International Journal of Computer Vision (2021) 129:656–680
https://doi.org/10.1007/s11263-020-01384-1

Entrack: Probabilistic Spherical Regression with Entropy Regularization for Fiber Tractography

Viktor Wegmayr[1] · Joachim M. Buhmann[1]

Received: 8 January 2020 / Accepted: 7 September 2020 / Published online: 6 November 2020

Abstract
White matter tractography, based on diffusion-weighted magnetic resonance images, is currently the only available in vivo method to gather information on the structural brain connectivity. The low resolution of diffusion MRI data suggests to employ probabilistic methods for streamline reconstruction, i.e., for fiber crossings. We propose a general probabilistic model for spherical regression based on the Fisher-von-Mises distribution, which efficiently estimates maximum entropy posteriors of local streamline directions with machine learning methods. The optimal precision of posteriors for streamlines is determined by an information-theoretic technique, the expected log-posterior agreement concept. It relies on the requirement that the posterior distributions of streamlines, inferred on retest measurements of the same subject, should yield stable results within the precision determined by the noise level of the data source.

Keywords Diffusion MRI · Brain · Tractography · Machine Learning · Maximum-entropy inference · Algorithm validation

1 Introduction

1.1 Cerebral White Matter and Diffusion MRI

The structural connectivity between different cortical brain regions is established by white matter, that is composed of myelinated axons to distribute action potentials as messages between communicating neurons. The functional importance of connectivity for cognition has been undisputedly recognized by neuroscience research (Bargmann and Marder 2013; Filley and Fields 2016).

The advent of diffusion-weighted magnetic resonance imaging (DWI) (Chilla et al. 2015; Soares et al. 2013) has empowered neuroscientists and neurologists to monitor changes in the structural connectivity with potential relevance for diagnosis, prognosis and therapy of neurodegenerative diseases (Oishi et al. 2011). DWI is currently the only non-invasive, and non-radiative imaging modality, which enables neurologists to investigate the connective micro-architecture of the white matter in a minimally invasive way. Its image contrast encodes the anisotropic diffusion of water in tissue (Beaulieu 2002), making it a highly informative probe of the fibrous white matter (Bihan and Iima 2015). The axon bundles of white matter locally exhibit clear preferential directions, as shown in Fig. 1a.

However, fiber tracking algorithms are required to reconstruct consistent long-range tissue connectivity from local, voxel-centric [1] DWI measurements.

1.2 Tractography

Long-range connections in the white matter are commonly referred to as streamlines, fibers or tracts. Algorithmic methods to computationally reconstruct such streamlines from DWI are known as *tractography* (Jeurissen et al. 2019; Nimsky et al. 2016). Schematically, tractography infers structural connections between voxels to answer questions like “Does there exist a structural connection between regions A and B?”. We show a prototypical tractography result, also referred to as tractogram, in Fig. 1b.

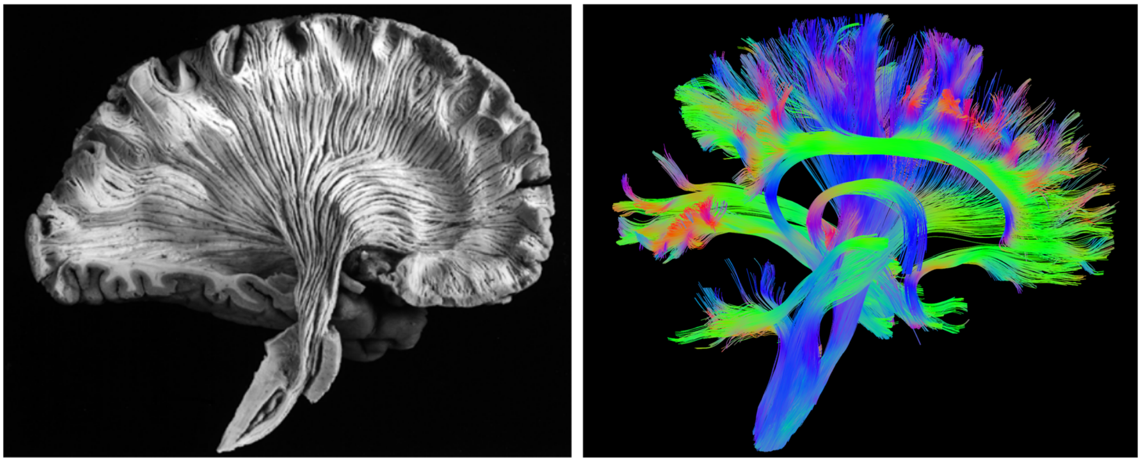
Tractography is clinically applied to gather health information for a number of neurological conditions, especially for preoperative planning of neurosurgery, and for research on stroke and dementia impact on brain function (Yamada

Communicated by Simone Frintrop.

✉ Viktor Wegmayr
vwegmayr@inf.ethz.ch

[1] Department of Computer Science, ETH Zurich, Zurich, Switzerland

[1] A voxel is the 3-D analogue of a pixel.



(a) The cerebral white matter (Williams et al., 1997)

(b) The ISMRM15 tractogram (Maier-Hein et al., 2017)

Fig. 1 Juxtaposition of (**a**) an anatomical sample of the cerebral white matter illustrating the fibrous white matter bundles, and (**b**) a reconstruction of white matter pathways by tractography from a DWI measurement. The RGB colors correspond to the projection of the local xyz fiber orientation (R = x, G = y, B = z)

et al. 2009). The data processing pipeline for tractography is composed of three stages, (i) DWI measurements of apparent diffusion, (ii) estimation of a local diffusion model per voxel, and (iii) inference of streamlines following the local diffusion model—as illustrated in Fig. 2.

Tractography poses a major challenge due to its ambiguity mostly caused by partial volume effects, since axon diameters rarely exceed few micrometers, while the DWI resolution is limited to the scale of millimeters. This lack of resolution severely complicates inference of streamlines since the superposition of diffusion information renders it difficult to disambiguate locations where fibers cross, touch, or fan apart (Jbabdi and Johansen-Berg 2011). These complex fiber configurations have been observed to be highly prevalent in the white matter of human brains (Jeurissen et al. 2013) which further impedes data analysis of white matter especially in neurology.

The majority of tractography algorithms reconstructs streamlines in a *local* manner, i.e. they proceed iteratively from a given seed point, and greedily determine the direction of the next step based only on the local diffusion features, and information from previous direction estimates. Tractography algorithms can be distinguished into *deterministic* and *probabilistic* methods depending on how they estimate the direction of the next step. While deterministic methods compute a point-estimate of the next direction in line with the most-likely direction, probabilistic methods infer a distribution over possible directions. Sampling from this distribution supports following multiple traces along different directions at every step. In particular, probabilistic methods are able to express the uncertainty of their predictions, which also renders them more robust in the presence of noise.

1.3 Contributions

Probabilistic Regression for Tractography

Recently, algorithms based on supervised machine learning (ML) have successfully extended the toolbox of local tractography methods. Even though these ML algorithms depend on the quality of the training streamlines, it has been shown by several works that ML models trained on fibers produced by another, unsupervised algorithm (teacher) can generalize very well to new DWI data, even improving over the teacher performance Wegmayr (2018), Neher et al. (2017), Benou and Riklin-Raviv (2019).

In this work, we present a probabilistic regression approach that avoids the conceptual problems of classification-based models such as direction discretization, and the lack of a closeness notion for directions. To define a proper regression model for d-dimensional vectors on the unit-sphere $\mathbf{s} \in \mathbb{S}_{d-1}$ in a probabilistic framework, we propose a learnable posterior based on the Fisher-von-Mises (FvM) distribution (Mardia and Jupp 2000). Conditioned on the feature vector $x \in \mathcal{X} \subseteq \mathbb{R}^p$, the posterior p^{FvM} is given by

$$p^{\text{FvM}}(\mathbf{s} \mid x) := C\big(\kappa(x)\big) \exp\left(\kappa(x)\langle \mathbf{s}, \boldsymbol{\mu}(x)\rangle\right), \tag{1}$$

where $\langle \mathbf{s}, \boldsymbol{\mu}(x)\rangle$ denotes the scalar product between the random output direction $\mathbf{s}$ and the mean direction $\boldsymbol{\mu}(x)$. $C\big(\kappa(x)\big)$ abbreviates the normalization constant of p^{FvM}. Besides the mean direction, the scalar concentration $\kappa(x)$ is also a function of the input x, which accounts for input-dependent noise, heteroscedastic noise. In the context of tractography, x represents the local diffusion features, whereas $\mathbf{s}$ should be understood as the latent local direction of the fiber bundle. The functions $\boldsymbol{\mu}(x)$, $\kappa(x)$ are represented by neural networks, and their parameters are learned by minimizing the negative log-likelihood of observed reference streamlines.

Parameter inference for such a probabilistic approach amounts to a model selection problem and has to be carefully regularized to avoid an unbounded increase of the concentration $\kappa(x)$ during model training, which would effectively reduce the model to its deterministic variant. While this effect is a common problem in many applications, both in tractography (Benou and Riklin-Raviv 2019), and outside of it (Sensoy et al. 2018; Kumar and Tsvetkov 2018), solutions are typically based on heuristics with ad-hoc penalty terms.

Instead, we derive a sound regularization scheme based on the information-theoretically optimal maximum entropy principle (Jaynes 1957). The resulting Gibbs distribution controls uncertainty by a *precision* parameter β that allows us to adapt the posterior uncertainty to the noise in the data. Even though the presented entropy-regularized FvM model applies to general spherical regression tasks with the need for uncertainty estimation, our focus is on applications to tractography, hence we refer to it as *Entrack*. Other pattern analysis applications of spherical or directional regression can be found in the prediction of word embedding vectors Kumar and Tsvetkov (2018), or object pose estimation from images Prokudin et al. (2018).

The Optimal Precision

While the precision β mentioned above enables us to regularize the global FvM posterior for streamline directions in all voxels, it is a priori not clear how to determine its optimal value. In particular, we are going to argue that common cross-validation techniques are not effective, because they measure the generalization error with respect to the posterior mean direction, whereas the precision only controls the width of the posterior distribution. Indeed, the smallest generalization error is achieved by the posterior distribution with infinite precision, which yields the well-known empirical risk minimizer as an estimator. Infinite precision implies minimal entropy, which means a sub-optimal robustness of the posterior distribution in the presence of noise. We also show that even more involved evaluation schemes, such as the Tractometer (Maier-Hein et al. 2017), are not a viable method to determine the optimal posterior precision, because their evaluation is still based on a single measurement instance, which is insufficient to estimate the influence of data fluctuations on the resulting tractograms.

Our model selection criterion requires at least two measurements to estimate the optimal posterior width relative to the data noise. Formally, this two instance scenario is described by the information-theoretic framework of *expected log-posterior agreement* (PA) (Buhmann 2010), which determines the optimal value of the precision parameter by maximizing the relative overlap between the posterior distributions on repeated measurements. We discuss its implementation in the context of tractography, and perform experiments on repeated DWI scans of the same subject to estimate the optimal precision.

Extension of Previous Work

This work is an extended version of our previous conference paper Wegmayr et al. (2019). We have extended, and reorganized the theoretical contributions about probabilistic directional regression, and entropy regularization (Sect. 3), including a novel annealing algorithm (Algorithm 1). Moreover, we propose the method of posterior agreement to determine the optimal precision (Sect. 3.5), and describe how to implement it for tractography (Sect. 4.3). The experiments are extended considerably, too, by investigating case studies of posterior estimation of local fiber direction (Sect. 5.3), and its relationship with fractional anisotropy (Sect. 6.2). Additionally, the evaluation on the Tractometer benchmark has been extended to include more competing methods (Sect. 6.3). Lastly, the experimental validation of posterior agreement on retest data also represents a new contribution (Sect. 6.5).

Overview

After summarizing related work in Sect. 2, we describe a probabilistic model for spherical regression, based on the FvM distribution, in Sect. 3.2.

To address the wide-spread problem of probabilistic overfitting, we introduce a regularized Gibbs free energy objective in Sect. 3.3, which controls the entropy of the posterior distribution via a precision parameter. We discuss its implications for model training, including an automatic annealing algorithm for parameter optimization in Sect. 3.4. Concluding the general description of methods, we present the expected log-posterior agreement for the FvM posterior in Sect. 3.5, which allows us to calibrate the precision parameter according to the noise level in the data.

Section 4 presents the described methods in the context of streamline tractography, which is also indicated by the term Entrack. In particular, we define the models for DWI data, and their interpretation in terms of tractograms. Based on a factorization of tractograms into independent, piecewise linear streamline segments, we use the entropy-regularized regression model in Sect. 4.1 to learn the relationship between local fiber direction, and the diffusion data.

Using a step-wise tracking algorithm, we show how to use the local Entrack posterior to obtain long-range streamlines

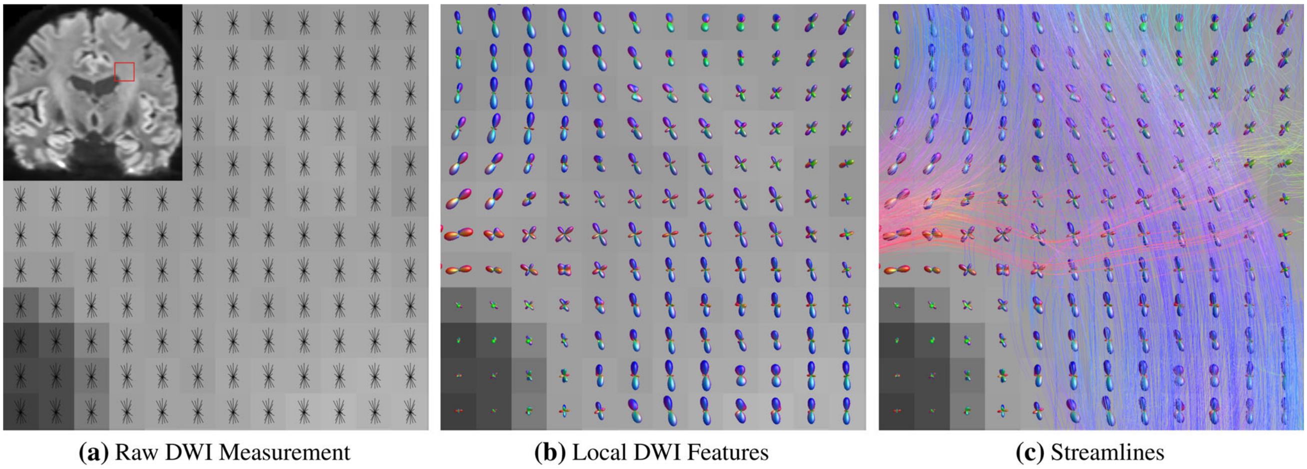

(a) Raw DWI Measurement **(b)** Local DWI Features **(c)** Streamlines

Fig. 2 Illustration of the different steps of a tractography pipeline. (a) Schematic DWI-measurement. (b) Fitting a per-voxel model to the DWI measurements, to obtain the diffusion features, here in terms of fiber orientation distributions (FOD). (c) Reconstruction of long-range streamlines by tractography.

in Sect. 4.2. The calculation of the PA for tractograms from repeated DWI measurements is described in Sect. 4.3.

After providing details about DWI data preprocessing (Sect. 5.1), and the neural network implementation of the Entrack posterior (Sect. 5.2), we perform case studies on prototypical diffusion profiles in Sect. 5.3, to investigate the patterns learned by the model.

We turn to whole-brain tractography in Sect. 6, including experiments on synthetic phantom data (Sects. 6.1, 6.3 and 6.4), and on real data (Sect. 6.5). Finally, we conclude in Sect. 7.

2 Related Work

2.1 Local Tractography Algorithms

Local tractography algorithms, as opposed to global tractography (Reisert et al. 2011), reconstruct streamlines independently, in an iterative way, based on the DWI signal in a close spatial neighborhood. This design strategy has served as the core idea of many tractography algorithms since streamline tractography originally used Runge-Kutta methods to integrate the streamline progression (see Basser et al. (2000)).

Later, the works of Behrens et al. (2003), Friman et al. (2006) introduced local, probabilistic tractography models based on mathematical models for the posterior distribution of the streamline direction. While these probabilistic methods have proven to be robust to noise, they are computationally expensive, because they need to re-estimate the high-dimensional integrals involved in the posterior for *every* voxel.

Very recently, a new generation of models based on supervised machine learning (ML) entered the scene, and they promise to solve deficiencies of traditional models (Poulin et al. 2019). (i) ML-methods solve the parameter estimation problem only *once* over a representative set of examples during their training phase. Afterwards during inference, the algorithm only requires arithmetic evaluation of the model function at each voxel, which is efficiently achieved.

(ii) Moreover, ML-methods are better suited to capture complex patterns between DWI data and fiber direction in a non-parametric way than traditional approaches, which are limited by the richness of parametric statistical models.

However, ML-methods rely on fiber tracking examples to yield supervision information which is not required for traditional (“unsupervised”) methods. To circumvent this issue, supervised approaches have been trained on the output of the previous state-of-the-art algorithms in traditional tractography; furthermore, well-curated training sets are becoming available in increasing numbers (Wasserthal et al. 2018; Essen et al. 2013).

Depending on the estimation technique for local streamline directions, ML models for tractography have been formulated either as regression problems or as classification tasks. Classification models (Neher et al. 2017; Benou and Riklin-Raviv 2019) are probabilistic in nature, but require categorical classes to approximate continuous directions. In contrast, regression models (Wegmayr 2018; Poulin et al. 2017) provide the more appropriate representation for continuous directions, but we are not aware of probabilistic regression models in the context of tractography.

2.2 Uncertainty Quantification

Uncertainty in statistical inference arises in two distinctly different flavors – epistemic and aleatoric uncertainty – as described by Kiureghian and Ditlevsen (2009) for general engineering.

Kendall and Gal (2017) discusses the estimation of epistemic and aleatoric uncertainty in the context of computer vision. The first one, epistemic uncertainty, refers to our uncertainty about the model parameters, and it decreases when more observed samples become available. The second one, aleatoric uncertainty, refers to input-dependent noise, which is inherent to the data distribution. As such, it is unaffected by the number of observed samples. Very recently, predictive models, which also incorporate estimation of aleatoric noise, have received increasing attention, e.g. for categorical classification (Sensoy et al. 2018).

Probabilistic regression has been addressed by Prokudin et al. (2018), who uses a mixture of 1-dimensional FvM distributions in the context of object-pose estimation, and by Kumar and Tsvetkov (2018) for sequence-to-sequence models for language generation. Similarly to this work, the latter proposes a probabilistic error function based on d-dimensional FvM distribution, however, their focus is rather on reducing model training time than on uncertainty quantification.

