Uncertainty mapping and probabilistic tractography using Simulation-Based Inference in diffusion MRI: A comparison with classical Bayes

J.P. Manzano-Patron[1], **Michael Deistler**[2], **Cornelius Schröder**[2], **Theodore Kypraios**[3], **Pedro J. Gonçalves**[4,5], **Jakob H. Macke**[2,6], **Stamatios S.N. Sotiropoulos**[1]

[1]Sir Peter Mansfield Imaging Centre (SMPIC), University of Nottingham, UK
[2]Machine Learning in Science, Excellence Cluster Machine Learning, University of Tübingen & Tübingen AI Center, Germany
[3]School of Mathematical Sciences, University of Nottingham, UK
[4]VIB-Neuroelectronics Research Flanders (NERF), Belgium
[5]Department of Computer Science, Department of Electrical Engineering, KU Leuven, Belgium
[6]Department of Empirical Inference, Max Planck Institute for Intelligent Systems, Germany

Stamatios.Sotiropoulos@nottingham.ac.uk

ABSTRACT

Simulation-Based Inference (SBI) has recently emerged as a powerful framework for Bayesian inference: Neural networks are trained on simulations from a forward model, and learn to rapidly estimate posterior distributions. We here present an SBI framework for parametric spherical deconvolution of diffusion MRI data of the brain. We demonstrate its utility for estimating white matter fibre orientations, mapping uncertainty of voxel-based estimates and performing probabilistic tractography by spatially propagating fibre orientation uncertainty. We conduct an extensive comparison against established Bayesian methods based on Markov-Chain Monte-Carlo (MCMC) and find that: a) in-silico training can lead to calibrated SBI networks with accurate parameter estimates and uncertainty mapping for both single and multi-shell diffusion MRI, b) SBI allows amortised inference of the posterior distribution of model parameters given unseen observations, which is orders of magnitude faster than MCMC, c) SBI-based tractography yields reconstructions that have a high level of agreement with their MCMC-based counterparts, equal to or higher than scan-rescan reproducibility of estimates. We further demonstrate how SBI design considerations (such as dealing with noise, defining priors and handling model selection) can affect performance, allowing us to identify optimal practices. Taken together, our results show that SBI provides a powerful alternative to classical Bayesian inference approaches for fast and accurate model estimation and uncertainty mapping in MRI.

Keywords Bayesian inference, dMRI, Artificial Neural Networks, Fibre Orientations, Markov-Chain Monte-Carlo, Ball & Sticks, Parametric Deconvolution

1 Introduction

Identifying the parameters of a model that best represent a set of observations is an omnipresent task in science. For Magnetic Resonance Imaging (MRI) in particular, this is a key task, as on many occasions biophysical quantities of interest are indirectly inferred from measurements. In diffusion MRI (dMRI) of the brain, measurements of the diffusion scatter pattern of water molecules within tissues can be mapped through models to neuronal tissue microstructure and brain connectivity Novikov (2021); Alexander et al. (2019); Jbabdi et al. (2015).

Deterministic approaches, from non-linear optimisation to machine learning methods (see Karimi and Warfield (2024) for a review), can be used to fit such models to data, providing *point estimates* of model parameters. However, these

estimates are affected by different sources of uncertainty, from noise in the observations (aleatoric uncertainty), to preprocessing errors, modelling assumptions, and incomplete sampling (epistemic uncertainty). Mapping uncertainty provides a principled manner for assessing confidence in the estimates Jones (2003), separating noise and signal, identifying covariance structures in the parameters and aiding experimental design Alexander (2008). In dMRI, uncertainty in fibre orientations has been the basis for probabilistic tractography Behrens et al. (2007); Kaden et al. (2007); Yendiki et al. (2011); Girard et al. (2024), which estimates spatial distributions of white matter bundles.

