensus, primarily aims to reduce data uncertainty, but we observe that it also reduces model uncertainty due to an interdependency of both. In either case, we fuse information from multiple fibre estimates to obtain a more reliable basis for fibre tractography.

Our experimental results demonstrate that each approach by itself makes dMRI tractography more accurate, as confirmed both by qualitative results and a formal statistical analysis. Even though combining both methods yields an additional benefit, it is worth keeping in mind that the additional computational expense from model averaging is marginal (in our experiments, it added 20 s to the pre-processing), while the computational effort for bootstrapping is substantial (in our experiments, more than 8 h per dataset).

We expect that, in many use cases, this will make model averaging the more pragmatic choice. Despite this, bootstrapping also pro-

vided further insights, most importantly, on the interaction between data and model uncertainty, and the effect of model averaging on the uncertainty in fibre direction estimates.

Acknowledgements

References

[AHL*01] ALEXANDER A. L., HASAN K. M., LAZAR M., TSURUDA J. S., PARKER D. L.: Analysis of partial volume effects in diffusion-tensor MRI. *Magnetic Resonance in Medicine 45* (2001), 770–780. http://doi.org/10.1002/mrm.1105

[ALGS17] ANKELE M., LIM L.-H., GROESCHEL S., SCHULTZ T.: Versatile, robust, and efficient tractography with constrained higher-order tensor fODFs. *International Journal of Computer Assisted Radiology and Surgery 12*, 8 (2017), 1257–1270. http://doi.org/10.1007/s11548-017-1593-6

[BBJ*07] BEHRENS T., BERG H. J., JBABDI S., RUSHWORTH M., WOOLRICH M.: Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? *NeuroImage 34*, 1 (2007), 144–155. http://doi.org/10.1016/j.neuroimage.2006.09.018

[BKN04] BRETTHORST G. L., KROENKE C. D., NEIL J. J.: Characterizing water diffusion in fixed baboon brain. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*. R. Fischer, R. Preuss and U. von TOUSSAINT (Eds.). American Inst. of Physics, Melville (2004), pp. 3–15. http://doi.org/10.1063/1.1835192

[BML94] BASSER P., MATTIELLO J., LEBIHAN D.: Estimation of the effective self-diffusion tensor from the NMR spin echo. *Journal of Magnetic Resonance, Series B 103*, 3 (1994), 247–254. http://doi.org/10.1006/jmrb.1994.1037

[BPtHV13] BRECHEISEN R., PLATEL B., ter Haar ROMENIJ B. M., VILANOVA A.: Illustrative uncertainty visualization of DTI fiber pathways. *The Visual Computer 29*, 4 (2013), 297–309. http://doi.org/10.1007/s00371-012-0733-9

[BVPtH09] BRECHEISEN R., VILANOVA A., PLATEL B., ter Haar ROMENIJ B. M.: Parameter sensitivity visualization for DTI fiber tracking. *IEEE Trans. on Visualization and Computer Graphics 15*, 6 (2009), 1441–1448. http://doi.org/10.1109/TVCG.2009.170

[CLH06] CHUNG S., LU Y., HENRY R. G.: Comparison of bootstrap approaches for estimation of uncertainties of DTI

parameters. *NeuroImage 33*, 2 (2006), 531–541. http://doi.org/10.1016/j.neuroimage.2006.07.001

[FÖK*07] Freidlin R. Z., Özarslan E., Komlosh M. E., Chang L.-C., Koay C. G., Jones D. K., Basser P. J.: Parsimonious model selection for tissue segmentation and classification applications: A study using simulated and experimental DTI data. *IEEE Transactions on Medical Imaging 26*, 11 (2007), 1576–1584. http://doi.org/10.1109/TMI.2007.907294

[Fri37] Friedman M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association 32*, 200 (1937), 675–701. http://doi.org/10.1080/01621459.1937.10503522

[GRNM16] Graumann A., Richter M., Nimsky C., Merhof D.: Exploring crossing fibers of the brain's white matter using directional regions of interest. In *Visualization in Medicine and Life Sciences III*. Springer, Cham (2016), pp. 179–194. http://doi.org/10.1007/978-3-319-24523-2_8

[GSWS21] Gillmann C., Saur D., Wischgoll T., Scheuermann G.: Uncertainty-aware visualization in medical imaging—a survey. *Computer Graphics Forum 40*, 3 (2021), 665–689. http://doi.org/10.1111/cgf.14333

[GvdVS21] Grün J., van der Voort G., Schultz T.: Reducing model uncertainty in crossing fiber tractography. In *Proceedings of the EG Workshop on Visual Computing for Biology and Medicine* (VCBM), pp. 55–64. http://doi.org/10.2312/vcbm.20211345