Lastly, we also mention the method of Hauberg et al. (2015) for shortest-path tractography, who uses probabilistic numerics to solve Gaussian-process ODEs.

2.3 Expected Log-Posterior Agreement

The framework of expected log-posterior agreement defines a model selection method for algorithms and it was originally derived from information-theoretic principles (Buhmann 2010). More precisely, as described by Buhmann (2013), it measures the trade-off between informativeness, and stability of a cost minimizing algorithm in terms of the overlap that its posterior distribution exerts between repeated measurements. An algorithm's posterior distribution is considered informative, if it narrows down the set of potential solutions for each measurement, and it is considered stable, if the posteriors obtained from repeated measurements agree with each other in spite of measurement noise.

The PA framework, sometimes also referred to as Approximation Set Coding, or Gibbs posterior agreement, has been applied in various settings such as singular value decomposition (Frank and Buhmann 2011), spectral clustering (Chehreghani et al. 2012), Gaussian process regression (Fischer et al. 2016), and combinatorial optimization problems (Buhmann et al. 2018).

The PA criterion has also been applied in the context of neuroscience, namely to determine the optimal number of clusters for cortex parcellation (Gorbach et al. 2018).

3 Entropy-Regularized Spherical Regression

3.1 The Fisher-von-Mises Distribution

The FvM distribution is a unimodal, directional distribution defined on the d-sphere $\mathbb{S}_{d-1}$. For random unit vectors $\mathbf{s} \in \mathbb{S}_{d-1}$, the FvM density is given by

$$p^{\mathrm{FvM}}(\mathbf{s} \mid \boldsymbol{\mu}, \kappa) := C(\kappa) \exp\left(\kappa \langle \mathbf{s}, \boldsymbol{\mu} \rangle\right) \tag{2}$$

with $\langle \mathbf{s}, \boldsymbol{\mu} \rangle = \sum_{i=1}^{d} s_i \mu_i$, and the normalizing constant

$$C(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)}, \tag{3}$$

where $I_n(.)$ denotes the modified Bessel function of the first kind. The FvM distribution is parameterized by the unit-length mean direction $\boldsymbol{\mu} \in \mathbb{S}_{d-1}$ and the scalar concentration $\kappa \in \mathbb{R}^+$. We illustrate the $d = 3$ dimensional FvM density for three different concentration parameters κ in Fig. 3. The norm of the first moment $W(\kappa)$, and the entropy $H(\kappa)$ of the FvM distribution are given by

$$\begin{aligned} W(\kappa) &:= \left\| \int_{\mathbb{S}_{d-1}} \mathbf{s}\, p^{\mathrm{FvM}}(\mathbf{s} \mid \boldsymbol{\mu}, \kappa) \mathrm{d}\mathbf{s} \right\|_2 \\ &= I_{d/2}(\kappa) \,/\, I_{d/2-1}(\kappa), \qquad (4a) \\ H(\kappa) &:= -\int_{\mathbb{S}_{d-1}} \log p^{\mathrm{FvM}}(\mathbf{s} \mid \boldsymbol{\mu}, \kappa) p^{\mathrm{FvM}}(\mathbf{s} \mid \boldsymbol{\mu}, \kappa) \mathrm{d}\mathbf{s} \\ &= -\log C(\kappa) - \kappa W(\kappa). \qquad (4b) \end{aligned}$$

We illustrate both functions for $d = 3$ in Fig. 4, and note that in contrast to the mean direction $\boldsymbol{\mu}$, the norm of the first moment, i.e. $W(\kappa)$, can be smaller than 1. Indeed it vanishes in the limit of very small concentration $\kappa \to 0$ when p^{FvM} approaches the uniform distribution proportional to the inverse surface of the d-sphere:

$$C(0) = \frac{\Gamma(d/2+1)}{d\pi^{d/2}}, \tag{5}$$

For very large concentration, the FvM distribution contracts at the mean direction:

$$p^{\mathrm{FvM}}(\mathbf{s} \mid \boldsymbol{\mu}, \kappa) \overset{\kappa \to \infty}{\longrightarrow} \delta(\mathbf{s} - \boldsymbol{\mu}), \tag{6}$$

where δ denotes the Dirac measure.

3.2 Probabilistic Regression with the FvM

In spherical regression, we want to estimate the regression function $\mathbf{y} : \mathcal{X} \to \mathbb{S}_{d-1},\ x \mapsto \mathbf{y}(x)$, which maps the feature

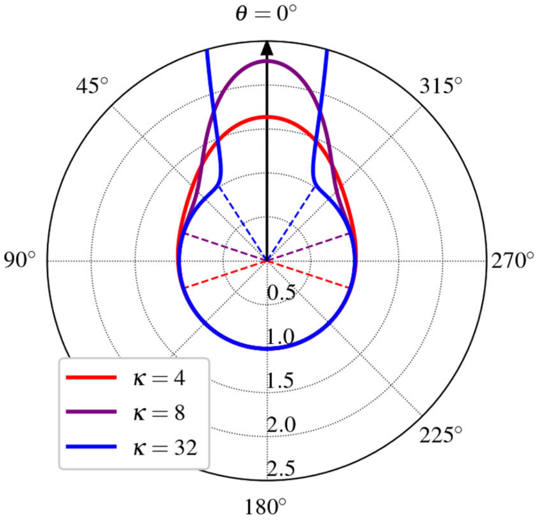


Fig. 3 Illustration of the FvM density distribution within the xz plane. The mean direction is along the z-axis (*arrow*, $\theta = 0°$). The radius of each polar plot is equal to $r(\theta) = 1 + p^{\mathrm{FvM}}(\mathbf{R}_\theta \mathbf{e}_z \mid \mathbf{e}_z, \kappa)$, where $\mathbf{R}_\theta$ is a rotation by θ degrees around the y-axis. The cones (*dashed*) correspond to the 99.5-percentile cones for the respective concentrations.

space $\mathcal{X} \subseteq \mathbb{R}^p$ to the d-sphere $\mathbb{S}_{d-1}$. As the feature vectors $x \in \mathcal{X}$ are drawn from a distribution $p(x)$, the observations $\mathbf{y}(x)$ are random variables. The estimated regression function is denoted as $\boldsymbol{\mu}$, and as the involved vectors have unit-length, the squared distance between a prediction $\boldsymbol{\mu}(x)$ and the corresponding observation $\mathbf{y}(x)$ effectively reduces to the negative cosine loss, when disregarding constant terms:

$$\ell\left(\boldsymbol{\mu}(x), \mathbf{y}(x)\right) = -\left\langle \boldsymbol{\mu}(x), \mathbf{y}(x) \right\rangle, \tag{7}$$

The loss in Eq. (7) is -1, if the prediction points into the same direction as the observation, and 1 if they are anti-parallel. To obtain the corresponding probabilistic regression model, we additionally introduce a function $\kappa : \mathcal{X} \to \mathbb{R}_+$, which acts as the concentration parameter of a predicted FvM distribution $p^{\mathrm{FvM}}(. \mid x) := p^{\mathrm{FvM}}(. \mid \boldsymbol{\mu}(x), \kappa(x))$. The loss of the functions $\boldsymbol{\mu}(x), \kappa(x)$ is the negative log-likelihood of the direction $\mathbf{y}(x)$ under the corresponding FvM distribution:

$$\begin{aligned} L\left(\mathbf{y}(x), p^{\mathrm{FvM}}(. \mid x)\right) &:= -\log p^{\mathrm{FvM}}\left(\mathbf{y}(x) \mid \boldsymbol{\mu}(x), \kappa(x)\right) \\ &= -\kappa(x)\left\langle \mathbf{y}(x), \boldsymbol{\mu}(x)\right\rangle - \log C\left(\kappa(x)\right). \end{aligned} \tag{8}$$

The functions $\boldsymbol{\mu}, \kappa$ are typically parametrized, e.g. in terms of neural networks. Given inputs $x \in \mathbb{R}^p$, the simplest such example would be

$$\begin{aligned} \boldsymbol{\mu}_{\varphi_\mu}(x) &:= (W_\mu x + b_\mu)/\|W_\mu x + b_\mu\|_2 \\ \kappa_{\varphi_\kappa}(x) &:= |\langle w_\kappa, x\rangle + b_\kappa|, \end{aligned} \tag{9}$$

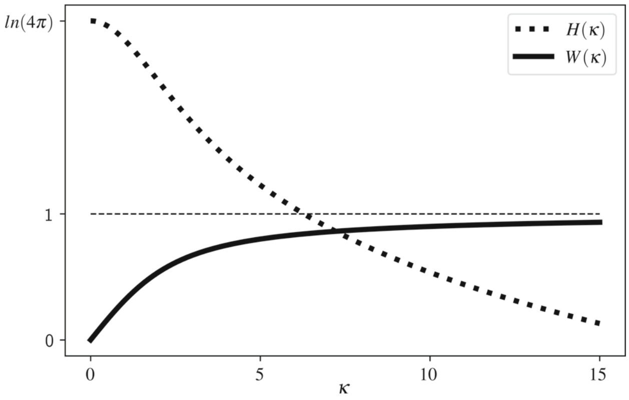


Fig. 4 Illustration of the entropy $H(\kappa)$ (*dotted*) and the norm of the first moment $W(\kappa)$ (*solid*) of a FvM distribution. Figure reproduced with slight modifications from Wegmayr et al. (2019)

with the weights and biases

$$\begin{aligned} \varphi &:= (\varphi_\mu, \varphi_\kappa) \\ &:= (W_\mu, b_\mu, w_\kappa, b_\kappa) \in \mathbb{R}^{d \times p} \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}, \end{aligned} \tag{10}$$

which define the parametrized FvM posterior

$$p_\varphi^{\mathrm{FvM}}(. \mid x) := p^{\mathrm{FvM}}\left(. \mid \boldsymbol{\mu}_{\varphi_\mu}(x), \kappa_{\varphi_\kappa}(x)\right). \tag{11}$$

In our experiments, the neural networks are more complex than Eq. (9) , but it effectively plays the same role. Given a training set $\{(x_i, \mathbf{y}_i)\}_{i=1\ldots n}$ $(\forall i : \mathbf{y}_i := \mathbf{y}(x_i))$, the parameters φ are estimated by minimizing the empirical risk function

$$\hat{\varphi} := \arg\min_{\varphi} \frac{1}{n} \sum_{i=1}^{n} L\left(\mathbf{y}_i, p_\varphi^{\mathrm{FvM}}(. \mid x_i)\right). \tag{12}$$

The risk function inherits the property of loss attenuation from the loss function in Eq. (8), which means that the loss caused by a large deviation $-\langle \mathbf{y}(x), \boldsymbol{\mu}(x)\rangle$ is reduced by a low certainty $\kappa(x)$. We illustrate the effect of loss attenuation for the FvM distribution ($d = 3$) in Fig. 5. The attenuation property ensures an increased robustness to outliers due to an adaptive sensitivity of the loss function. Furthermore, the concentration $\kappa(x)$ is a function of the input and this fact enables us to assess the certainty of the predicted direction for any sample x.

However, in practice, these benefits of a probabilistic formulation will be severely reduced by overfitting. When the model complexity is large, e.g. for neural networks, the posterior p_φ^{FvM} can perfectly minimize the training risk, in particular its concentration estimates will be biased towards large values, as we can see from the gradient of the risk with respect to the parameters φ_κ:

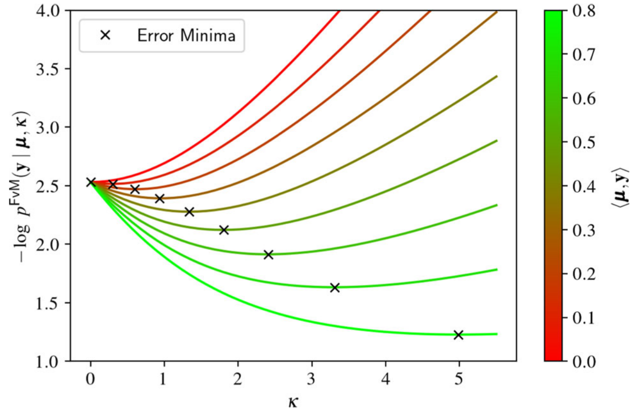


Fig. 5 Illustration of loss attenuation for the FvM distribution. Each line shows the negative log-likelihood as a function of κ for a fixed value of $\langle \mathbf{y}(x), \boldsymbol{\mu}(x) \rangle$. For high deviations between prediction $\boldsymbol{\mu}(x)$ and target $\mathbf{y}(x)$ (*red lines*), the minimum loss (*black crosses*) is realized at a low concentration κ, while small deviations (*green lines*) have their minima at higher concentrations. Figure reproduced with modifications from Wegmayr et al. (2019) (Color figure online).

$$\begin{aligned}
&\nabla_{\varphi_\kappa} \frac{1}{n} \sum_{i=1}^{n} L\left(\mathbf{y}_i, p_\varphi^{\mathrm{FvM}}(. \mid x_i)\right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{\partial L}{\partial \kappa} \left(\mathbf{y}_i, p_\varphi^{\mathrm{FvM}}(. \mid x_i)\right) \nabla_{\varphi_\kappa} \kappa_{\varphi_\kappa}(x_i) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\left\langle \mathbf{y}_i, \boldsymbol{\mu}_{\varphi_\mu}(x_i) \right\rangle - W\left(\kappa_{\varphi_\kappa}(x_i)\right)\right) \nabla_{\varphi_\kappa} \kappa_{\varphi_\kappa}(x_i), \qquad (13)
\end{aligned}$$

which tends to zero asymptotically $\forall i : \boldsymbol{\mu}_{\varphi_\mu}(x_i) \rightarrow \mathbf{y}_i$, and $\forall i : \kappa_{\varphi_\kappa}(x_i) \rightarrow \infty$, recalling that $\lim_{\kappa \rightarrow \infty} W(\kappa) = 1$. This trend is also documented in Fig. 5, where the concentration that minimizes the loss evolves towards large values for improved model fit. Eventually, the FvM distribution fitted to the training data will concentrate all its probability mass on the mean direction. As a consequence, the probabilistic model degenerates to the deterministic limit of Eq. (6), and this behavior is undesirable for down-stream information processing, where often access to typical samples is required rather than simply extracting the most-likely direction.

Moreover, at large concentrations the entropy of the output distributions is also minimal, which is detrimental for its robustness to noise, as we will discuss, together with a principled solution, in the next section.

3.3 Entropy Regularization

In the previous section we argued and demonstrated that probabilistic models can treat experimental settings with noise more effectively than their deterministic counterparts. Still the essential question remains open: How robust are different parametric distributions to the fluctuations generated by a particular data source?

The Maximum Entropy Principle, well known in physics and information theory (Jaynes 1957), provides an answer to this question of model uncertainty. In contrast to maximum-likelihood estimation, which requires assumptions about the parametric form of the desired distribution, the maximum-entropy approach is based on the knowledge about moments of the desired distribution. The maximum-entropy distribution obtains its robustness from the fact that it is the least informative distribution, which still fulfills the known constraints.

To put it differently, the change of the maximum-entropy distribution with respect to perturbations of the constraints is the least possible, as it avoids any overspecification, which is not supported by the data. In the context of spherical regression, the constraints are represented by the observations x and the observed direction $\mathbf{y}(x)$. Each observation provides a constraint in the entropy maximization over posterior distributions $p(\mathbf{s} \mid x)$ in the following sense:

$$\max_{p(.|x)} \mathbb{E}_{p(\mathbf{s}|x)}\big(-\log p(\mathbf{s} \mid x)\big) \quad \text{s.t.} \quad \mathbb{E}_{p(\mathbf{s}|x)}\mathbf{s} = w\, \mathbf{y}(x), \qquad (14)$$

where $\mathbb{E}_{p(\mathbf{s}|x)}$ denotes the expectation with respect to the distribution $p(\mathbf{s} \mid x)$. The parameter $w \in [0, 1]$ controls the width of the distribution $p(\mathbf{s} \mid x)$, i.e. $p(\mathbf{s} \mid x) = \delta\big(\mathbf{s} - \mathbf{y}(x)\big)$, if $w = 1$, and $p(\mathbf{s} \mid x) = C(0)$, if $w = 0$. The constraint can also be written as $\mathbb{E}_{p(\mathbf{s}|x)}\big\langle \mathbf{s}, \mathbf{y}(x) \big\rangle = w$, which also shows that w should be interpreted as the amount of spread that $p(\mathbf{s} \mid x)$ has around the observations $\mathbf{y}(x)$. Using a Lagrange multiplier $\beta \geq 0$, we can rewrite the constrained optimization problem in Eq. (14) as an unconstrained problem $\min_{p(.|x)} g_\beta\big(\mathbf{y}(x), p(. \mid x)\big)$, with

$$g_\beta\left(\mathbf{y}(x), p(. \mid x)\right) = -\mathbb{E}_{p(\mathbf{s}|x)} \left\langle \mathbf{s}, \mathbf{y}(x) \right\rangle + \frac{1}{\beta} \mathbb{E}_{p(\mathbf{s}|x)} \log p(\mathbf{s} \mid x) \qquad (15)$$

The functional g_β represents the Gibbs free energy, which is the difference between the expected cost $-\mathbb{E}_{p(\mathbf{s}|x)}\big\langle \mathbf{s}, \mathbf{y}(x) \big\rangle$, and the entropy $-\mathbb{E}_{p(\mathbf{s}|x)} \log p(\mathbf{s} \mid x)$ divided by the Lagrange factor β, controlling the precision of the direction estimation. The precision is determined by the value w in the constraint of Eq. (14), and needs to be considered as a hyper-parameter if we only observe $\mathbf{y}(x)$.