Classical approaches for mapping uncertainty in dMRI include bootstrapping and Bayesian inference. Bootstrapping Pajevic and Basser (2003) involves iteratively fitting the same model to randomly resampled (or perturbed) versions of the original dataset. This provides an empirical distribution for each model parameter, whose width denotes the respective estimation uncertainty Jones (2008); Bernstein et al. (2019); Gu et al. (2019); Ning et al. (2021), but can be sensitive to how resampling is performed Whitcher et al. (2008); Kauermann et al. (2009). On the other hand, Bayesian approaches offer a principled framework for estimating the posterior distribution of the model parameters given the data, providing measures of confidence and covariance in the estimates. For instance, Variational Bayes Chappell et al. (2009); Kaden et al. (2008); Blei et al. (2017), or Markov Chain Monte Carlo (MCMC) sampling Behrens et al. (2003); Kaden and Kruggel (2012); Sotiropoulos et al. (2013) provide flexibility in modelling both signal and noise components, and make it possible to explicitly incorporate prior knowledge. These approaches have two main limitations: 1) they require the likelihood of the model to be evaluated analytically, which is not always available or it is too expensive to compute; and 2) although great speed-ups have been achieved (e.g. utilising GPUs Hernandez-Fernandez et al. (2019) and/or improving the algorithmic design Manzano Patron (2023); Harms et al. (2024)), they are highly iterative in nature, which can result in substantial computational complexity and limited scalability.

Simulation-based inference (SBI) Cranmer et al. (2020) has been developed to tackle such limitations in Bayesian inference. The basic idea is that artificial neural networks are trained on simulations from a forward model, in a way that allows them to learn Bayesian posterior distributions. Importantly, this approach can *amortise* inference: After initial training, performing inference on unseen data only requires a single forward pass through the pretrained neural network, without the need for costly iterative sampling and likelihood evaluations. This amortisation property is particularly effective in scenarios where the same problem needs to be solved many times, as in the voxel-wise inference done in microstructural models of dMRI.

In this work, we demonstrate how SBI can be used for mapping orientations and uncertainty in dMRI and, for the first time, all the way up to probabilistic tractography. We comprehensively compare SBI with a classical MCMC-based and heavily-used implementation (FSL-bedpostX) Behrens et al. (2007); Jbabdi et al. (2012); Hernández et al. (2013), with respect to both voxel-wise fibre orientation uncertainty estimates and whole-brain white matter tract reconstructions.

1.1 Related work

In Bayesian inference, the posterior distribution $p(\mathbf{\Omega}|Y_{obs})$ of model parameters $\mathbf{\Omega}$ given the measured data Y_{obs} is proportional to the product of the prior $p(\mathbf{\Omega})$ and the likelihood function $p(Y_{obs}|\mathbf{\Omega})$, the conditional distribution of observations given a model. For problems where the likelihood is intractable or expensive to compute Tavaré et al. (1997); Beaumont (2010), Approximate Bayesian Computation (ABC) or Likelihood-Free Inference (LFI) approaches have been proposed Sisson et al. (2018); Beaumont et al. (2002). In ABC, the posterior is approximated by simulating large amounts of data Y_{sim} from a forward stochastic simulator $f_Y(\mathbf{\Omega})$ (synthetic likelihood), with parameters $\mathbf{\Omega}$ randomly drawn from the defined priors $p(\mathbf{\Omega})$. Samples for which a distance $\rho(Y_{sim}, Y_{obs})$ is below a pre-defined threshold are accepted as posterior samples (otherwise rejected). SBI recasts this as a density estimation problem, where neural networks are used to learn a mapping between observations and the posterior distribution Beaumont et al. (2002); Blum and François (2009). This avoids inefficient sampling-rejection schemes and allows problems with high-dimensional observations to be tackled. Such networks, known as Neural Density Estimators (NDEs) Magdon-Ismail and Atiya (1998), have enabled inference in a purely data-driven fashion on different target (conditional) densities, such as the posterior distribution of model parameters (Neural Posterior Estimation or NPE) pap (2016); Lueckmann et al. (2017); Papamakarios et al. (2017); Greenberg et al. (2019), a surrogate likelihood (Neural Likelihood Estimation) Papamakarios et al. (2019); Durkan et al. (2018), or variants based on denoising diffusion models Sharrock et al. (2022); Geffner et al. (2023); Wildberger et al. (2024); Gloeckler et al. (2024).