[HS05] Hlawitschka M., Scheuermann G.: HOT-lines: Tracking lines in higher order tensor fields. In *Proceedings of the IEEE Visualization* (2005), C. Silva, E. Gröller and H. Rushmeier (Eds.), pp. 27–34. http://doi.org/10.1109/VISUAL.2005.1532773

[JBBS02] Jenkinson M., Bannister P., Brady M., Smith S.: Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage 17*, 2 (2002), 825–841. http://doi.org/10.1006/nimg.2002.1132

[JDML19] Jeurissen B., Descoteaux M., Mori S., Leemans A.: Diffusion MRI fiber tractography of the brain. *NMR in Biomedicine 32*, 4 (2019), e3785. http://doi.org/10.1002/nbm.3785

[JLJ*11] Jeurissen B., Leemans A., Jones D. K., Tournier J.-D., Sijbers J.: Probabilistic fiber tracking using the residual bootstrap with constrained spherical deconvolution. *Human Brain Mapping 32* (2011), 461–479. http://doi.org/10.1002/hbm.21032

[JLT*12] Jeurissen B., Leemans A., Tournier J.-D., Jones D. K., Sijbers J.: Investigating the prevalence of complex fiber configurations in white matter tissue with diffusion magnetic resonance imaging. *Human Brain Mapping 34*, 11 (2012), 2747–2766. http://doi.org/10.1002/hbm.22099

[JM99] Jupp P. E., Mardia K. V.: *Directional Statistics*. Wiley, Chichester, West Sussex, England, 1999.

[Jon08] Jones D. K.: Tractography gone wild: Probabilistic fibre tracking using the wild bootstrap with diffusion tensor MRI. *IEEE Transactions on Medical Imaging 27*, 9 (2008), 1268–1274. http://doi.org/10.1109/TMI.2008.922191

[JTD*14] Jeurissen B., Tournier J.-D., Dhollander T., Connelly A., Sijbers J.: Multi-tissue constrained spherical deconvolution for improved analysis of multi-shell diffusion {MRI} data. *NeuroImage 103* (2014), 411–426. http://doi.org/10.1016/j.neuroimage.2014.07.061

[JTE*05] Jones D. K., Travis A. R., Eden G., Pierpaoli C., Basser P. J.: PASTA: pointwise assessment of streamline tractography attributes. *Magnetic Resonance in Medicine 53*, 6 (2005), 1462–1467. http://doi.org/10.1002/mrm.20484

[KK08] Konishi S., Kitagawa G.: *Information Criteria and Statistical Modeling*. Springer Science & Business Media, New York, 2008. http://doi.org/10.1007/978-0-387-71887-3

[KKA07] Kaden E., Knösche T. R., Anwander A.: Parametric spherical deconvolution: Inferring anatomical connectivity using diffusion MR imaging. *NeuroImage 37* (2007), 474–488. http://doi.org/10.1016/j.neuroimage.2007.05.012

[Kum37] Kummer E. E.: De integralibus quibusdam definitis et seriebus infinitis. *Journal für die reine und angewandte Mathematik 17* (1837), 228–242. http://doi.org/10.1515/crll.1837.17.228

[Kum80] Kumaraswamy P.: A generalized probability density function for double-bounded random processes. *Journal of Hydrology 46*, 1-2 (1980), 79–88. http://doi.org/10.1016/0022-1694(80)90036-0

[LBL*86] Le Bihan D., Breton E., Lallemand D., Grenier P., Cabanis E., Laval-Jeantet M.: MR imaging of intravoxel incoherent motions: Application to diffusion and perfusion in neurologic disorders. *Radiology 161*, 2 (1986), 401–407. http://doi.org/10.1148/radiology.161.2.3763909

[MCCvZ99] Mori S., Crain B. J., Chacko V. P., van Zijl P. C. M.: Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging. *Annals of Neurology 45*, 2 (1999), 265–269. http://doi.org/10.1002/1531-8249(199902)45:2<265::aid-ana21>3.0.co;2-3

[MNH*17] Maier-Hein K. H., Neher P. F., Houde J.-C., Côté M.-A., Garyfallidis E., Zhong J., Chamberland M., Yeh F.-C., Lin Y.-C., Ji Q., Reddick W. E., Glass J. O., Chen D. Q., Feng Y., Gao C., Wu Y., Ma J., He R., Li Q., Westin C.-F., Deslauriers-Gauthier S., González J. O. O., Paquette M., St-Jean S., Girard G., Rheault F., Sidhu J., Tax C. M. W., Guo F., Mesri H. Y., Dávid S., Froeling M., Heemskerk A. M., Leemans A., Boré A., Pinsard B., Bedetti C., Desrosiers M., Brambati S., Doyon J., Sarica A., Vasta R., Cerasa A., Quattrone A., Yeatman J., Khan A. R., Hodges W.,