By variational calculus, we can derive the distribution that minimizes Eq. (15) for *one* particular x:

$$p(\mathbf{s} \mid x) = C(\beta) \exp\left(\beta \langle \mathbf{s}, \mathbf{y}(x) \rangle\right) \quad \text{with} \quad W\big(\beta\big) = w, \qquad (16)$$

which corresponds to the FvM distribution with a fixed concentration β around the mean direction $\mathbf{y}(x)$. While it is well known that for the constraints in Eq. (14) the maximum-entropy distribution is given by the FvM (Mardia 1975), we are rather interested in learning the conditional distribution

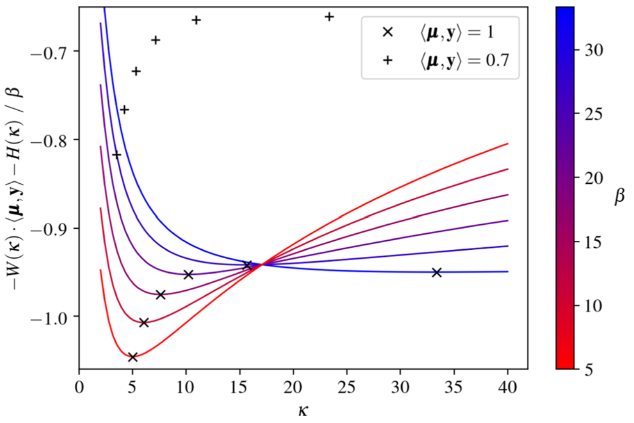


Fig. 6 Illustration of the Gibbs free energy minima (Eq. (18), *crosses*) as a function of κ, at different precisions β. We show two different, but fixed deviations $\langle \mathbf{y}, \boldsymbol{\mu} \rangle$ ($\times$ *and* $+$). Colors indicates low β (*red*), and high β (*blue*). We omitted the solid lines for $\langle \mathbf{y}, \boldsymbol{\mu} \rangle = 0.7$ to avoid clutter. Figure reproduced with modifications from Wegmayr et al. (2019) (Color figure online).

$p(. \mid x)$ over *all* x, which minimizes the *expected* Gibbs free energy

$$G_\beta := \mathbb{E}_{p(x)} g_\beta \left(\mathbf{y}(x),\, p^{\mathrm{FvM}}(. \mid x) \right), \tag{17}$$

for an arbitrary, but fixed data distribution $p(x)$.

Based on the observation that the FvM functional in Eq. (16) minimizes g_β, we make the ansatz $p(. \mid x) = p^{\mathrm{FvM}}\left(. \mid \boldsymbol{\mu}(x), \kappa(x)\right)$, which replaces the constant β with the input-dependent concentration $\kappa(x)$, and the unknown function $\mathbf{y}(x)$ with the estimator $\boldsymbol{\mu}(x)$. If we plug this into Eq. (15), we obtain the proposed entropy-regularized loss function for the estimators $\boldsymbol{\mu}(x), \kappa(x)$:

$$g_\beta \left(\mathbf{y}(x),\, p^{\mathrm{FvM}}(. \mid x) \right) = -W\left(\kappa(x)\right) \left\langle \mathbf{y}(x), \boldsymbol{\mu}(x) \right\rangle - \frac{1}{\beta} H\left(\kappa(x)\right). \tag{18}$$

The entropy-regularized objective in Eq. (18) has a similar loss attenuating property as the loss based on the FvM log-likelihood from Eq. (8). As shown in Fig. 6 for $d = 3$, the free energy is minimized at lower concentration $\kappa(x)$, when the deviation $\langle \mathbf{y}(x), \boldsymbol{\mu}(x) \rangle$ increases.

However, whereas the concentration diverges in the maximum-likelihood approach when $\langle \mathbf{y}(x), \boldsymbol{\mu}(x) \rangle \to 1$, it remains finite with the maximum-entropy method even in this case, because the precision parameter β limits the concentration, as we will discuss in more detail in the next section.

3.4 Automatic Annealing Schedule

In analogy to Sect. 3.2, we denote the parameters of the parametrized FvM posterior as φ_β. We propose to determine these parameters by minimizing the expected Gibbs free energy G_β (Eq. (17)). In a learning setting, we substitute the data distribution $p(x)$ in G_β by the empirical distribution of the observed data x_i and $\mathbf{y}_i := \mathbf{y}(x_i)$. The estimated parameters $\hat{\varphi}_\beta$ of the FvM distribution are

$$\hat{\varphi}_\beta = \arg\min_{\varphi} \frac{1}{n} \sum_{i=1}^{n} g_\beta \left(\mathbf{y}_i,\, p_\varphi^{\mathrm{FvM}}(. \mid x_i) \right). \tag{19}$$

To determine the optimal precision parameter β, we need to obtain the posterior parameters $\hat{\varphi}_\beta$ at different values of β. While this sounds conceptually straightforward, more considerations are necessary in practice. To compare models at different precision *values*, we need to assert, that the optimization of the model parameters φ_β has indeed *equilibrated* at the given precision value β. More formally, we consider the model parameters φ_β in equilibrium at a given precision β, when the following condition holds:

$$\frac{1}{\beta} = \frac{1}{n} \sum_i \frac{\left\langle \mathbf{y}_i, \boldsymbol{\mu}(x_i) \right\rangle}{\kappa(x_i)}. \tag{20}$$

This condition can be motivated by the gradient of the risk with respect to the parameters of the concentration, i.e.

$$\begin{aligned} &\nabla_{\varphi_\kappa} \frac{1}{n} \sum_{i=1}^{n} g_\beta \left(\mathbf{y}_i,\, p_\varphi^{\mathrm{FvM}}(. \mid x_i) \right) \\ &\quad = -\frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{\beta} - \frac{\left\langle \mathbf{y}_i, \boldsymbol{\mu}_{\varphi_\mu}(x_i) \right\rangle}{\kappa_{\varphi_\kappa}(x_i)} \right) \frac{\partial H}{\partial \kappa} \left(\kappa_{\varphi_\kappa}(x_i) \right) \nabla_{\varphi_\kappa} \kappa_{\varphi_\kappa}(x_i), \end{aligned} \tag{21}$$

which shows that the equilibrium condition in Eq. (20) is fulfilled, if the gradient vanishes. Additionally, we see again that the concentration κ remains finite, and is limited by the precision β, in contrast to Eq. (13).

In practice, it will depend on the optimization parameters (learning rate, batch size, etc.), and the precision itself, if the condition Eq. (20) is approximately fulfilled. Besides the cumbersome tuning of optimization parameters, it is very time-consuming to re-run the optimization for each precision value with a new initialization of φ.

Thus, we propose a robust, automatic annealing schedule to efficiently produce models in equilibrium at different precision values. The detailed annealing procedure is described by Algorithm 1, and its effect is illustrated in Fig. 7. Effectively, the progress of optimization is automatically paced by using Eq. (20) as a control criterion. As long as equilibrium is not established for a particular precision value, this value is held constant until the optimization of the model has equilibrated. This computational strategy renders model adaptation more robust to the choice of optimization parameters.

Moreover, we can efficiently extract models for different precision values during the course of a single training run, without re-initializing the parameters.

Figure 8 illustrates the joint distribution of $\kappa(x_i)$, and $\left\langle \mathbf{y}_i, \boldsymbol{\mu}(x_i) \right\rangle$ at one point during a typical annealing process,

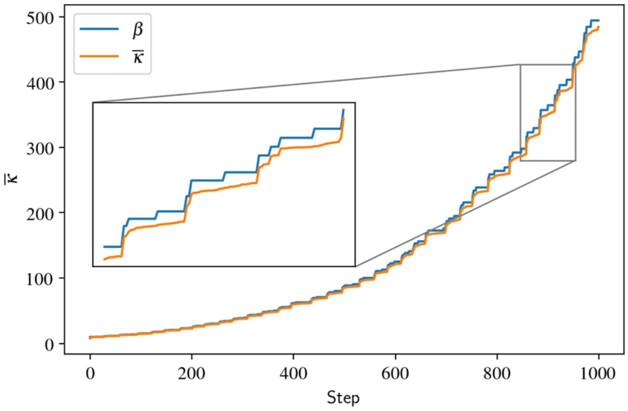


Fig. 7 Typical training run: Evolution of the average concentration $\overline{\kappa} = \frac{1}{n}\sum_i \kappa(x_i)$ (*orange*) in response to the automatic annealing schedule described in Algorithm 1. The precision β (*blue*) is paced automatically in accordance with the optimization progress (Color figure online).

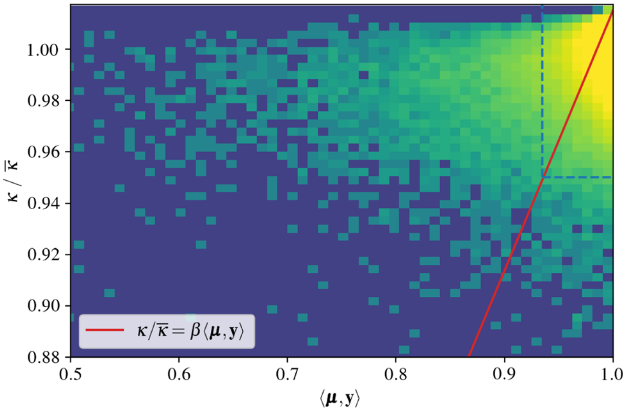


Fig. 8 Joint distribution of per-sample concentration (*y-axis, scaled by average concentration* $\overline{\kappa} = \frac{1}{n}\sum_i \kappa(x_i)$), and per-sample cosine deviation (*x-axis*) during a typical training run. The average per-sample concentration is concentrated around the precision, i.e. $\overline{\kappa}/\beta = 0.985$. For reference, the dashed-blue box contains 95.6% of the samples. The theoretical line of equilibrium is given by $\kappa = \beta\langle\mathbf{y},\boldsymbol{\mu}\rangle$ (*red line*) (Color figure online).

which shows that the model is indeed approximately at equilibrium. To see the effect of β on the optimization of the parameters of $\boldsymbol{\mu}$ more clearly as well, we also consider the gradient of the loss in Eq. (18) with respect to $\boldsymbol{\mu}$. To maintain the unit-length constraint, we assume that $\boldsymbol{\mu} = \mathbf{z}/\|\mathbf{z}\|_2$. Thus, the gradient is given by

$$\nabla_{\mathbf{z}}\, g_\beta\left(\mathbf{y}(x),\, p^{\mathrm{FvM}}(.\mid x)\right) = W(\kappa)\left(\frac{\mathbf{y}}{\|\mathbf{z}\|_2} - \langle\mathbf{y},\mathbf{z}\rangle\frac{\mathbf{z}}{\|\mathbf{z}\|_2^3}\right), \tag{22}$$

where we have written $\mathbf{z}, \kappa, \mathbf{y}$ instead of $\mathbf{z}(x), \kappa(x), \mathbf{y}(x)$ for brevity. The gradient vanishes when $\mathbf{z} = \mathbf{y}$, and the optimum with respect to $\boldsymbol{\mu}$ does not depend on the precision β. In practice, however, when we use gradient-descent to optimize the risk function, the magnitude of the gradient will be multiplied with $W(\kappa)$, which clearly depends on β. So while the optimum for $\boldsymbol{\mu}$ is still the same, we see that the precision influences the *effective learning rate* for the parameters of $\boldsymbol{\mu}$. At the beginning of the annealing schedule, the factor $W(\kappa)$ is small, because the precision is small (see also Fig. 7), i.e. the parameters of $\boldsymbol{\mu}$ are less susceptible to the deviation from the target $\mathbf{y}$. As the precision is gradually increased, the effective learning increases as well, and the gradient updates of $\boldsymbol{\mu}$ will push it stronger towards $\mathbf{y}$.

Algorithm 1 Annealing Schedule

Require: Training set $\{(x_i, \mathbf{y}_i)\}_{i=1\ldots n}$, learning rate η, start and end precision $\beta_0 < \beta_s$, growth rate $\gamma > 1$, tolerance $\epsilon \in (0, 1)$.
Ensure: Posterior parameters at equilibrium $\Phi = (\hat{\varphi}_{\beta_0}, \ldots, \hat{\varphi}_{\beta_s})$.

```
$\Phi \leftarrow \emptyset$
$\varphi \leftarrow \varphi_0$                                   ▷ Parameter initialization
$\overline{\beta} \leftarrow 0$
$\beta \leftarrow \beta_0$
while $\beta < \beta_s$ do
    while $|1 - \overline{\beta}/\beta| > \epsilon$ do                ▷ Check equilibrium eq. (20)
        $\varphi \leftarrow \varphi - \eta\nabla_\varphi \frac{1}{n}\sum_{i=1}^n g_\beta(\mathbf{y}_i, p_\varphi^{\mathrm{FvM}}(.\mid x_i))$
        $\overline{\beta} \leftarrow \frac{1}{n}\sum_i \langle\mathbf{y}_i, \boldsymbol{\mu}(x_i)\rangle/\kappa(x_i)$
    end while
    $\Phi \leftarrow \Phi \cup \varphi$
    $\beta \leftarrow \gamma \cdot \beta$
end while
return $\Phi$
```

To recapitulate, we have discussed how the precision parameter β constrains the *average* concentration of the FvM posterior during training, and how we can consistently, and efficiently obtain model parameters at different levels of precision. In the next section, we address the question of how to determine the optimal precision based on the noise in the data.

3.5 Optimal Precision by Posterior Agreement

We have seen in the previous two sections, that the proposed entropy regularization effectively limits the concentration $\kappa(x)$, however, it introduces the undetermined precision hyper-parameter β.

A common strategy to determine hyper-parameters would be cross-validation with respect to the generalization error $-\sum_i \langle\mathbf{y}_i, \boldsymbol{\mu}(x_i)\rangle$ on a validation set. However, cross-validation of this kind does not provide a solution here, because we have shown in the last section that the optimum of the function $\boldsymbol{\mu}(x)$ is not affected by the precision (Eq. (22)).

One could object, that the generalization error does not entirely reflect the learned posterior, but only its mean direction, and that we should rather compute the *expected* generalization error $\sum_i \rho_\beta(x_i, \mathbf{y}_i)$ with

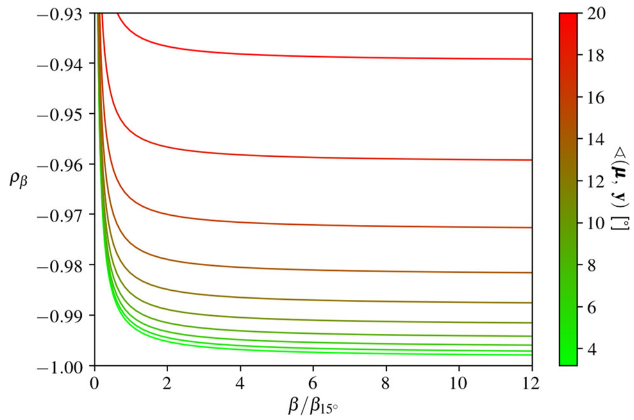


Fig. 9 Generalization error ρ_β of the FvM posterior. While ρ_β decreases with the deviation between $\boldsymbol{\mu}$ and $\mathbf{y}$, its minimum is always achieved for $\beta \to \infty$

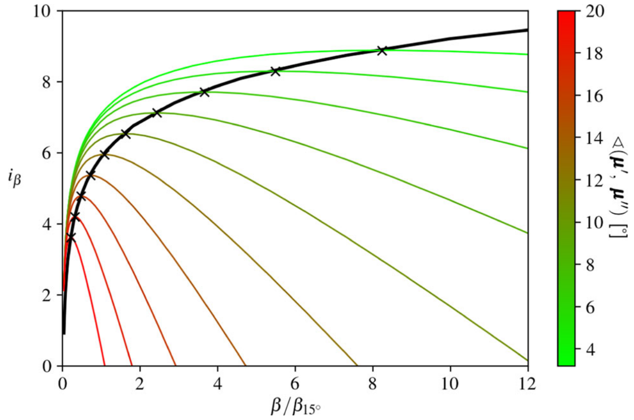


Fig. 10 The optimal precision (*crosses*) of the FvM posterior agreement i_β is high for low deviation between $\boldsymbol{\mu}'$ and $\boldsymbol{\mu}''$ (*green*), and decreases for larger deviations (*red*). The agreement optima coincide with the number of effective, conic bits (*black line*), given by Eq. (30) (Color figure online).

$$
\begin{aligned}
\rho_\beta(x,\mathbf{y}) &= -\int_{\mathbb{S}_{d-1}} \langle \mathbf{y},\mathbf{s}\rangle p_\beta^{\mathrm{FvM}}(\mathbf{s}\mid x)\mathrm{d}\mathbf{s} \\
&= -\langle \mathbf{y},\boldsymbol{\mu}(x)\rangle W\left(\kappa_\beta(x)\right)
\end{aligned}
\tag{23}
$$

where we have defined $p_\beta^{\mathrm{FvM}} := p_{\hat{\varphi}_\beta}^{\mathrm{FvM}}$, and κ_β analogously. Moreover, note again that $\boldsymbol{\mu}$ does not carry the precision subscript to indicate that it does not depend on β. Even though the expected generalization error depends on the precision, it does *not* have an optimum for finite β, as illustrated in Fig. 9. Instead, the minimum is always achieved at $\beta \to \infty$, which corresponds to the well-known empirical risk minimizer. This argument shows that we can not determine the optimal width of the posterior by minimizing risk since it introduces a bias which underestimates uncertainty.

Instead, we need a criterion, which can assess the stability of the posterior distribution with respect to data fluctuations. For this purpose, we propose a method motivated by the information-theoretic framework of expected log-posterior agreement. It is applicable to any Gibbs posterior distribution, if we have access to repeated measurements, which are used to assess the noise level in the data, and to calibrate the precision β accordingly. Specifically, we require a validation set, which provides two independent realizations x_i', x_i'' for each measurement i, i.e. a set $\{(x_i', x_i'')\}_{i=1\dots n}$. In the context of spherical regression, we define the PA for one repeated measurement as

$$
\begin{aligned}
& i_\beta(x',x'') := \\
& \quad \log_2\left(\max\left\{C(0)^{-1}\int_{\mathbb{S}_{d-1}} p_\beta^{\mathrm{FvM}}(\mathbf{s}\mid x')p_\beta^{\mathrm{FvM}}(\mathbf{s}\mid x'')\mathrm{d}\mathbf{s}\ ,\, 1\right\}\right),
\end{aligned}
\tag{24}
$$

where $C(0)$ is the normalization of the uniform distribution as defined in Eq. (5). Performing the integration over all directions $\mathbf{s}$, the PA reads

$$
i_\beta(x',x'') = \log_2\left(\max\left\{C(0)^{-1}\frac{C(\kappa_\beta')C(\kappa_\beta'')}{C\left(\|\kappa_\beta'\boldsymbol{\mu}' + \kappa_\beta''\boldsymbol{\mu}''\|_2\right)}\ ,\, 1\right\}\right),
\tag{25}
$$

where we have written $\kappa_\beta' = \kappa_\beta(x')$, etc. for brevity. The maximal agreement is realized between the following two limiting cases:

(i) When the posterior distribution is very broad, i.e. it does not contain any information about the mean direction, the PA is zero:

$$
\lim_{\beta\to 0} i_\beta(x',x'') = 0.
\tag{26}
$$

(ii) For highly peaked posteriors with different mean directions due to noisy measurements, the agreement also vanishes:

$$
\boldsymbol{\mu}' \neq \boldsymbol{\mu}'' \Rightarrow \exists\beta > 0 : \frac{C(\kappa_\beta')C(\kappa_\beta'')}{C\left(\|\kappa_\beta'\boldsymbol{\mu}' + \kappa_\beta''\boldsymbol{\mu}''\|_2\right)} \leq C(0)
\tag{27}
$$

Indeed, if we increase β sufficiently high, the agreement integral drops below the value $C(0)$ achieved in the uniform case, and the posterior agreement $i_\beta(x',x'')$ becomes zero again. We note, that the PA also vanishes for $\beta > 0$, if either of the two measurements is uninformative, i.e. uniform by either $\kappa_\beta' = 0$ or $\kappa_\beta'' = 0$.