A few recent studies have explored inferring parameters from dMRI models using such approaches, demonstrating how SBI can assist in fitting models that are typically challenging to invert. Jallais et al. (2022) fitted the SANDI model of grey matter Palombo et al. (2020) to direction-averaged dMRI signal, using the width of the SBI-estimated posteriors to identify regions where the model presented indeterminacies or biases. A generalisation of this work was presented in Jallais and Palombo (2024), proposing a toolbox applicable to a broader range of models with examples of fitting to direction-averaged dMRI signals. In Eggl and Santis (2024), the authors used SBI mean posterior to demonstrate that it can provide robust point-estimates with fewer observations than deterministic non-linear fitting. In another

recent work, Karimi et al. (2024) presented a likelihood-free framework to obtain the full posterior of parameters in a multi-tensor model. An extension of this work for the standard model of white matter Novikov et al. (2018) was presented in Consagra et al. (2024), demonstrating the ability to obtain fiber-specific microstructural features.

However, despite the great potential revealed for SBI from these studies, a comprehensive comparison to classical Bayesian approaches, which have been used in dMRI for more than two decades, is still missing. In particular, Bayesian frameworks for mapping orientation uncertainty and performing uncertainty-based probabilistic tractography Behrens et al. (2003, 2007); Jbabdi et al. (2012); Hernandez-Fernandez et al. (2019) have been collectively used in thousands of studies before Girard et al. (2024). Therefore, they naturally provide a "gold-standard" against which SBI approaches can be calibrated, optimised and validated.

1.2 Contributions

In this study we close this gap, expanding on our previous preliminary work Manzano-Patron et al. (2022) and complementing recent work by others Jallais et al. (2022); Jallais and Palombo (2024); Karimi et al. (2024); Consagra et al. (2024). We devise an SBI framework for mapping fibre orientation uncertainty from diffusion MRI data using parametric spherical deconvolution models. We subsequently use the SBI-derived voxel-wise posterior distributions to perform probabilistic tractography for the first time. Given the inherent difficulty in validating uncertainty estimates, we comprehensively evaluate the accuracy and precision of the estimates against an established MCMC-based classical inference approach, crucially using the same models and priors, as implemented in FSL-bedpostx Behrens et al. (2007); Jbabdi et al. (2012); Hernández et al. (2013). We evaluate SBI both for voxel-wise and whole-brain level inference, exploring how estimated uncertainty is spatially propagated.

We find that in-silico training can lead to SBI networks that yield accurate mean parameter estimates and uncertainty for unseen observations in both single- and multi-shell dMRI acquisitions, while offering orders of magnitude computational speed-ups for inference. We also show that SBI-based probabilistic tractography for a representative set of more than 40 white matter bundles exhibits a high level of agreement with MCMC-based counterparts; and that this level of agreement is within the range of individual scan-rescan reproducibility of estimates. We further identify optimal SBI design features (such as how to account for noise, how to restrict and constraint priors and how to deal with model selection) that improve training, increase performance and similarity to MCMC estimates. Our findings augment recent studies suggesting that SBI provides new and exciting opportunities for amortised inference in MRI models.

2 Methods

2.1 NPE theory and network architecture

In classical Bayesian inference (like MCMC, see Fig. 1), given a signal and noise model, a likelihood function $p(Y_{obs}|\mathbf{\Omega})$ can be defined. Any prior knowledge of the parameters $\mathbf{\Omega}$ can be embedded in the prior $p(\mathbf{\Omega})$ and the two can be used to estimate the posterior distribution of the parameters given the observed data, i.e. $p(\mathbf{\Omega}|Y_{obs}) \sim p(Y_{obs}|\mathbf{\Omega})p(\mathbf{\Omega})$. This typically necessitates highly iterative processes and potentially expensive likelihood calculations. Furthermore, parameter estimation given a dataset starts from scratch for new observations Y_{obs} (e.g. data for different voxels).