ALEXANDER S., ROMASCANO D., BARAKOVIC M., AURÍA A., ESTEBAN O., LEMKADDEM A., THIRAN J.-P., CETINGUL H. E.,

ODRY B. L., MAILHE B., NADAR M. S., PIZZAGALLI F., PRASAD G., VILLALON-REINA J. E., GALVIS J., THOMPSON P. M., REQUEJO F. D. S., LAGUNA P. L., LACERDA L. M., BARRETT R., DELL'ACQUA F., CATANI M., PETIT L., CARUYER E., DADUCCI A., DYRBY T. B., HOLLAND-LETZ T., HILGETAG C. C., STIELTJES B., DESCOTEAUX M.: The challenge of mapping the human connectome based on diffusion tractography. *Nature Communications 8*, 1 (2017). http://doi.org/10.1038/s41467-017-01285-x

[NDH*15] NEHER P. F., DESCOTEAUX M., HOUDE J.-C., STIELTJES B., MAIER-HEIN K. H.: Strengths and weaknesses of state of the art fiber tractography pipelines. A comprehensive in-vivo and phantom evaluation study using tractometer. *Medical Image Analysis 26*, 1 (2015), 287–305. http://doi.org/10.1016/j.media.2015.10.011

[Nem63] NEMENYI P. B.: *Distribution-Free Multiple Comparison*. PhD thesis, 1963.

[PSS*12] PANAGIOTAKI E., SCHNEIDER T., SIOW B., HALL M. G., LYTHGOE M. F., ALEXANDER D. C.: Compartment models of the diffusion MR signal in brain white matter: A taxonomy and comparison. *NeuroImage 59* (2012), 2241–54. http://doi.org/10.1016/j.neuroimage.2011.09.081

[Sch78] SCHWARZ G.: Estimating the dimension of a model. *The Annals of Statistics 6*, 2 (Mar. 1978), 461–464. http://doi.org/10.1214/aos/1176344136

[Sch12] SCHULTZ T.: Learning a reliable estimate of the number of fiber directions in diffusion MRI. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI) Part III, LNCS*. N. Ayache, et al. (Ed.). (2012), vol. *7512*, pp. 493–500. http://doi.org/10.1007/978-3-642-33454-2_61

[SDM*19] SCHILLING K. G., DADUCCI A., MAIER-HEIN K., POUPON C., HOUDE J.-C., NATH V., ANDERSON A. W., LANDMAN B. A., DESCOTEAUX M.: Challenges in diffusion MRI tractography lessons learned from international benchmark competitions. *Magnetic Resonance Imaging 57* (2019), 194–209. http://doi.org/10.1016/j.mri.2018.11.014

[SHV21] SIDDIQUI F., HÖLLT T., VILANOVA A.: A progressive approach for uncertainty visualization in diffusion tensor imaging. *Computer Graphics Forum 40*, 3 (2021), 411–422. http://doi.org/10.1111/cgf.14317

[SJX*13] SOTIROPOULOS S. N., JBABDI S., XU J., ANDERSSON J. L., MOELLER S., AUERBACH E. J., GLASSER M. F., HERNANDEZ M., SAPIRO G., JENKINSON M., FEINBERG D. A., YACOUB E., LENGLET C., ESSEN D. C. V., UGURBIL K., BEHRENS T. E.: Advances in diffusion MRI acquisition and processing in the Human Connectome Project. *NeuroImage 80* (2013), 125–143. http://doi.org/10.1016/j.neuroimage.2013.05.057

[SS08] SCHULTZ T., SEIDEL H.-P.: Estimating crossing fibers: A tensor decomposition approach. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1635–1642. http://doi.org/10.1109/TVCG.2008.128

[SSSS13] SCHULTZ T., SCHLAFFKE L., SCHÖLKOPF B., SCHMIDT-WILCKE T.: HiFiVE: A Hilbert space embedding of fiber variability estimates for uncertainty modeling and visualization. *Computer Graphics Forum 32*, 3 (2013), 121–130. http://doi.org/10.1111/cgf.12099

[STR*22] SCHILLING K. G., TAX C. M. W., RHEAULT F., LANDMAN B. A., ANDERSON A. W., DESCOTEAUX M., PETIT L.: Prevalence of white matter pathways coming into a single white matter voxel orientation: The bottleneck issue in tractography. *Human Brain Mapping 43*, 4 (2022), 1196–1213. http://doi.org/10.1002/hbm.25697

[SV19] SCHULTZ T., VILANOVA A.: Diffusion MRI visualization. *NMR in Biomedicine 32*, 4 (2019), e3902. http://doi.org/10.1002/nbm.3902