In Fig. 10 we illustrate the behavior of i_β when $\kappa_\beta' = \kappa_\beta'' = \beta$, showing clearly the discussed trade-off between low precision and high precision.

To determine the optimal precision based on a validation set, we compute the average over all n repeated measurements

$$\mathcal{I}_\beta = \frac{1}{n}\sum_{j=1} i_\beta(x'_j, x''_j), \tag{28}$$

and maximize it with respect to β:

$$\hat{\beta} = \arg\max_{\beta\in[0,\infty)} \mathcal{I}_\beta \,. \tag{29}$$

Moreover, we can assign an interesting interpretation to the numeric values of $i_{\hat{\beta}}$. Let us consider the solid angle $\Omega_p(\beta)$ centered on an arbitrary, but fixed $\boldsymbol{\mu}$, which contains p% of the probability mass of a general $p^{\mathrm{FvM}}(\mathbf{s} \mid \boldsymbol{\mu}, \beta)$ distribution. We use it to partition the unit sphere into $4\pi/\Omega_p(\beta)$ effective, conic *bins*. In this sense, the precision β determines a quantization angle on the sphere. If we measure the number of effective, conic *bits* with the binary logarithm, we find that

$$i_{\hat{\beta}} \simeq \log_2\left(\frac{4\pi}{\Omega_{99.5}(\hat{\beta})}\right) \tag{30}$$

as shown by the black line in Fig. 10.

This result is interesting, because it suggests that the value of i_β corresponds to the bit-rate of a noisy communication scenario, as described in the original, information-theoretic derivation by Buhmann (2010).

Besides, we can use it to define an intuitive scale for the directional precision by calibrating it with respect to β_θ, where θ is the aperture angle of the solid angle $\Omega_{99.5}(\beta)$. More precisely, the relationship between β and θ is given by

$$\Omega_{99.5}(\beta) = 2\pi(1-\cos\theta). \tag{31}$$

Unless indicated otherwise, we scale all experimental precision values with respect to $\beta_{15^\circ} = 114.40$.

4 The Entrack Posterior for Tractography

In this section, we apply the entropy-regularized, probabilistic regression objective from Eq. (18) to streamline tractography on DWI measurements.

A DWI measurement I records the diffusion signal at every location $\mathbf{r}$ in the measurement volume $\mathcal{V} \subset \mathbb{R}^3$ [2] , i.e. $I : \mathcal{V} \times (\mathbb{S}_2)^N \to \mathbb{R}_+,\ (\mathbf{r}, \mathbf{g}_n) \mapsto I_\mathbf{r}(\mathbf{g}_n)$. Essentially, $I_\mathbf{r}(\mathbf{g}_n)$ corresponds to the magnitude of the *local* diffusion signal along one of the N experimentally *fixed* magnetic gradient directions $\mathbf{g}_n$. The diffusion signal exhibits an invariance to inversion of the gradient directions $\mathbf{g}_n$, i.e. $I_\mathbf{r}(-\mathbf{g}_n) = I_\mathbf{r}(\mathbf{g}_n)$. In practice, it is common to work with a lower-dimensional feature representation $\mathbf{X}_f$ of the high-dimensional DWI measurement, i.e. $\mathbf{X}_f : \mathcal{V} \to \mathbb{R}^p,\ \mathbf{r} \mapsto \omega_\mathbf{r}$ such that $f(\mathbf{g}_n \mid \omega_\mathbf{r}) = I_\mathbf{r}(\mathbf{g}_n)$. The function f is an experimental aspect of tractography, and we discuss its concrete implementation in Sect. 5.1, but for the following considerations we assume it as fixed, i.e. $\mathbf{X} := \mathbf{X}_f$.

By measuring the DWI features $\mathbf{X}$ of the underlying brain tissue $\mathcal{T}$, the goal of a tractography algorithm $\mathcal{A}$ is to recover the corresponding long-range tissue connections $\mathbf{T}$, also referred to as *tractogram*:

$$\mathcal{T} \xrightarrow{I,f} \mathbf{X} \xrightarrow{\mathcal{A}} \mathbf{T}$$

A tractogram $\mathbf{T}$ is a set of $i = 1,\ldots,n$ variable-length streamlines $\mathbf{t}_i = (\mathbf{r}_{i,1},\ldots,\mathbf{r}_{i,n_i}) \in (\mathbb{R}^3)^{n_i}$, which should be understood as a *representation* of tissue connectivity rather than an anatomically faithful image of individual axons.

To learn the tractography mapping $\mathcal{A} : \mathbf{X} \to \mathbf{T}$, we factorize the joint posterior $p(\mathbf{T} \mid \mathbf{X})$ of an entire tractogram into the product of independent streamlines $\mathbf{t}_i$. Moreover, we factorize each streamline into the product of its segments $\mathbf{y}_{i,j} \propto \mathbf{r}_{i,j} - \mathbf{r}_{i,j-1}$, but retain nearest-neighbor interactions between successive segments. Thus, the posterior probability of the direction $\mathbf{y}_{i,j}$, described by the FvM distribution, is conditioned on the diffusion data $\mathbf{X}(\mathbf{r}_{i,j-1})$ at the location $\mathbf{r}_{i,j-1}$, and the incoming direction $\mathbf{y}_{i,j-1}$:

$$\begin{aligned} p^{\mathrm{trk}}(\mathbf{y}_{i,j} \mid \mathbf{X}(\mathbf{r}_{i,j-1}), \mathbf{y}_{i,j-1}) :=& \\ p^{\mathrm{FvM}}\big(\mathbf{y}_{i,j} \mid \boldsymbol{\mu}(\mathbf{X}(\mathbf{r}_{i,j-1}), \mathbf{y}_{i,j-1}), &\kappa(\mathbf{X}(\mathbf{r}_{i,j-1}), \mathbf{y}_{i,j-1})\big). \end{aligned} \tag{32}$$

Due to the tractography context, we refer to the posterior p^{trk} in Eq. (32) as *Entrack* posterior. Under these assumptions, we can write the joint tractogram posterior as

$$\begin{aligned} p(\mathbf{T} \mid \mathbf{X}) &= \prod_{i=1}^{n} p(\mathbf{t}_i \mid \mathbf{X}) \\ &= \prod_{i=1}^{n} p\big(\mathbf{y}_{i,1} \mid \mathbf{X}(\mathbf{r}_{i,1})\big)\, p(\mathbf{r}_{i,1}) \\ &\prod_{j=2}^{n_i} p^{\mathrm{trk}}(\mathbf{y}_{i,j} \mid \mathbf{X}(\mathbf{r}_{i,j-1}), \mathbf{y}_{i,j-1}), \end{aligned} \tag{33}$$

where we also have made explicit the need for priors of the fiber seed points $\mathbf{r}_{i,1}$, and the initial directions $\mathbf{y}_{i,1}$.

[2] In practice, $\mathcal{V} \subset \mathbb{Z}^3$, but we assume that a continuous measurement can be achieved, e.g. by interpolation.

4.1 Entrack: Learning the Local Posterior

Given a measurement of DWI features $\mathbf{X}$, and a corresponding reference tractogram $\mathbf{T}$ as supervision information for training, our goal is to learn the posterior distribution of local streamline direction $p^{\text{trk}}(\mathbf{y} \mid \boldsymbol{x}, \mathbf{y}^{in})$ based on the entropy-regularized Gibbs free energy presented in Eq. (18). Note, that we have denoted $\boldsymbol{x} \in \mathbb{R}^p$ as the variable for the local diffusion data, and $\mathbf{y}^{in} \in \mathbb{S}_2$ as the variable for the incoming fiber direction. Together, they constitute the input vector $x = (\boldsymbol{x}, \mathbf{y}^{in}) \in \mathbb{R}^p \times \mathbb{S}_2$, which the Entrack posterior conditions on.

In the following, we discuss how to decompose the data set $(\mathbf{X}, \mathbf{T})$ such that it can be used with the risk function introduced in Eq. (19). Specifically, we need to construct samples $\big((\boldsymbol{x}_i, \mathbf{y}_i^{in}), \mathbf{y}_i\big)$ to capture the relationship between the target direction $\mathbf{y}$ and the input $(\boldsymbol{x}, \mathbf{y}^{in})$. We detail the corresponding sample generation process in Algorithm 2, and illustrate it in Fig. 11. With an accordingly generated training set $\mathbb{X}$, we can use the entropy-regularized risk function from Eq. (19) to estimate the parameters of $p_\beta^{\text{trk}} := p_{\hat{\varphi}_\beta}^{\text{trk}}$.

However, we also need to account for the inversion symmetry of DWI data, which makes it equivalent to traverse a streamline in forward, and backward direction. To ensure that the posterior can learn this symmetry from the data, i.e. $p^{\text{trk}}(-\mathbf{y} \mid \boldsymbol{x}, -\mathbf{y}^{in}) = p^{\text{trk}}(\mathbf{y} \mid \boldsymbol{x}, \mathbf{y}^{in})$, we incorporate this invariance explicitly in the risk function by adding forwards $(u = +1)$, *and* backwards direction $(u = -1)$:

$$\hat{\varphi}_\beta = \arg\min_{\varphi} \frac{1}{2N} \sum_{k=1}^{N} \sum_{u \in \{\pm 1\}} g_\beta\big(u \cdot \mathbf{y}_k,\ p_\varphi^{\text{trk}}(. \mid \boldsymbol{x}_k, u \cdot \mathbf{y}_k^{in})\big), \tag{34}$$

where N refers to the number of sample streamline segments.

Algorithm 2 Sample Generation

```
Require: DWI features $\mathbf{X}$, tractogram $\mathbf{T} = \{(\mathbf{r}_{i,1}, \dots, \mathbf{r}_{i,n_i})\}_{i=1\dots n}$.
Ensure: Training set $\mathbb{X} = \Big\{\big((\boldsymbol{x}_k, \mathbf{y}_k^{in}), \mathbf{y}_k\big)\Big\}_{k=1\dots N}$
$\mathbb{X} \leftarrow \emptyset$
for $i = 1, \dots, n$ do
    for $j = 3, \dots, n_i$ do
        $\mathbf{y} \leftarrow (\mathbf{r}_{i,j} - \mathbf{r}_{i,j-1})/\|\mathbf{r}_{i,j} - \mathbf{r}_{i,j-1}\|_2$
        $\boldsymbol{x} \leftarrow \mathbf{X}(\mathbf{r}_{i,j-1})$
        $\mathbf{y}^{in} \leftarrow (\mathbf{r}_{i,j-1} - \mathbf{r}_{i,j-2})/\|\mathbf{r}_{i,j-1} - \mathbf{r}_{i,j-2}\|_2$
        $\mathbb{X} \leftarrow \mathbb{X} \cup \big((\boldsymbol{x}, \mathbf{y}^{in}), \mathbf{y}\big)$
    end for
end for
return $\mathbb{X}$
```

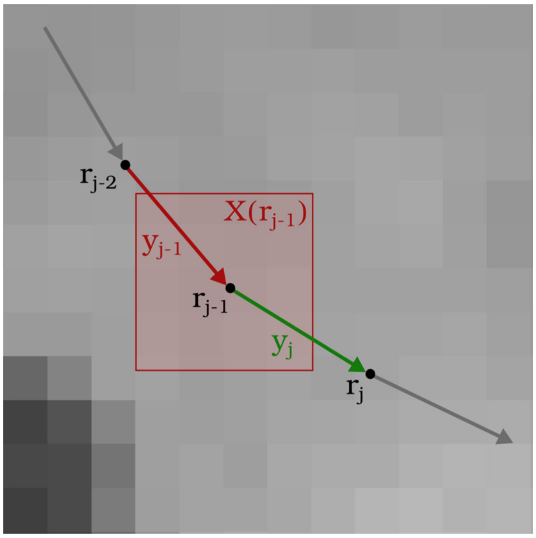


Fig. 11 Generation of a training sample from a reference streamline. For each point along a fiber we extract the local diffusion data $\mathbf{X}(\mathbf{r}_{j-1})$ (*red square*), and the incoming direction $\mathbf{y}_{j-1}$ (*red arrow*) as inputs, whereas the outgoing direction $\mathbf{y}_j$ serves as target (*green arrow*) (Color figure online).

4.2 Entrack: Streamline Inference

The trained FvM posterior p_β^{trk} is employed by the iterative tracking algorithm as described by Algorithm 3.

To construct a streamline, we start from a seed point $\mathbf{r}_1 \in \mathcal{V}$, and obtain the local DWI features $\mathbf{X}(\mathbf{r}_1)$. Provided with a prior direction $\mathbf{y}_1 \in \mathbb{S}_2$, e.g. from a diffusion-tensor fit, we can establish the next point $\mathbf{r}_2$ of the streamline by sampling a direction $\mathbf{y}_2$ from $p_\beta^{\text{trk}}\big(\mathbf{y} \mid \mathbf{X}(\mathbf{r}_1), \mathbf{y}_1\big)$, and setting $\mathbf{r}_2 = \mathbf{r}_1 + \alpha \mathbf{y}_2$, with a step size α. This iteration repeats, until a termination criterion is met, such as a thresholds on fiber length, strength of the diffusion signal, fiber bending angle, or leaving a predefined region of interest (ROI). The corresponding streamline $\mathbf{t}_i$ simply consists of the traversed points, i.e. $\mathbf{t}_i = (\mathbf{r}_{i,1}, \dots, \mathbf{r}_{i,n_i})$.

To obtain a dense tractogram $\mathbf{T} = \{\mathbf{t}_i\}_{i=1\dots n}$, we place n seed points within a specified ROI, e.g. within a white matter mask for whole-brain tractography.

4.3 Entrack: Posterior Agreement

In Sect. 3.5 we introduced the PA for general directional regression with the FvM posterior, and now we describe how we implement it for tractography to determine the optimal β for p_β^{trk}.

Given two independent DWI measurements $\mathbf{X}', \mathbf{X}''$ of the same subject, we denote the corresponding tractograms, obtained with Algorithm 3 in conjunction with p_β^{trk}, as $\mathbf{T}'_\beta, \mathbf{T}''_\beta$. The tractograms carry the subscript β, because they implicitly depend on the precision via the Entrack posterior

Algorithm 3 Iterative Streamline Tractography

Require: DWI features $\mathbf{X}$, seed point $\mathbf{r}_1$, prior direction $\mathbf{y}_1$, posterior of local fiber direction $p_\beta^{\mathrm{trk}}(\mathbf{y} \mid \boldsymbol{x}, \mathbf{y}^{in})$, step size α.
Ensure: Predicted streamline $\mathbf{t}$.

```
$\mathbf{r} \leftarrow \mathbf{r}_1$
$\mathbf{t} \leftarrow \{\mathbf{r}_1\}$
$\mathbf{y}^{in} \leftarrow \mathbf{y}_1$
$\mathbf{y} \sim p_\beta^{\mathrm{trk}}(. \mid \mathbf{X}(\mathbf{r}_1), \mathbf{y}_1)$
while terminate$(\mathbf{r}, \mathbf{y}, \mathbf{y}^{in}, \mathbf{X}(\mathbf{r}), \mathbf{t}) \neq$ True do ▷ Tracking Iteration
    $\mathbf{r} \leftarrow \mathbf{r} + \alpha\mathbf{y}$
    $\mathbf{t} \leftarrow \mathbf{t} \cup \mathbf{r}$
    $\mathbf{y}^{in} \leftarrow \mathbf{y}$
    $\mathbf{y} \sim p_\beta^{\mathrm{trk}}(. \mid \mathbf{X}(\mathbf{r}), \mathbf{y}^{in})$
end while
return $\mathbf{t}$
```

p_β^{trk}. We recall that the PA $i_\beta(x', x'')$ from Eq. (25) depends on two measurements x', x'' of the same input, and the input to the Entrack posterior consists of two components, i.e. $x = \big(\mathbf{X}(\mathbf{r}), \mathbf{y}^{in}\big)$. Given proper image registration between $\mathbf{X}'$ and $\mathbf{X}''$, it is straightforward to match the repeated measurements of the DWI data by considering the same location, i.e. $\mathbf{X}'(\mathbf{r}), \mathbf{X}''(\mathbf{r})$.

To obtain $\mathbf{y}^{in}(\mathbf{r})', \mathbf{y}^{in}(\mathbf{r})''$ from the discrete streamlines of the two tractograms $\mathbf{T}'_\beta, \mathbf{T}''_\beta$ we consider a small volume around the location $\mathbf{r}$, and compute $\mathbf{y}^{in}(\mathbf{r})', \mathbf{y}^{in}(\mathbf{r})''$ based on the corresponding streamlines which pass through this *voxel*. Thus, the continuous measurement volume $\mathcal{V}$ is decomposed into little cuboids with voxel size $a > 0$, that are indexed by their discrete location inside the measurement volume, i.e. $\mathbf{z} \in \mathcal{V}_a = \mathcal{V} \cap \{z \cdot a \mid z \in \mathbb{Z}^3\}$. We refer to the volume of a voxel as $v_\mathbf{z} = \{\mathbf{r} : \|\mathbf{r} - \mathbf{z}\|_\infty \leq a/2\}$. Even though we can now compute $\mathbf{y}^{in}(\mathbf{z} \mid \mathbf{T}) \propto \sum_{i,j} \mathbb{I}\{r_{i,j} \in v_\mathbf{z}\}\mathbf{y}_{i,j}$, we still need to take into account that independent streamline bundles may cross at the same voxel, and they must not be confused with each other.