Within SBI, this problem can be addressed using Neural Posterior Estimation (NPE) pap (2016); Lueckmann et al. (2017); Greenberg et al. (2019). In NPE, the mapping between observations and posterior is learned once using a training dataset and can be applied on new unseen data without any expensive calculations, a property known as *amortised inference*. Being able to evaluate the likelihood function is not necessary as long as we have a forward process $f_Y(\mathbf{\Omega})$ to generate simulated data Y. As in all Bayesian inference schemes, we can code prior knowledge into our prior distribution $p(\mathbf{\Omega})$ to define what regions of the parameter space are sampled for such forward predictions (or *simulations*). NPE approximates the target posterior using a parameterised conditional density distribution $q_\Phi(\mathbf{\Omega}|Y)$ represented by neural networks with parameters Φ. Hence, having a training set of pairs (Ω_i, Y_i), a NPE is trained to learn the mapping function $q_\Phi(\mathbf{\Omega}|Y) \approx p(\mathbf{\Omega}|Y)$. After training, we can use $q_\Phi(\mathbf{\Omega}|Y)$ to obtain samples from the (approximate) posterior $p(\mathbf{\Omega}|Y_{obs})$ with a single forward pass through the network.

More formally, we minimise the expected Kullback-Leibler divergence between true posterior distribution $p(\mathbf{\Omega}|Y)$ and the approximation $q_\Phi(\mathbf{\Omega}|Y)$ pap (2016); Papamakarios et al. (2021):

$$\min_{\Phi} \mathcal{L}(\Phi) = \mathbb{E}_{y\sim p(Y)} \left[D_{\mathrm{KL}} \left(p(\mathbf{\Omega}|Y) \parallel q_\Phi(\mathbf{\Omega}|Y) \right) \right]. \tag{1}$$

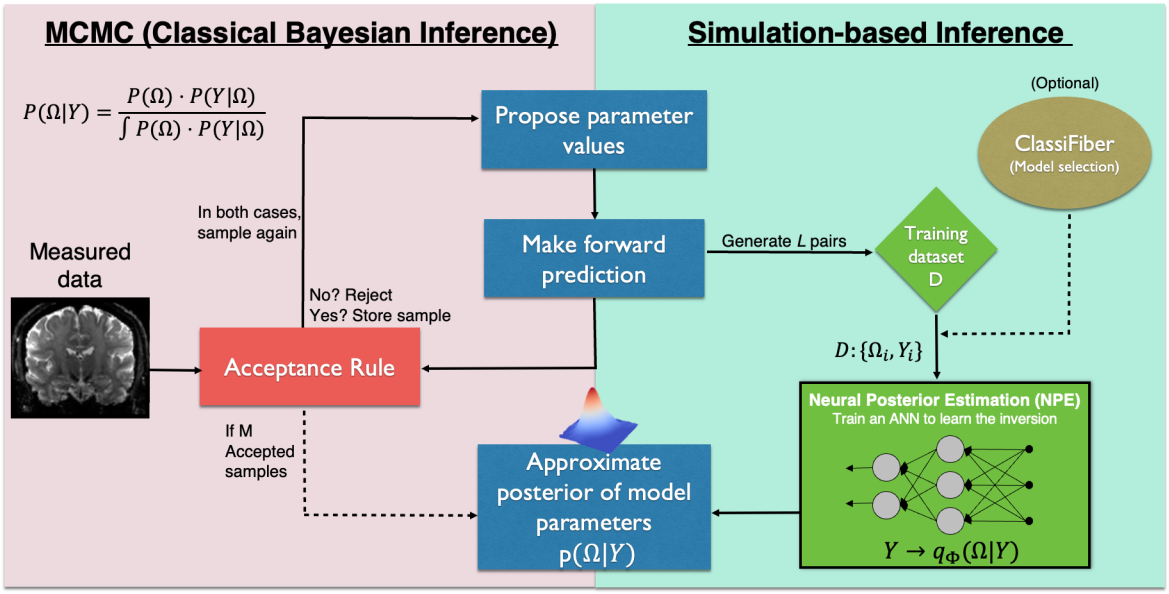

Figure 1: **Schematic overview of Markov-Chain Monte-Carlo (MCMC) (an example of Classical Bayesian Inference) and Simulation-Based Inference (SBI)**. For NPE, an artificial neural network (ANN) is trained on a set $\{\boldsymbol{\Omega_i}, \boldsymbol{Y_i}\}$ obtained from the forward simulator, to learn the latent mapping q_Φ between observations and posterior samples.