[SVBK14] SCHULTZ T., VILANOVA A., BRECHEISEN R., KINDLMANN G.: Fuzzy fibers: Uncertainty in dmri tractography. In *Scientific Visualization: Uncertainty, Multifield, Biomedical, and Scalable Visualization*. C. Hansen, M. Chen, C. Johnson, A. Kaufman and H. Hagen (Eds.). Springer, London (2014), pp. 79–92. http://doi.org/10.1007/978-1-4471-6497-5_8

[SWK10] SCHULTZ T., WESTIN C.-F., KINDLMANN G.: Multi-diffusion-tensor fitting via spherical deconvolution: A unifying framework. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), LNCS.* T. Jiang, N. Navab, J. P. W. Pluim and M. A. Viergever (Eds.). Springer, (2010), vol. *6361*, pp. 673–680. http://doi.org/10.1007/978-3-642-15705-9_82

[TCC07] TOURNIER J.-D., CALAMANTE F., CONNELLY A.: Robust determination of the fibre orientation distribution in diffusion MRI: Non-negativity constrained super-resolved spherical deconvolution. *NeuroImage 35*, 4 (2007), 1459–1472. http://doi.org/10.1016/j.neuroimage.2007.02.016

[TCWP16] TAKEMURA H., CAIAFA C. F., WANDELL B. A., PESTILLI F.: Ensemble tractography. *PLoS Computational Biology 12*, 2 (2016), e1004692. http://doi.org/10.1371/journal.pcbi.1004692

[TSH*18] TOBISCH A., STIRNBERG R., HARMS R. L., SCHULTZ T., ROEBROECK A., BRETELER M. M., STÖCKER T.: Compressed sensing diffusion spectrum imaging for accelerated diffusion microstructure MRI in long-term population imaging. *Frontiers in Neuroscience 12* (2018), 650. http://doi.org/10.3389/fnins.2018.00650

[VSB*13] VAN ESSEN D. C., SMITH S. M., BARCH D. M., BEHRENS T. E., YACOUB E., UGURBIL K.: The Wu-Minn Human connectome project: An overview. *NeuroImage 80* (2013), 62–79. Mapping the Connectome. http://doi.org/10.1016/j.neuroimage.2013.05.041

[VVL13] VOS S. B., VIERGEVER M. A., LEEMANS A.: Multi-fiber tractography visualizations for diffusion MRI data. *PLOS*

ONE 8, 11 (2013), e81453. http://doi.org/10.1371/journal.pone. 0081453

[WCP*07] Wakana S., Caprihan A., Panzenboeck M. M., Fallon J. H., Perry M., Gollub R. L., Hua K., Zhang J., Jiang H., Dubey P., Blitz A., van Zijl P., Mori S.: Reproducibility of quantitative tractography methods applied to cerebral white matter. *NeuroImage 36* (2007), 630–644. http://doi.org/10.1016/j.neuroimage.2007.02.049

[WKL99] Weinstein D., Kindlmann G., Lundberg E.: Tensorlines: Advection-diffusion based propagation through diffusion tensor fields. In *Proceedings of the IEEE Visualization* (1999), pp. 249–253. http://doi.org/10.1109/VISUAL.1999.809894

[WNM18] Wasserthal J., Neher P., Maier-Hein K. H.: TractSeg—fast and accurate white matter tract segmentation. *NeuroImage 183* (2018), 239–253. http://doi.org/10.1016/j.neuroimage.2018.07.070

[WSSS14] Wiens V., Schlaffke L., Schmidt-Wilcke T., Schultz T.: Visualizing uncertainty in HARDI tractography using superquadric streamtubes. In *Proceedings of the EG Conference on Visualization (EuroVis) Short Papers* (2014), pp. 37–41. http://doi.org/10.2312/eurovisshort. 20141154

[YYPC21] Yang J. Y.-M., Yeh C.-H., Poupon C., Calamante F.: Diffusion MRI tractography for neurosurgery: The basics, current state, technical reliability and challenges. *Physics in Medicine & Biology* (2021). http://doi.org/10.1088/1361-6560/ac0d90

[ZDL03] Zhang S., Demiralp C., Laidlaw D. H.: Visualizing diffusion tensor MR images using streamtubes and streamsurfaces. *IEEE Transactions on Visualization and Computer Graphics 9*, 4 (2003), 454–462. http://doi.org/10.1109/TVCG.2003. 1260740

[ZSGA12] Zhang H., Schneider T., Gandini Wheeler-Kingshott C., Alexander D.: NODDI: Practical in vivo neurite orientation dispersion and density imaging of the human brain. *NeuroImage 61* (Mar. 2012), 1000–1016. http://doi.org/10.1016/j.neuroimage.2012.03.072