Instead we need to consider each bundle b separately, and condition the local direction on the bundle, too, i.e. $\mathbf{y}^{in}(\mathbf{z}, b \mid \mathbf{T})$. More precisely, we consider a bundle b as a set of coherent streamlines, which are similar in the sense that the average pointwise distance is small for each pair of streamlines in a bundle. This way, we can partition a tractogram into a set of bundles $B(\mathbf{T}) = \{b_1, \ldots, b_k\}$ such that $\forall i, j : b_i \cap b_j = \emptyset \wedge \bigcup_i b_i = \mathbf{T}$. We also refer to Garyfallidis et al. (2012) for more details about the practical grouping of tractograms into bundles. Using the partitioned tractogram B, we can compute the local streamline direction per-bundle, up to normalization to unit-length, as

$$\mathbf{y}^{in}(\mathbf{z}, b \mid \mathbf{T}) \propto \sum_{i=1}^{|\mathbf{T}|} \sum_{j=1}^{n_i} \mathbb{I}\{\mathbf{t}_i \in b\}\mathbb{I}\{\mathbf{r}_{i,j} \in v_\mathbf{z}\}\mathbf{y}_{i,j}, \tag{35}$$

with $\mathbf{y}_{i,j} = (\mathbf{r}_{i,j} - \mathbf{r}_{i,j-1})/\|\mathbf{r}_{i,j} - \mathbf{r}_{i,j-1}\|_2$. Essentially, we have decomposed the tractogram $\mathbf{T}$ into the directions of its individual fiber bundles at each voxel, which is also known as *fixel* representation (Raffelt et al. 2017), referring to *"a specific fiber bundle within a specific voxel"*. Consequently, we can think of each tuple $(\mathbf{z}, b)$ as the coordinates of one fixel. We define the posterior mean direction of such a fixel as

$$\boldsymbol{\mu}_\beta(\mathbf{z}, b \mid \mathbf{X}, \mathbf{T}_\beta) := \boldsymbol{\mu}\big(\mathbf{X}(\mathbf{z}), \mathbf{y}^{in}(\mathbf{z}, b \mid \mathbf{T}_\beta)\big), \tag{36}$$

where $\boldsymbol{\mu}$ is the mean direction of the Entrack posterior p_β^{trk}.

Lastly, when we compute the posterior concentration of a fixel, i.e $\kappa_\beta(\mathbf{z}, b \mid \mathbf{X}, \mathbf{T}_\beta)$, we also need to take into account the number of streamlines which represent the summary direction $\mathbf{y}^{in}(\mathbf{z}, b \mid \mathbf{T}_\beta)$. Intuitively, we should be more certain about the summary direction, if it is represented by many fibers, i.e. the concentration should be increased [3]. Formally, the fixel concentration is scaled by the streamline density, i.e.

$$\kappa_\beta(\mathbf{z}, b \mid \mathbf{X}, \mathbf{T}_\beta) := n(\mathbf{z}, b \mid \mathbf{T}_\beta)\kappa_\beta\big(\mathbf{X}(\mathbf{z}), \mathbf{y}^{in}(\mathbf{z}, b \mid \mathbf{T}_\beta)\big), \tag{37}$$

where the streamline density is defined as

$$n(\mathbf{z}, b \mid \mathbf{T}) := \frac{1}{a^3} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \mathbb{I}\{\mathbf{t}_i \in b\}\mathbb{I}\{\mathbf{r}_{i,j} \in v_\mathbf{z}\} \tag{38}$$

In particular, the fixel concentration $\kappa_\beta(\mathbf{z}, b \mid \mathbf{X}, \mathbf{T}_\beta)$ is zero, i.e. the posterior doesn't contain any information about the fixel direction, when we don't observe any streamline. Putting everything together, we obtain the posterior agreement for one fixel, based on Eq. (25), as

$$2^{i_\beta(\mathbf{z},b)} = \max\left\{4\pi \frac{C\big(\kappa'_\beta(\mathbf{z}, b)\big)C\big(\kappa''_\beta(\mathbf{z}, b)\big)}{C\big(\|\kappa'_\beta(\mathbf{z}, b)\boldsymbol{\mu}'_\beta(\mathbf{z}, b) + \kappa''_\beta(\mathbf{z}, b)\boldsymbol{\mu}''_\beta(\mathbf{z}, b)\|_2\big)}, 1\right\} \tag{39}$$

where we have defined $\kappa'_\beta(\mathbf{z}, b) := \kappa_\beta(\mathbf{z}, b \mid \mathbf{X}', \mathbf{T}'_\beta)$, etc. for brevity. Consequently, the average PA over all fixels is given by

$$\mathcal{I}_\beta = \frac{1}{|\mathcal{V}_a|} \sum_{\mathbf{z} \in \mathcal{V}_a} \frac{1}{|B_\mathbf{z}|} \sum_{b \in B_\mathbf{z}} i_\beta(\mathbf{z}, b), \tag{40}$$

with the set of bundles that intersect a particular voxel denoted as $B_\mathbf{z} = \{b \in B(\mathbf{T}'_\beta \cup \mathbf{T}''_\beta) : \exists \mathbf{t} \in b : \exists \mathbf{r} \in \mathbf{t} : \mathbf{r} \in v_\mathbf{z}\}$.

[3] Refer to appendix A for a formal justification.

5 Experiments

We provide the entire code implementing this work at https://github.com/vwegmayr/tractography, which includes code for managing data acquisition, data preprocessing, sample generation, model training, model inference, and evaluation.

5.1 Data and Preprocessing

In the following we summarize the most important details about the DWI data and its preprocessing, i.e. how we obtain the DWI features $\mathbf{X}$.

ISMRM15 Data

The simulated DWI data, that was also used in the ISMRM15 challenge, can be obtained from http://tractometer.org/. We use the DWI data referred to as "basic data" on the challenge website. The corresponding DWI image has the shape $90 \times 108 \times 90 \times 33$, with 32 gradient directions $b = 1000$ s/mm^2, plus one acquisition with $b = 0$ s/mm^2. The voxel size is 2 mm.

We preprocess the DWI image according to the standard preprocessing pipeline described by Glasser et al. (2013), using the MRtrix tool (https://mrtrix.org/). This procedure includes the following steps, where we indicate the corresponding MRtrix commands in parentheses:

1. Basic denoising (*dwidenoise*)
2. Eddy current & motion correction (*dwipreproc*)
3. B$_0$ intensity normalization (*dwinormalize*)

After preprocessing, we estimate the DWI features for every voxel $\mathbf{z}$ in terms of fiber orientation distribution (FOD) coefficients $\mathbf{X}_{FOD} : \mathcal{V} \to \mathbb{R}^{15}; \mathbf{r} \mapsto \{D^{lm}_{\mathbf{r}}\}$: [4]

1. Response function estimation (*dwi2response*)
2. Constrained spherical deconvolution (*dwi2fod*)
3. Log-Domain intensity normalization (*mtnormalise*)

HCP Data

The HCP diffusion data, accessible at the website https://db.humanconnectome.org/, are already preprocessed according to the standard preprocessing pipeline by Glasser et al. (2013). The DWI image has the shape $145 \times 174 \times 145 \times 108$, and we extract 90 gradient directions with $b = 1000$ s/mm^2, plus 18 interlaced acquisitions with $b = 0$ s/mm^2. The voxel size is 1.25 mm.

We perform the same procedure to estimate the per-voxel FOD as described for the ISMRM15 data.

TractSeg Streamlines

The TractSeg dataset (Wasserthal et al. 2018) is a collection of high-quality white matter reference tracts for 105 subjects, whose diffusion data is also included in the HCP dataset. It can be downloaded at https://zenodo.org/record/1477956. Each reference tractogram contains $\sim$ 1.7 million fibers, grouped into 72 reference bundles, which amount to $\sim$ 70 million fiber segments in total.

For training, we use the tractogram of subject 992774, and reduce it to 20% of its size by sub-sampling the streamlines, weighted by bundle-size to ensure that small bundles are not under-represented.

5.2 Entrack Model Architecture and Training

In this section, we discuss our implementation of the Entrack posterior $p^{\text{trk}}\big(\mathbf{y} \mid \boldsymbol{\mu}(\mathbf{X}(\mathbf{r}), \mathbf{y}^{in}), \kappa(\mathbf{X}(\mathbf{r}), \mathbf{y}^{in})\big)$, in particular the implementation of the functions $\boldsymbol{\mu}, \kappa$.

While the general formulation supports a wide range of possible functions, we chose a deep neural network model due to its superior ability to extract patterns automatically (Goodfellow et al. 2015). Moreover, neural networks (NN) naturally support modular architectures, which allows us to readily formulate $\boldsymbol{\mu}, \kappa$ in terms of two output modules NN_μ, NN_κ, based on a shared NN module NN_z:

$$
\begin{aligned}
\boldsymbol{\mu}\big(\mathbf{X}(\mathbf{r}), \mathbf{y}^{in}\big) &= \text{NN}_\mu\Big(\mathbf{z}\big(\mathbf{X}(\mathbf{r}), \mathbf{y}^{in}\big)\Big) \\
\kappa\big(\mathbf{X}(\mathbf{r}), \mathbf{y}^{in}\big) &= \text{NN}_\kappa\Big(\mathbf{z}\big(\mathbf{X}(\mathbf{r}), \mathbf{y}^{in}\big)\Big) \\
\mathbf{z}\big(\mathbf{X}(\mathbf{r}), \mathbf{y}^{in}\big) &= \text{NN}_z\big(\mathbf{X}(\mathbf{r}), \mathbf{y}^{in}\big)
\end{aligned}
\tag{41}
$$

Specifically, each NN module is a series of fully-connected layers, as shown in Fig. 12, together with the detailed parameters. The inputs $\mathbf{X}(\mathbf{r}), \mathbf{y}^{in}$ are both flattened, and concatenated to form a 408-dimensional input vector, i.e. 3 dimensions for $\mathbf{y}^{in}$, and $405 = (3 \times 3 \times 3) \times 15$ dimensions for $\mathbf{X}(\mathbf{r})$, which represents the 15 FOD coefficients ($l = 4$) for each voxel in a $3 \times 3 \times 3$ cube centered on the location $\mathbf{z} = a[\mathbf{r}/a]$, where a is the voxel size.

Model Training

The described NN model for the Entrack posterior is trained on samples obtained from the TractSeg streamlines of subject 992774, using the sample generation procedure described by Algorithm 2.

[4] Please refer to appendix C for more details.

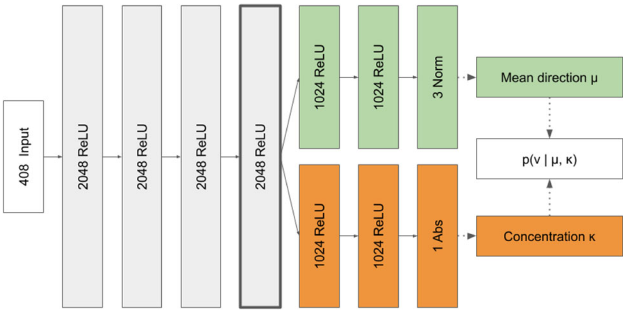


Fig. 12 Neural network implementation of the parametrized Entrack posterior. The input layer is the 408-dimensional vector $(\boldsymbol{x}, \mathbf{y}^{in})$. The shared layers (*gray boxes*) consist of 4 fully-connected layers, each with 2048 hidden units, using the ReLU activation function. The two output branches for $\boldsymbol{\mu}$ (*green boxes*), and κ (*orange boxes*), share the same hidden representation $\mathbf{z}$ (*bold margin*). Each of them has its appropriate activation function, i.e. normalization to unit-norm for $\boldsymbol{\mu}$, and absolute value for the positive κ (Color figure online).

For parameter optimization, we use the annealing scheme described in Algorithm 1 with $\eta = 0.99$, $\epsilon = 0.01$, $\beta_0 = 10$, $\beta_s = 1000$. Moreover, we use the Adam optimizer (Kingma and Ba 2014) with a learning rate of $2 \cdot 10^{-4}$, and a batch size of 512. Note, the number of training epochs depends on how fast the annealing proceeds, but in our experiments it typically reached the target precision within 30 epochs.

5.3 Local Case Studies of Entrack Posterior

To better understand which patterns the Entrack model has recognized in the training data, we perform a series of experiments on prototypical inputs.

Single Fiber Direction

In this setup, we investigate how $\boldsymbol{\mu}\left(\mathbf{X}(\mathbf{r}), \mathbf{y}^{in}\right)$, and $\kappa\left(\mathbf{X}(\mathbf{r}), \mathbf{y}^{in}\right)$ behave when we rotate $\mathbf{y}^{in}$, while keeping $\mathbf{X}(\mathbf{r})$, and β fixed. For this purpose, we select DWI features $\mathbf{X}(\mathbf{r})$ from a voxel in the corpus callosum (see inset of Fig. 13a), which exhibits a clear, unidirectional DWI signal visualized by the gray FOD in Fig. 13b.

In Fig. 13a we consider the change of κ, when $\mathbf{y}^{in}$ is rotated in-plane, relative to the fixed DWI input. More precisely, the figure shows the function $\kappa\left(\mathbf{X}(\mathbf{r}), \mathbf{R}_\theta \mathbf{e}_x\right)$, where $\mathbf{R}_\theta$ is a 3×3 rotation matrix whose rotation axis is perpendicular to the plane of view, and $\mathbf{e}_x = (1, 0, 0)$.

We observe, that the concentration of the Entrack posterior is largest along the DWI main direction, and decreases when perpendicular. This behavior makes sense, because we expect the uncertainty to be small, when the incoming fiber direction agrees with the local diffusion data, and to increase when they disagree.

Moreover, the angular profile is approximately inversion-symmetric, as should be expected from the properties of DWI data. In Fig. 13b we consider the probability of proceeding along $\mathbf{y}^{in}$, i.e. $\log p^{\text{trk}}(\mathbf{y}^{in} \mid \mathbf{X}(\mathbf{r}), \mathbf{y}^{in})$, again with $\mathbf{y}^{in} = \mathbf{R}_\theta \mathbf{e}_x$.

As expected, the probability to follow the previous direction, is the highest when it is aligned with the diffusion data.

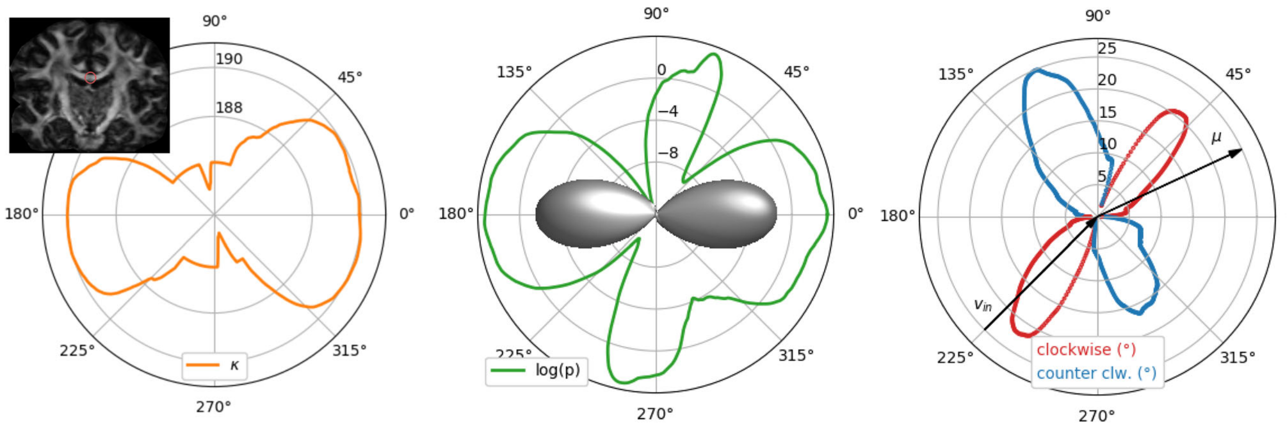
However, it is also high, when $\mathbf{y}^{in}$ is *perpendicular* to the main direction of diffusion. We can understand this unexpected behavior in the sense, that the model recognizes situations where the incoming direction clearly contradicts the present direction of diffusion. But instead of predicting a suboptimal superposition between the previous direction and the DWI main direction, the non-linear model favors continuity with respect to the incoming direction.

This interpretation is also supported by Fig. 13c, which shows the amount and direction of deflection of $\boldsymbol{\mu}$ from $\mathbf{y}^{in}$. The two black arrows in the figure represent one exemplaric pair $\mathbf{y}^{in}$ and $\boldsymbol{\mu}(\boldsymbol{x}, \mathbf{y}^{in})$ to illustrate the deflection. The exemplaric incoming direction has an incidence angle of about $\theta = 45°$, for which we read off a deflection of ca. $+20°$, as shown by the radius and color of the intersecting lobe. Thus, the mean-direction predicted by the model is rotated 20° *clockwise* with respect to the incoming direction.

This example shows, that the model pushes the incoming direction closer to the main direction of diffusion, *if they sufficiently agree*. In contrast, when the incoming direction does not relate to the diffusion data (e.g. at $\theta \approx 90°$), the model predicts no deflection, but rather follows the previous direction, effectively implementing a continuity prior. We provide a similar case study, which supports the same conclusions, but for crossing fiber directions, in appendix F.

Influence of Precision

In this experiment, we investigate how the local log-probability profile from Fig. 13b changes as a function of the precision β. For this purpose, we visualize the profile of $\log p^{\text{trk}}_\beta(\mathbf{y}^{in} \mid \mathbf{X}(\mathbf{r}), \mathbf{y}^{in})$ for different values of β in Fig. 14. As expected, the sensitivity of the posterior to the details of the data increases with the precision. More precisely, the dependence of $\log p^{\text{trk}}_\beta\left(\mathbf{R}_\theta \mathbf{e}_x \mid \mathbf{X}(\mathbf{r}), \mathbf{R}_\theta \mathbf{e}_x\right)$ on θ is strongly modulated by the DWI data for high precision β, and tends to be isotropic, i.e. insensitive, for small values of β. This observation illustrates nicely the concept of precision: At low precision, the output distribution is broadened, taking into account only the strongest part of the data signal. On one hand, this smoothing renders the posterior robust to data fluctuations, on the other hand, it suppresses fine details in patterns. It only starts to take into account more details, when we increase the precision. Thus, the posterior will capture higher-order patterns in the data, but it will also be more susceptible to noise.



(a) Concentration $\kappa(\mathbf{X}(\mathbf{r}), \mathbf{y}^{in})$ **(b)** $\log p_\beta^{\text{trk}}(\mathbf{y}^{in} \mid \mathbf{X}(\mathbf{r}), \mathbf{y}^{in})$ **(c)** Deflection of $\boldsymbol{\mu}(\mathbf{X}(\mathbf{r}), \mathbf{y}^{in})$ from $\mathbf{y}^{in}$

Fig. 13 Case study of $p_\beta^{\text{trk}}(\mathbf{y}^{in} \mid \mathbf{X}(\mathbf{r}), \mathbf{y}^{in})$ at $\beta/\beta_{15^\circ} = 1.58$, when $\mathbf{X}(\mathbf{r})$ is a fixed location chosen inside the corpus callosum (*black inset, left*). The predominant FOD is strongly unimodal (*gray dumbbell, center*). Each sub-figure relates to the same FOD, and in each case the polar angle θ indicates the in-plane orientation of $\mathbf{y}^{in} = \mathbf{R}_\theta \mathbf{e}_x$ (Color figure online).

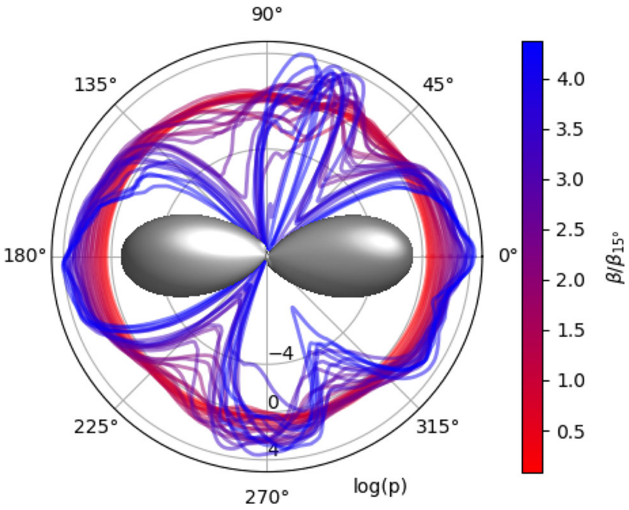


Fig. 14 Log-likelihood $\log p_\beta^{\text{trk}}(\mathbf{y}^{in} \mid \mathbf{X}(\mathbf{r}), \mathbf{y}^{in})$ for low precision β (*red*), and high precision (*blue*). The inset FOD (*gray dumbells, center*) illustrates the fixed input data $\mathbf{X}(\mathbf{r})$. The polar angle θ controls the orientation of $\mathbf{y}^{in} = \mathbf{R}_\theta \mathbf{e}_x$ (Color figure online).

6 Whole-Brain Tractography

In the previous section we studied local properties of the Entrack posterior $p_\beta^{\text{trk}}(\mathbf{y} \mid \mathbf{X}(\mathbf{r}), \mathbf{y}^{in})$. In this section, we focus on its performance in the context of whole-brain tractography, i.e. the iterative tracking procedure of Algorithm 3. In particular, we are interested to determine the optimal value for the precision β. For all whole-brain tractography experiments we use the posterior model p_β^{trk} trained on HCP subject 992774, as described in Sect. 5.2.

Experimental Parameters We describe the concrete experimental parameters required for Algorithm 3 to produce a whole-brain tractogram. Besides the posterior p_β^{trk}, several other, influential factors are involved in the iterative prediction:

- **Interpolation**: The original ISMRM data comes at a voxel size of 2 mm, we up-sample it to the resolution of the HCP data (1.25 mm), using trilinear interpolation.
- **Seeds**: We place one seed point at the center of every voxel inside a white-matter mask, which was thresholded at a value of 0.1.
- **Prior**: We use the main principal axis of a diffusion-tensor fit as initial incoming direction $\mathbf{y}_1$.
  To address the ambiguity about the sign of the prior direction, each streamline is propagated in both directions.
- **Step Size**: We use a step size of 0.25 mm, i.e. 1/5 of the voxel size.
- **Length Constraints**: Fibers are automatically terminated after 800 steps, and we only retain streamlines with a length between 30 mm and 200 mm.
- **Fiber Termination**: Besides termination by length, we also terminate fibers when they arrive at a voxel outside of the white matter mask.
- **Fiber Filtering**: Besides the length restriction, we do not further filter the predicted fibers, e.g. by curvature, etc.

ISMRM15 Phantom and The Tractometer The *Tractometer* (TM) is an evaluation tool for tractography results (Côté et al. 2013; Maier-Hein et al. 2017), and it served also as comparison measure in the ISMRM15 tractography challenge. It is based on a simulated DWI phantom of the brain, which was generated using 25 carefully prepared fiber bundles, which mimic the complex fiber arrangement in the white matter

(Poupon et al. 2010; Neher et al. 2014). A cross-sectional view of these bundles is shown in Fig. 1b.

The TM defines two sets of metrics, which assess on one hand the quality of long-range connectivity, and on the other hand the bundle fidelity of predicted fibers. The first group of metrics includes *valid bundles* (VB), *invalid bundles* (IB), *valid connections* (VC), *invalid connections* (IC), and *non connections* (NC). The second group of metrics includes *mean overlap* (OL), *mean overreach* (OR), and *mean F1 score* (F1). Please refer to appendix D for more details about these evaluation scores.

6.1 Precision Dependence of TM Scores

In Fig. 15, we present the TM metrics as a function of the precision β. As a major observation, the TM scores do not seem to suggest a consistent optimal precision. The VC-score increasingly saturates to a maximum value of 0.52 ($\beta/\beta_{15^\circ} = 1.58$), while the maximum F1-score of 0.54 ($\beta/\beta_{15^\circ} = 0.42$) marginally decreases by 2.5% over the same range. The VB-score toggles between 23 and 24, and remains stable otherwise. The IB-score is the least consistent, with a local minimum of 123 at $\beta/\beta_{15^\circ} = 0.66$, however its total variation over the range of β is also very small $\approx 5\%$. This observation indicates, that besides the lack of a clear optimum with respect to the precision, the TM scores are also not very sensitive to β, except the VC-score, which increases by 20%. The behavior of the VC score rather suggests a comparison with the generalization error, described in Eq. (23), which also saturates for $\beta \to \infty$, and does not have an optimum at finite precision.

6.2 Posterior Concentration and Fractional Anisotropy

A common quantitative measure of how much the diffusion signal is confined along a single direction is the fractional anisotropy FA $\in [0, 1]$: [5] It relates to the eccentricity of the diffusion ellipsoid, and it is 0 for an isotropic sphere, which is characteristic for voxels with ambiguous DWI measurements, whereas it is 1 for voxels with a diffusion ellipsoid clearly elongated in one direction. In Fig. 16, we show that the posterior concentration $\kappa\left(\boldsymbol{x}, \mathbf{y}^{in}\right)$ has indeed a strong correlation with FA$(\boldsymbol{x})$ ($r = 0.42$), which means that there is a strong link between model certainty, and the articularity of the diffusion signal.

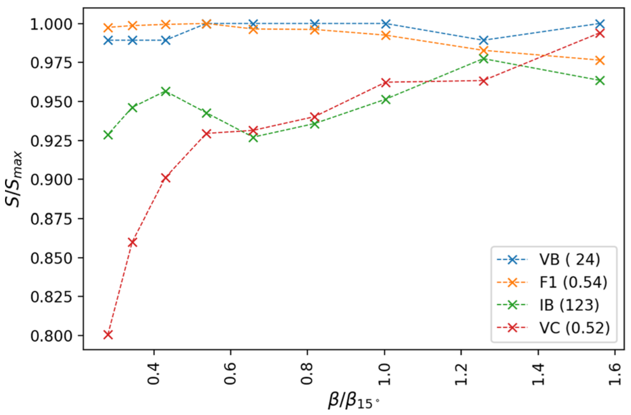


Fig. 15 Tractometer scores S as function of the precision β. Each score is scaled by its maximum S_{max} (*see legend for values*) for better comparability.

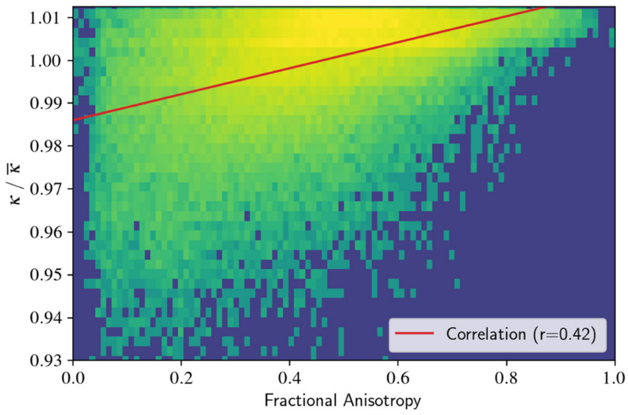


Fig. 16 Joint histogram of the predicted concentrations $\kappa(\boldsymbol{x}, \mathbf{y}^{in})$ (*y-axis, scaled by average* $\overline{\kappa}$), and the corresponding fractional anisotropy FA$(\boldsymbol{x})$ of the diffusion data. The predicted concentrations show a strong correlation with the fractional anisotropy of the diffusion data (r=0.42).

6.3 Comparison to the State-of-the-Art

To provide an absolute reference point for the presented TM-scores, we include an overview of the scores achieved by current tractography solutions based on supervised machine learning in table 1. Besides the row “ISMRM15 $\varnothing$”, which represents the average over all teams of the ISMRM15 challenge (ML and non-ML), we have divided the results into two groups. The first group, marked with an asterisk, represents results where the model was trained on the synthetic DWI phantom, and these data are also used for evaluation. The work of Neher et al. (2017) refers to this setting as *in silico*→*in silico*, meaning that training fibers for these algorithms were obtained on the phantom data by another state-of-the-art algorithm. Even though the training fibers do not exactly correspond to the evaluation fibers, there exists a strong statistical dependence and we can not consider this setting as a valid generalization test. Instead, the model should be trained, and evaluated on two independent instances of (synthetic) DWI data. But as the ISMRM challenge provides only one instance, models should be trained on real DWI data, e.g. from the HCP. As expected, the results in this setting fall

[5] Please refer to appendix C for more details.

Table 1 Tractometer scores on the synthetic ISMRM data set.

Model	VB↑	IB↓	VC↑	OL↑	OR↓	F1↑
ISMRM15 ∅	21	88	54	31	23	44
Neher (2017)	23	94	52	**59**	37	n.a.
Wegmayr (2018)	23	**57**	**72**	16	**28**	n.a.
Entrack (sample)	**24**	123	52	58	39	**54**
Entrack (mean)	**24**	116	54	45	35	47
FvM (sample)	23	154	44	53	36	51
FvM (mean)	23	112	55	48	34	49
Detrack	23	133	43	44	34	48
Classifier (mean)	22	133	36	45	33	46
Poulin (2017)*	23	130	42	64	35	64
Benou (2019)*	**25**	**56**	**71**	**69**	**23**	**70**
Entrack (mean)*	24	126	65	62	36	59
Entrack (sample)*	24	117	65	60	36	58

Up/down-arrows indicate higher/lower-is-better, and best scores for each subgroup (w/*: *in silico*→*in silico*, w/o*: *in vivo*→*in silico*) are shown in bold.

behind compared to the results in the *in silico*→*in silico* setting, which should be considered as (overly) optimistic estimates of the generalization performance. Aside from this criticism, it is apparent that all algorithms fail to consistently outperform the competitors in the realistic *in vivo*→*in silico* setting.

Instead, we can observe a strong trade-off between OL and VC/IB. On one hand, the work of Wegmayr (2018) achieves very good VC/IB (72%/57), but poor OL (16%), on the other hand Entrack (sample) and Neher et al. (2017) achieve much better OL (58% and 59%, respectively), but poorer VC (52% and 52%, respectively). The results for the Entrack model were obtained at $\beta/\beta_{15^\circ} = 1.58$. A similar trade-off is seen between the (sample)/(mean) variants, which refer to how the fiber directions are obtained from the posterior during streamline progression. At each tracking step, the (sample) variant draws a random direction from the posterior, while the (mean) variant always chooses the most likely direction. The Entrack (sample) method achieves superior bundle coverage compared to the (mean) method (OL 58% vs. 45%), but at the cost of more false-positives (IB 123 vs. 116). The same is true for the FvM model, which has the same architecture as the Entrack model, but is trained without entropy regularization, i.e. using the probabilistic regression objective in Eq. (12). The bundle coverage of the FvM (sample) model is also better than its (mean) variant (OL 53% vs. 48%), but also at the cost of more false-negative bundles (IB 154 vs. 112).

Moreover, we highlight that the TM scores support a model ranking Entrack (sample) $\succ$ FvM (sample) $\succ$ Detrack $\succ$ Classifier (mean), which denotes the respective benefits of entropy regularization, a probabilistic loss, and a regression model. The Detrack model has the same neural network architecture as FvM and Entrack, but without the output for κ, and it is trained with the standard negative cosine loss of Eq. (7). The Classifier model also shares the same neural network architecture as all the other models, but it has a softmax output over directions, and is trained by the usual cross-entropy loss for classification. We note, that each of the listed models should be understood as a module in the complete pipeline described by Algorithm 3. Such a pipeline is controlled by various other significant influence factors (training data, seed points, etc.) that are different in each case, thereby limiting comparability. We can only assert that the results of Classifier, Detrack, FvM, and Entrack have been conditioned on the same pipeline, so that their differences can be attributed indeed to the respective choices of the objective functions.

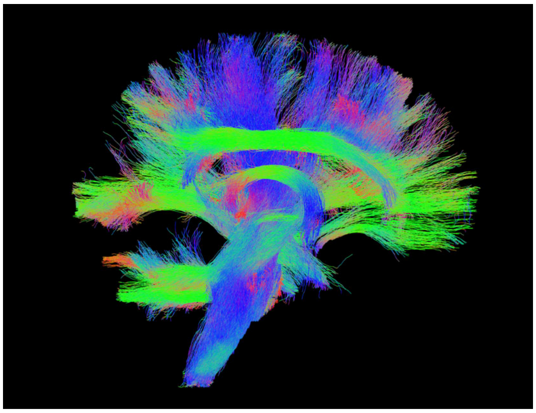


Fig. 17 Whole-brain tractogram predicted by the proposed Entrack model on the synthetic ISMRM data, at a precision of $\beta/\beta_{15^\circ} = 1.58$. For comparison, consider the ground-truth tractogram in Fig. 1b.

6.4 Qualitative Results on ISMRM

In addition to the evaluation metrics provided by the Tractometer, we present qualitative tractogram visualizations. In Fig. 17 we show an overview-section of the whole-brain tractogram obtained with the Entrack model on the ISMRM data, which should be compared to the ground-truth fibers in Fig. 1b.

Additionally, to facilitate a more detailed analysis, we have computed the voxel mask of the predicted corticospinal tract (CST), and visualize its overlap, overreach, and underreach with respect to the ground-truth bundle in Appendix G. Lastly, we demonstrate visualizations of the heteroscedastic uncertainty estimated by the Entrack posterior in Fig. 18. On one hand, we can visualize the spatial dependence of $\kappa\left(\mathbf{X}(\mathbf{r}), \mathbf{y}^{in}\right)$ (Figs. 18a and b), and on the other hand, we can compute per-fiber statistics, such as the log-probability per streamline, i.e.

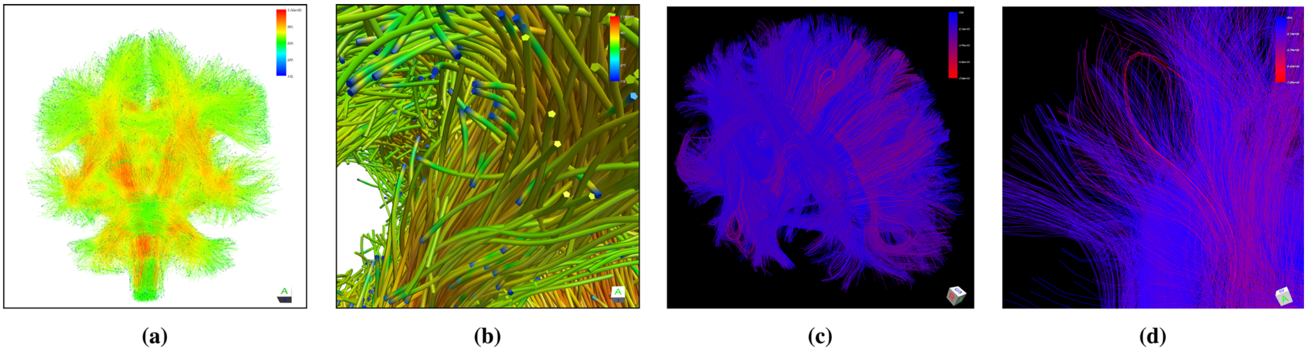


Fig. 18 Per-point, and per-fiber statistics on the ISMRM reference fibers. (**a**) Illustration of the spatial variation of the local certainty κ. Bundle cores tend to be more certain (*red*), while the certainty of their outlines is reduced (*green*). (**b**) Enhanced view of increased uncertainty at fiber ends (*blue*). (**c**) Illustration of fibers with high average log-probability $\overline{\log p}$(*blue*), and low $\overline{\log p}$ (*red*). (**d**) Enhanced view of an implausible loop (*red, center*). Figure reproduced from Wegmayr et al. (2019) (Color figure online).

$$\overline{\log p} = \frac{1}{L}\sum_{j=2}^{L} \log p^{\mathrm{trk}}(\mathbf{y}_j \mid \mathbf{X}(\mathbf{r}_j), \mathbf{y}_{j-1}), \tag{42}$$

shown in Figs. 18c and d. As we have discussed before, the concentration parameter measures the degree of certainty that the model encodes on the fiber direction at a given location. In Fig. 18a we can observe that the concentration/certainty is larger at the core of bundles than in the periphery, which agrees with the fact that the diffusion data is less ambiguous at the bundle cores than at the boundaries. Areas that are closer to the white matter boundary, such as bundle outlines, have lower concentration, because the diffusion signal is weaker in those areas. In particular, fiber end points exhibit the lowest concentrations, as they are located right at the white matter boundary, as shown in Fig. 18b.

In addition to per-point statistics, the per-fiber statistic $\overline{\log p}$ can be used to automatically detect fiber outliers, as shown in Fig. 18c and d. Clearly, without marking fibers in comparison to the average log-likelihood, it is highly error-prone to discover such outliers visually in a tangled whole-brain tractogram. The illustrated implausible loop was found in the ISMRM ground-truth fibers, which are otherwise well prepared. This finding also underlines the difficulty of preparing high-quality reference standards for tractography. Lastly, we note that in contrast to e.g. curvature based outlier-detection, the Entrack model acts as a data-informed filter, i.e. it can recognize fibers, which are strongly bending, but supported by the diffusion data, whereas these fibers would be discarded by a curvature dominated filter.

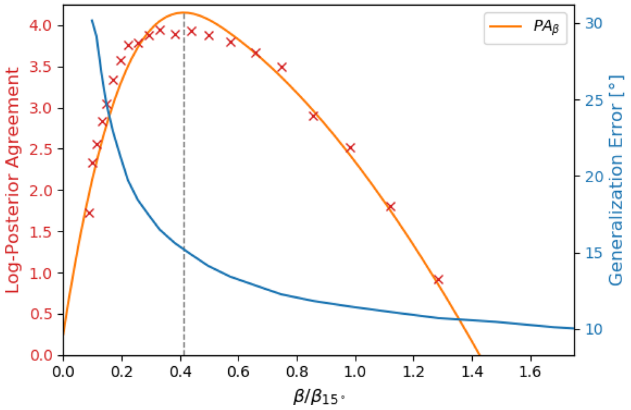


Fig. 19 Average log-posterior agreement (*red crosses*), and generalization error (*blue*). The phenomenological model PA_β (Eq. (43), *orange, solid*) agrees well with the empirical PA, and achieves its maximum value $i_{\hat{\beta}} = 4.14$, at $\hat{\beta}/\beta_{15^\circ} = 0.41$.

6.5 HCP Retest Data and Posterior Agreement

In this section, we show the results for the optimal precision obtained with the PA criterion from Sect. 4.3, based on two independent DWI measurements of the HCP subject 917255. In Fig. 19, we show the measured values of the expected posterior agreement $\mathcal{I}_\beta$ from Eq. (40), and the expected generalization error ρ_β from Eq. (23). In contrast to the Tractometer scores, we observe a clear optimum of $\mathcal{I}_\beta$ at $\hat{\beta}/\beta_{15^\circ} = 0.41$. As anticipated by our discussion in Sect. 3.5, the generalization error ρ_β suggests $\hat{\beta} \to \infty$ and thus fails to provide a finite estimate for the precision.

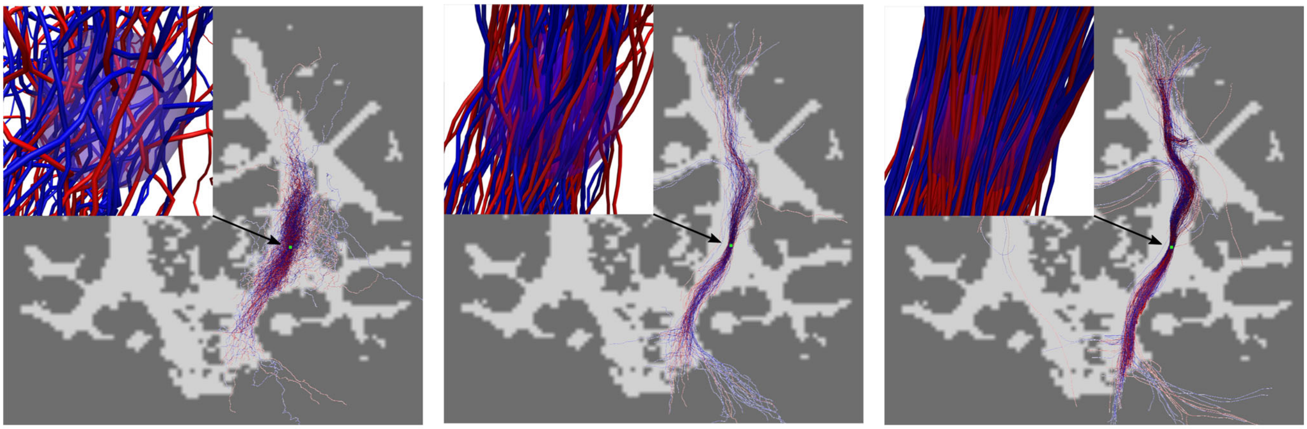
Furthermore, we are interested to explain the empirical PA with a phenomenological model of the form

$$PA_\beta(\bar{\theta}, \bar{n}) = \log_2 4\pi \frac{C(\beta\bar{n}_\beta)^2}{C\left(\beta\bar{n}_\beta\sqrt{2(1+\cos\bar{\theta})}\right)}, \tag{43}$$

which depends only on the summary statistics $\bar{\theta}, \bar{n}$. These are the average number of fibers per fixel

$$\bar{n}_\beta = \frac{W(\beta/\lambda)}{2\sum_{\mathbf{z}}|B_{\mathbf{z}}|}\sum_{\mathbf{z}}\sum_{b\in B_{\mathbf{z}}}\left(n'(\mathbf{z},b) + n''(\mathbf{z},b)\right), \tag{44}$$

where $W(.)$ was introduced in Eq. (4a), and the average deviation between the fixel direction on the two instances:



(a) $\beta/\beta_{15^\circ} = 0.1$ **(b)** $\hat{\beta}/\beta_{15^\circ} = 0.41$ **(c)** $\beta/\beta_{15^\circ} = 1.3$

Fig. 20 Comparison of predicted fibers obtained on two measurements of HCP subject 917255, at different precisions. The fibers predicted on the two measurements are distinguished by color (*red, blue*). Moreover, the shown streamlines all pass through one particular voxel (*black arrow*). The inset shows an enlarged view at the location of this particular voxel in the corticospinal tract (*green dot*).

$$\cos\bar{\theta} := \frac{1}{\sum_{\mathbf{z}}|B_{\mathbf{z}}|}\sum_{\mathbf{z}}\sum_{b\in B_{\mathbf{z}}}\left\langle \boldsymbol{\mu}'_{\beta}(\mathbf{z},b), \boldsymbol{\mu}''_{\beta}(\mathbf{z},b)\right\rangle. \tag{45}$$

This description has one free parameter λ, which essentially captures how fast the local variance of a fiber bundle increases when the precision is decreased. We refer to appendix B for more details about the origin of the parameter λ. It is a joint property of the iterative tracking together with the local posterior, and the DWI data distribution. In particular, it can be considered as a measure of how fast bundles produced by a particular tracking algorithm, on a particular DWI source, tend to disintegrate when the precision is lowered, as illustrated by Fig. 20. In our case, using $\lambda = 10$ provides a good match between the measured PA, and the phenomenological model, as shown by the orange curve in Fig. 19. Its maximum value is $i_{\hat{\beta}} = 4.14$, which means, at the given noise level, we can contract the Entrack posterior up to a concentration, which is equivalent to the partition of the sphere into $\approx 16 = 2^{i_{\hat{\beta}}}$ equally sized cones. A higher resolution can not be argued for, since it would increasingly reduce the agreement between the posteriors on the two measurements. This effect is not captured by the expected generalization error ρ_β (blue curve in Fig. 19), showing again that ρ_β is not an appropriate measure to determine a finite optimal precision, which is necessary to maintain the benefits of probabilistic models.

Qualitative Results on HCP Data

We provide some qualitative tractography examples obtained at the optimal precision $\hat{\beta}$ in Fig. 21. The selected fiber tracts were selected according to examples previously shown in the literature, e.g. in Neher et al. (2017), Poulin et al. (2017).

7 Discussion & Conclusion

We have presented a general probabilistic model for spherical regression based on the Fisher von Mises distribution, with an application to connectomics and the underlying inference of streamlines in white matter of the brain. Our theoretical considerations advocate the model to address loss attenuation, and heteroscedastic uncertainty quantification. For the proposed FvM model, we investigate the issue of probabilistic overfitting in tractography, which is commonly encountered in different probabilistic models, but only addressed by ad-hoc solutions. For instance, Kumar and Tsvetkov (2018) experiment with different regularization terms for the concentration, but it remains unclear which should be recommended in other applications. The classification model for tractography by Benou and Riklin-Raviv (2019) suggests a label smoothing heuristic to assert finite concentrations, and to establish a notion of angular closeness between direction “classes”.

Instead, we advocate a regularization based on the maximum entropy principle. Specifically, we derive the Gibbs free energy for the FvM distribution, and discuss its theoretical properties, in particular the role of the precision parameter β. In contrast to tuning hyper-parameters in regularization heuristics, the meaning of β is clearly motivated as the inverse width of the posterior distribution.

Based on the free energy objective, we also propose an automatically paced annealing scheme for model training, in the spirit of the deterministic annealing algorithm (Hofmann and Buhmann 1997), which is used to find superior global optima of non-convex optimization problems. Apart from the maximum entropy approach, we argue that it is inherently impossible to determine the precision parameter β with com-

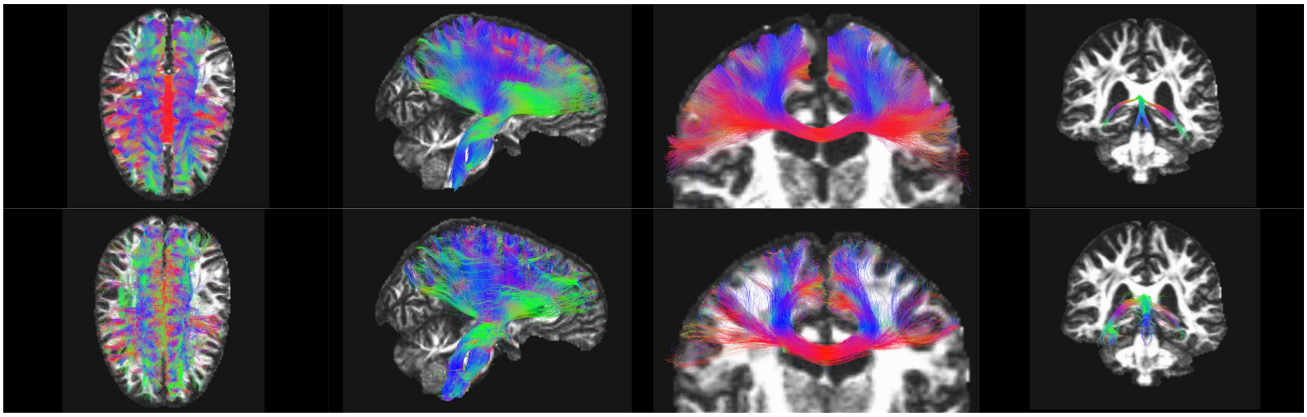


Fig. 21 Selected fiber tracts obtained on HCP data by Entrack at $\hat{\beta}$. *Top row*: TractSeg reference fibers (Wasserthal et al. 2018). *Bottom row*: Fibers predicted by Entrack at $\hat{\beta}$. From left to right: 1. Corpus Callosum, 2. Parieto-Occipital Pontine + Corticospinal + Fronto-Pontine tracts, 3. Frontal Part of Corpus Callosum, 4. Cingulum.

mon cross-validation techniques, since they do not bias the mode of the posterior distribution, but only its width. For this reason we propose a method, which takes into account the stability of the posterior *distribution* with respect to repeated measurements of the data, because the agreement between *normalized* distributions clearly depends on their width.

In the context of our tractography experiments, we refer to the entropy-regularized posterior distribution as *Entrack* model. Firstly, we study its capability of uncertainty quantification with prototypical cases of DWI data, which show that the Entrack model, parametrized by a neural network, has learned non-trivial patterns of streamline progression. Secondly, we employ the Entrack posterior, which describes the distribution of *local* streamline direction, for iterative whole-brain tractography to reconstruct long-range tissue connectivity. On one hand, we show that it produces competitive results in the Tractometer evaluation, which is based on the synthetic ISMRM15 phantom with known ground-truth. In particular, our ablation study shows a progressive improvement of Tractometer scores with respect to the baseline classification model. It is outperformed by deterministic regression, which is moreover improved by its probabilistic formulation, and even more so by the proposed entropy-regularized Entrack model. This model ranking indicates the respective benefits of regression over classification, probabilistic over deterministic, and entropy-regularized statistical inference over the maximum likelihood technique.

However, as expected, the Tractometer evaluation, based on one data instance, does not support to determine a *finite* optimal precision, which is essential to maintain the benefits of probabilistic models. Instead, we show that the posterior agreement, computed from two independent DWI measurements, defines a finite optimal precision, which takes into account the stability of tractograms under data fluctuations. We complete our study with qualitative examples of whole-brain tractography on both, the synthetic ISMRM15 data, and real HCP data.

In summary, the study documents a supervised approach to infer streamlines from DWI data and it validates the results by monitoring the stability of tractograms for repeated DWI measurements. Our modeling strategy generalizes to other data analysis challenges in biomedicine where a gold standard is difficult to establish and standard approaches fail to provide uncertainty calibration in accordance with data noise.

Acknowledgements We thank Amirreza Bahreini for supporting the computational implementation of this work.

Funding Open access funding provided by Swiss Federal Institute of Technology Zurich.

Compliance with Ethical Standards

Conflicts of interest The authors declare that they have no conflict of interest.

A Fixel Posterior

Consider a fixel $(\mathbf{z}, b)$, represented by the average direction $\mathbf{y}^{in}_{\mathbf{z},b} := \mathbf{y}^{in}(\mathbf{z}, b \mid \mathbf{T})$ of $n_{\mathbf{z},b} := n(\mathbf{z}, b \mid \mathbf{T})$ streamlines $\mathbf{y}_j$.

To represent the fixel posterior as a joint posterior over its streamlines, we write

$$p^{\text{trk}}_{joint}\big(\mathbf{y} \mid \mathbf{X}(\mathbf{z}), \mathbf{y}^{in}_{\mathbf{z},b}\big) \propto \prod_{j=1}^{n_{\mathbf{z},b}} p^{\text{trk}}\big(\mathbf{y} \mid \mathbf{X}(\mathbf{z}), \mathbf{y}_j\big). \tag{46}$$

After normalization, the joint concentration can be approximated as

$$\begin{aligned}\kappa^{joint} &= \Big\| \sum_{j=1}^{n_{\mathbf{z},b}} \kappa(\mathbf{X}(\mathbf{z}), \mathbf{y}_j)\boldsymbol{\mu}(\mathbf{X}(\mathbf{z}), \mathbf{y}_j)\Big\|_2 \\ &= \kappa(\mathbf{X}(\mathbf{z}), \mathbf{y}^{in}_{\mathbf{z},b}) \Big\| \sum_{j=1}^{n_{\mathbf{z},b}} \boldsymbol{\mu}(\mathbf{X}(\mathbf{z}), \mathbf{y}_j)\Big\|_2 \\ &= \kappa(\mathbf{X}(\mathbf{z}), \mathbf{y}^{in}_{\mathbf{z},b}) n_{\mathbf{z},b}\end{aligned} \tag{47}$$

where we have assumed in the first step, that the concentration is similar for all fibers of one fixel, and in the second step that the local fiber means are approximately aligned in the same direction.

B Precision-Dependence of $\bar{n}$

To estimate the precision dependence of $\bar{n}$, defined in Eq. (44), we need to estimate the precision dependence of the term $\| \sum_{j=1}^{n_{\mathbf{z},b}} \boldsymbol{\mu}(\mathbf{X}(\mathbf{z}), \mathbf{y}_j)\|_2$ used in the approximation Eq. (47) for the concentration of the fixel posterior.

We make the assumption that the posterior means of the fibers entering some voxel $\mathbf{z}$, and belonging to the same bundle b, are effectively distributed according to the same FvM with concentration β:

$$\boldsymbol{\mu}(\mathbf{X}(\mathbf{z}), \mathbf{y}_j) \sim p^{\text{FvM}}(\kappa = \beta), \tag{48}$$

so that we can approximate the norm of their empirical sum with the norm of their expectation:

$$\Big\| \sum_{j=1}^{n_{\mathbf{z},b}} \boldsymbol{\mu}(\mathbf{X}(\mathbf{z}), \mathbf{y}_j)\Big\|_2 = W(\beta) n_{\mathbf{z},b}. \tag{49}$$

Moreover, we introduce the free parameter λ to arrive at the phenomenological approximation

$$\bar{n}_\beta = W(\beta/\lambda)\bar{n}. \tag{50}$$

C DWI Feature Representations

In this section, we provide some details about commonly used feature representations $\mathbf{X}_f$ of DWI measurements. More details can be found in introductory texts, e.g. by Alexander (2006).

The diffusion tensor (DT) is arguably the most popular feature representation (Basser et al. 1994) for DWI measurements I. It is essentially a Gaussian model of the diffusion signal:

$$f(\mathbf{g}_n \mid D_\mathbf{r}) = I_0 \exp\big(-b\mathbf{g}_n^T \mathbf{D}_\mathbf{r} \mathbf{g}_n\big) \tag{51}$$

where $\mathbf{D}_\mathbf{r}$ is the positive definite, symmetric 3×3 diffusion tensor at location $\mathbf{r}$, b an experimental constant, and I_0 the unattenuated reference intensity.

The DT representation compresses the DWI signal at each voxel from N directions to the three orthogonal principal directions $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \boldsymbol{\epsilon}_3$ of $\mathbf{D}_\mathbf{r}$, and their respective positive eigenvalues $\lambda_1 > \lambda_2 > \lambda_3$. [6]. If we condense these features into one vector, we have that $\mathbf{X}_{DT} : \mathcal{V} \to \mathbb{R}^6$.

The fractional anisotropy (FA) of $\mathbf{D}$ is given by

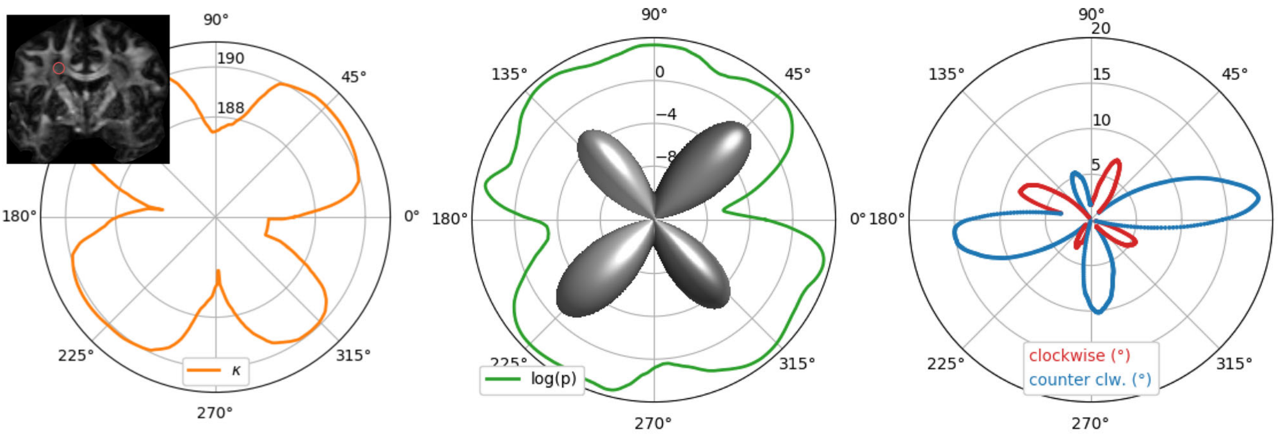
$$FA = \sqrt{\frac{(\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_3 - \lambda_1)^2}{2(\lambda_1^2 + \lambda_2^2 + \lambda_3^2)}}. \tag{52}$$

While the DT representation proves to be fairly robust, it can not properly account for complex fiber configurations, which require a multi-modal representation of directions. For this purpose, an angular expansion in terms of spherical harmonic functions Y_{lm} is commonly used. This representation is also referred to as fiber orientation distribution (FOD):

$$f(\mathbf{g}_n \mid \{D^{lm}_\mathbf{r}\}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} D^{lm}_\mathbf{r} Y_{lm}(\mathbf{g}_n). \tag{53}$$

Due to the inversion symmetry of the DWI signal, the odd coefficients $l = 1, 3, 5, \ldots$ are zero. If we retain coefficients up to $l = 4$, we have $\mathbf{X}_{FOD} : \mathcal{V} \to \mathbb{R}^{15}; \mathbf{r} \mapsto \{D^{lm}_\mathbf{r}\}$.

[6] For notational simplicity, we suppressed the location dependence $\mathbf{r}$ of the eigensystem.



(a) Concentration κ. **(b)** Log-Likelihood of $\mathbf{y}^{in}$. **(c)** Deflection of $\boldsymbol{\mu}$ from $\mathbf{y}^{in}$.

Fig. 22 Case study of a voxel representing a fiber crossing ($\beta/\beta_{15^\circ} = 1.58$).

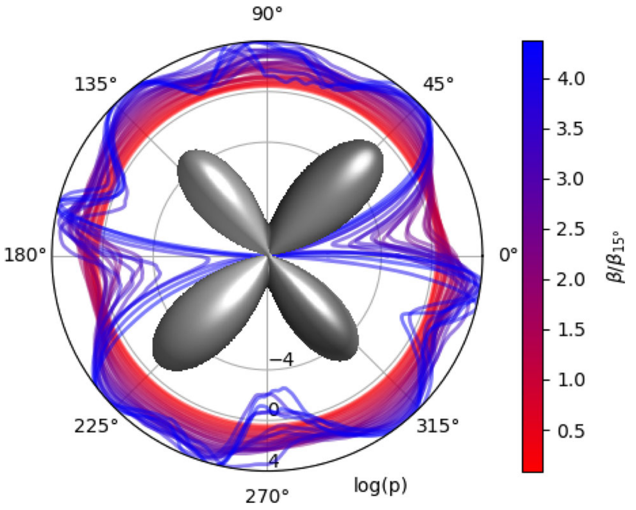


Fig. 23 Log-likelihood $\log p_\beta^{\mathrm{trk}}(\mathbf{y}^{in} \mid \mathbf{X}(\mathbf{r}), \mathbf{y}^{in})$ for low precision β (*red*), and high precision (*blue*). The inset FOD (*gray dumbells, center*) illustrates the fixed input data $\mathbf{X}(\mathbf{r})$. The polar angle θ controls the orientation of $\mathbf{y}^{in} = \mathbf{R}_\theta \mathbf{e}_x$.

D The Tractometer Evaluation

In the first step of the evaluation, the Tractometer tool identifies fibers which connect correct pairs of ground-truth ROIs (VC), fibers which connect incorrect pairs of ROIs (IC), and such that don't connect any pair of ROIs (NC).

Next, IC fibers which are shorter than 35 mm, are also assigned to NC. The VC, IC, and NC metrics simply report the relative size of each set. Furthermore, the sets of VC and IC fibers are clustered independently into bundles of coherent fibers. The number of IC bundles constitutes the IB metric.

In contrast, the identified VC bundles are matched to the 25 ground-truth bundles, and the VB metric reports the number of successful matches. To obtain the bundle fidelity metrics, the identified valid bundles are converted to volumetric binary masks, which are compared to the corresponding ground-truth bundle masks.

The OL metric reports the relative intersection between the predicted bundle mask B, and the corresponding ground-truth bundle mask $\hat{B}$, i.e. $\mathrm{OL} = |B \cap \hat{B}|/|\hat{B}|$. Similarly, the OR metric reports the relative bundle overreach, i.e. $\mathrm{OR} = |B \setminus \hat{B}|/|\hat{B}|$. Lastly, the F1 metric is simply the harmonic mean of OL, and 1-OR.

E FvM-Functions for $d = 3$

In reference to Eq. (4), we provide the explicit formulas for the first moment norm, and entropy of the FvM distribution in three dimensions:

$$C(\kappa) = \kappa/(4\pi \sinh\kappa)\ . \tag{54a}$$

$$W(\kappa) = \coth(\kappa) - \frac{1}{\kappa}\ . \tag{54b}$$

$$H(\kappa) = 1 - \kappa\coth(\kappa) - \log C(\kappa)\ . \tag{54c}$$

F Case Study: Fiber Crossing

In addition to the unimodal case study of the Entrack posterior in Fig. 13, we show a bimodal case study in Figs. 22 and 23.

G ISMRM: CST Bundle Masks

We illustrate volumetric masks of the left CST on the ISMRM phantom in Fig. 24.

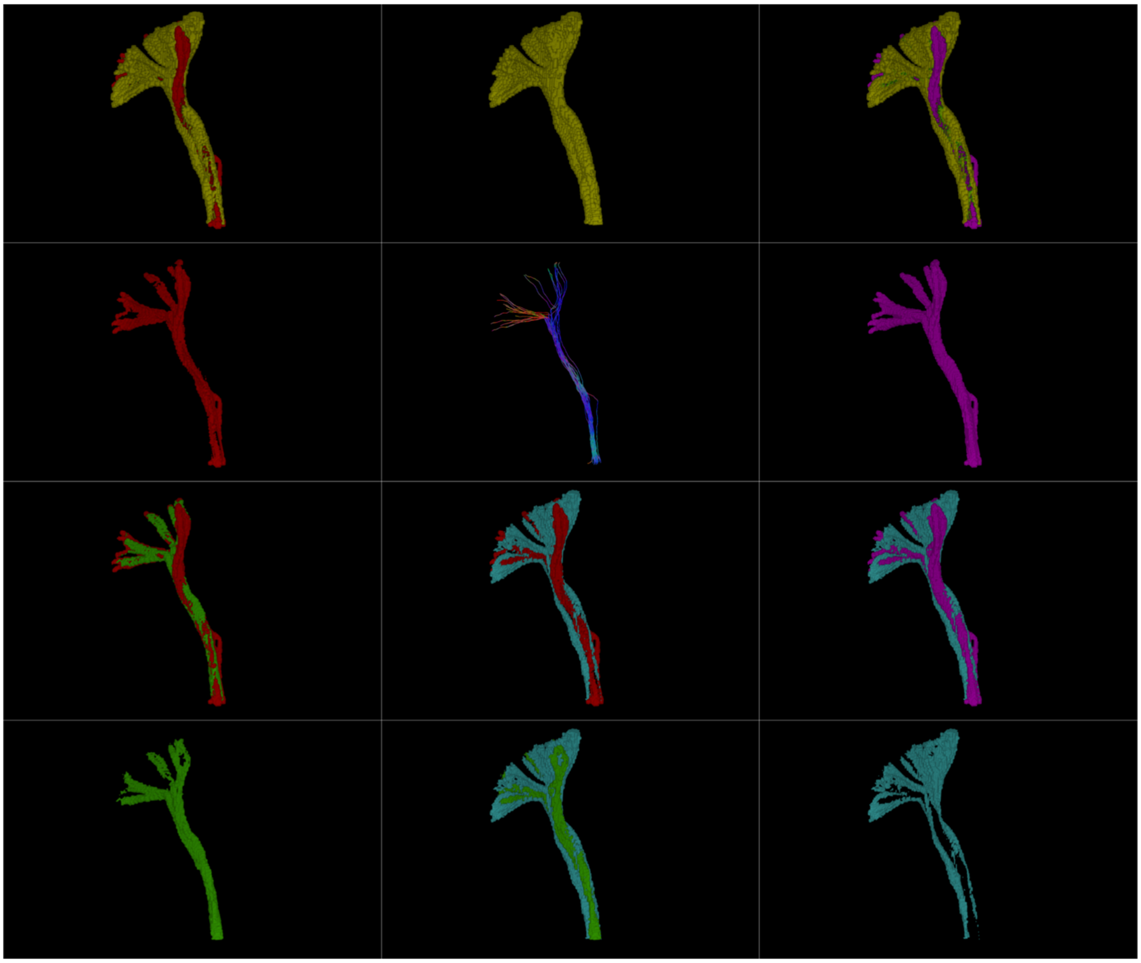


Fig. 24 Bundle-mask analysis of the left CST. The ground-truth mask (*yellow*) is shown in the first row, together with overlays of the predicted voxel mask (*purple*), and its overreach (*red*). In the center of the second row, we illustrate the streamlines, which the predicted mask is based on. The third row contains overlays of the overreach with the overlap (*green*), and the predicted mask with its underreach (*blue*). Underreach refers to voxels covered by the ground-truth bundle, but not by the predicted bundle. The precision is $\beta/\beta_{15^\circ} = 1.58$.

References

Alexander, D. C. (2006). An Introduction to Computational Diffusion MRI: the Diffusion Tensor and Beyond. In *Visualization and Processing of Tensor Fields*.

Bargmann, C. I., & Marder, E. (2013). From the connectome to brain function. *Nature Methods*, *10*, 483–490.

Basser, P., Mattiello, J., & LeBihan, D. (1994). Estimation of the effective self-diffusion tensor from the NMR spin echo. *Journal of Magnetic Resonance Series B*, *103*(3), 247–54.

Basser, P., Pajevic, S., Pierpaoli, C., Duda, J., & Aldroubi, A. (2000). In vivo fiber tractography using DT-MRI data. *Magnetic Resonance in Medicine*, *44*(4), 625–32.

Beaulieu, C. (2002). The basis of anisotropic water diffusion in the nervous system: A technical review. *NMR in Biomedicine*, *15*(7–8), 435–55.

Behrens, T., Woolrich, M., Jenkinson, M., Johansen-Berg, H., Nunes, R., Clare, S., et al. (2003). Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magnetic Resonance in Medicine*, *50*(5), 1077–88.

Benou, I., & Riklin-Raviv, T. (2019). DeepTract: A probabilistic deep learning framework for white matter fiber tractography. In *MICCAI*

Bihan, D. L., & Iima, M. (2015). Diffusion magnetic resonance imaging: What water tells us about biological tissues. *PLoS Biology*, *13*, e1002203.

Buhmann, J. (2010). Information theoretic model validation for clustering. In *2010 IEEE international symposium on information theory* (pp. 1398–1402).

Buhmann, J., Dumazert, J., Gronskiy, A., & Szpankowski, W. (2018). Posterior agreement for large parameter-rich optimization problems. *Theoretical Computer Science*, *745*, 1–22.

Buhmann, J.M. (2013). SIMBAD: Emergence of Pattern Similarity. In: *Similarity-Based Pattern Analysis and Recognition*.

Chehreghani, M.H., Busetto, A.G., & Buhmann, J.M. (2012). Information theoretic model validation for spectral clustering. In: *AISTATS*.

Chilla, G. S. V. N., Tan, C. H., Xu, C., & Poh, C. (2015). Diffusion weighted magnetic resonance imaging and its recent trend-a survey. *Quantitative Imaging in Medicine and Surgery*, *5*(3), 407–22.

Côté, M. A., Girard, G., Boré, A., Garyfallidis, E., Houde, J., & Descoteaux, M. (2013). Tractometer: Towards validation of tractography pipelines. *Medical Image Analysis*, *17*(7), 844–57.

Essen, D., Smith, S., Barch, D., Behrens, T. E. J., Yacoub, E., & Ugurbil, K. (2013). The WU-Minn human connectome project: an overview. *NeuroImage*, *80*, 62–79.

Filley, C., & Fields, R. (2016). White matter and cognition: Making the connection. *Journal of Neurophysiology*, *116*(5), 2093–2104.

Fischer, B., Gorbach, N.S., Bauer, S., Bian, Y., & Buhmann, J.M. (2016). Model Selection for Gaussian Process Regression by Approximation Set Coding. In *GCPR*.

Frank, M., & Buhmann, J. (2011). Selecting the rank of truncated SVD by maximum approximation capacity. In *2011 IEEE international symposium on information theory proceedings* (pp 1036–1040).

Friman, O., Farnebäck, G., & Westin, C. (2006). A Bayesian approach for stochastic white matter tractography. *IEEE Transactions on Medical Imaging*, *25*, 965–978.

Garyfallidis, E., Brett, M., Correia, M., Williams, G. B., & Nimmo-Smith, I. (2012). QuickBundles, a method for tractography simplification. *Frontiers in Neuroscience*, *6*, 175.

Glasser, M., Sotiropoulos, S., Wilson, J., Coalson, T. S., Fischl, B., Andersson, J., et al. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, *80*, 105–124.

Goodfellow, I., Bengio, Y., & Courville, A. C. (2015). Deep learning. *Nature*, *521*, 436–444.

Gorbach, N. S., Tittgemeyer, M., & Buhmann, J. (2018). Pipeline validation for connectivity-based cortex parcellation. *NeuroImage*, *181*, 219–234.

Hauberg, S., Schober, M., Liptrot, M.G., Hennig, P., & Feragen, A. (2015). A random riemannian metric for probabilistic shortest-path tractography. In *MICCAI*.

Hofmann, T., & Buhmann, J. (1997). Pairwise Data Clustering by Deterministic Annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*, 1–14.

Jaynes, E. (1957). Information Theory and Statistical Mechanics. *Physical Review*, *106*, 620–630.

Jbabdi, S., & Johansen-Berg, H. (2011). Tractography: Where do we go from here? *Brain Connectivity*, *1*(3), 169–83.

Jeurissen, B., Leemans, A., Tournier, J., Jones, D., & Sijbers, J. (2013). Investigating the prevalence of complex fiber configurations in white matter tissue with diffusion magnetic resonance imaging. *Human Brain Mapping*, *34*(11), 2747–66.

Jeurissen, B., Descoteaux, M., Mori, S., & Leemans, A. (2019). Diffusion MRI fiber tractography of the brain. *NMR Biomedicine*, *32*(4), e3785.

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS*.

Kingma, D.P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *ICLR*.

Kiureghian, A., & Ditlevsen, O. D. (2009). Aleatory or epistemic? Does it matter? *Structural Safety*, *31*, 105–112.

Kumar, S., & Tsvetkov, Y. (2018). Von Mises-Fisher Loss for Training Sequence to Sequence Models with Continuous Outputs. *ICLR*.

Maier-Hein, K., Neher, P., Houde, J., Côté, M. A., Garyfallidis, E., Zhong, J., et al. (2017). The challenge of mapping the human connectome based on diffusion tractography. *Nature Communications*, *8*, 13.

Mardia, K., & Jupp, P. (2000). *Directional statistics*., Wiley series in probability and statistics Hoboken: Wiley.

Mardia, K.V. (1975). Characterizations of directional distributions. In *Statistical Distributions in Scientific Work*

Neher, P., Laun, F., Stieltjes, B., & Maier-Hein, K. (2014). Fiberfox: Facilitating the creation of realistic white matter software phantoms. *Magnetic Resonance in Medicine*, *72*(5), 1460–70.

Neher, P., Côté, M. A., Houde, J., Descoteaux, M., & Maier-Hein, K. (2017). Fiber tractography using machine learning. *NeuroImage*, *158*, 417–429.

Nimsky, C., Bauer, M., & Carl, B. (2016). Merits and Limits of Tractography Techniques for the Uninitiated. *Advances and Technical Standards in Neurosurgery*, *43*, 37–60.

Oishi, K., Mielke, M., Albert, M., Lyketsos, C., & Mori, S. (2011). DTI analyses and clinical applications in Alzheimer's disease. *Journal of Alzheimer's disease: JAD*, *26*(Suppl 3), 287–96.

Poulin, P., Côté, M.A., Houde, J.C., Petit, L., Neher, P.F., Maier-Hein, K.H., Larochelle, H., & Descoteaux, M. (2017). Learn to track: deep learning for tractography. In *MICCAI*.

Poulin, P., Jörgens, D., Jodoin, P. M., & Descoteaux, M. (2019). Tractography and machine learning: Current state and open challenges. *Magnetic Resonance Imaging*, *64*, 37–48.

Poupon, C., Laribiere, L., Tournier, G., Bernard, J., Fournier, D., Fillard, P., Descoteaux, M., & Mangin, J.F. (2010). A diffusion hardware phantom looking like a coronal brain slice. In *ISMRM 18th Scientific Meeting and Exhibition, Stockholm, Sweden*

Prokudin, S., Gehler, P.V., & Nowozin, S. (2018). Deep Directional Statistics: Pose Estimation with Uncertainty Quantification. *ECCV*.

Raffelt, D., Tournier, J., Smith, R., Vaughan, D. N., Jackson, G., Ridgway, G., et al. (2017). Investigating white matter fibre density and morphology using fixel-based analysis. *Neuroimage*, *144*, 58–73.

Reisert, M., Mader, I., Anastasopoulos, C., Weigel, M., Schnell, S., & Kiselev, V. (2011). Global fiber reconstruction becomes practical. *NeuroImage*, *54*, 955–962.

Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. In *NIPS*.

Soares, J., Marques, P., Alves, V., & Sousa, N. (2013). A hitchhiker's guide to diffusion tensor imaging. *Frontiers in Neuroscience*, *7*, 31.

Wasserthal, J., Neher, P., & Maier-Hein, K. (2018). TractSeg - Fast and accurate white matter tract segmentation. *Neuroimage*, *183*, 239–253.

Wegmayr, V. (2018). Data-driven fiber tractography with neural networks. *ISBI* pp 1030–1033.

Wegmayr, V., Giuliari, G., & Buhmann, J.M. (2019). Entrack: A data-driven maximum-entropy approach to fiber tractography. In *GCPR*.

Williams, T. H., Gluhbegovic, N., & Jew, J. Y. (1997). The virtual hospital. http://163.178.103.176/Temas/Temab2N/APortal/FisoNerCG/LaUII/Neuro/BrainAn/Ch5Text/Section12.html, Accessed 24 Nov 2019.

Yamada, K., Sakai, K., Akazawa, K., Yuen, S., & Nishimura, T. (2009). MR tractography: A review of its clinical applications. *Magnetic Resonance in Medical Sciences : MRMS : An Official Journal of Japan Society of Magnetic Resonance in Medicine*, *8*(4), 165–74.

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