This is equivalent to minimise the negative log-likelihood of the posterior approximation Greenberg et al. (2019); Papamakarios et al. (2021):

$$\min_{\Phi} \mathcal{L}(\Phi) \approx -\frac{1}{L} \sum_{i=1}^{L} \log \left(q_\Phi(\Omega_i | Y_i) \right), \tag{2}$$

where L denotes the number of pair-vectors $(\boldsymbol{\Omega}_i, Y_i)$ used to train the networks, where $\boldsymbol{\Omega} \sim p(\boldsymbol{\Omega})$ and $Y_i \sim f_Y(\boldsymbol{\Omega}_i)$. During training, the parameters Φ of the network are optimised to minimise the loss in eq. 2.

Different model choices have been made to parameterise the posterior distribution $q_\Phi(\boldsymbol{\Omega}|Y)$ in previous applications to dMRI, such as Mixture Density Networks (MDNs) Karimi et al. (2024); Consagra et al. (2024) and Masked Autoencoders (MADE) Jallais et al. (2022); Jallais and Palombo (2024). In our work, we use Normalising Flows (NFs) as the basis structure for the density estimators given their flexibility and superior performance in other applications Lueckmann et al. (2021); Dockhorn et al. (2020); Manzano-Patron et al. (2022). NFs can produce complex probability distributions by transforming a base distribution through a series of invertible transformations f_k. A key features of NFs is their use of differentiable closed-form functions that support efficient computation of both the transformations and their Jacobian determinants, ensuring rapid density estimation and sampling. In this paper, we use Neural Spline Flows (NSF) Durkan et al. (2019), as implemented in the SBI package Tejero-Cantero et al. (2020). NSFs comprise of K rational-quadratic splines, each parameterised by a neural network.

2.2 Forward model $f_Y(\boldsymbol{\Omega})$

In this work, we use variants of the Ball & Sticks model Behrens et al. (2007); Jbabdi et al. (2012) as the forward simulator $f_Y(\boldsymbol{\Omega})$. Our goal is to infer the posterior of parameters $\boldsymbol{\Omega}$ from dMRI data in each voxel, with $Y_{obs} = \{Y_{1,obs}, \ldots, Y_{j,obs}\}$ of $j = 1, \ldots, J$ signals, each corresponding to a different diffusion-sensitising gradient. This model has been heavily used for performing parametric spherical deconvolution and uncertainty-based probabilistic tractography over the past 20 years. An MCMC-based implementation for classical Bayesian inference on this model has been developed by members of our team and is part of FSL Behrens et al. (2007); Hernández et al. (2013); Hernandez-Fernandez et al. (2019) and will be used for direct comparisons with SBI.

The forward model is multi-compartment, comprising of a weighted sum of N anisotropic compartments (sticks), representing distinct white matter fibre compartments, and an isotropic compartment (ball), representing partial volume and non-axonal contributions to the signal. The voxel-wise dMRI signal attenuation A in single-shell data (with

relatively low b-value) is described as:

$$f_Y(\mathbf{\Omega}) \rightarrow A_j = S_j/S_0 = \left(1 - \sum_{n=1}^{N} f_n\right) e^{-b_j d} + \sum_{n=1}^{N} f_n e^{-b_j d(\boldsymbol{g}_j \cdot \mathbf{v}_n^T)^2}, \tag{3}$$

where S_j and S_0 are the signals measured with and without diffusion-sensitising gradients, respectively, with b_j value and orientation $\boldsymbol{g}_j$ (protocol-defined parameters). The unknown model parameters to estimate are the mean diffusivity d, the vector $\mathbf{v}_n$ describing the nth fibre orientation, represented in spherical coordinates by $[\theta_n, \phi_n]$ (θ is the angle from the positive z-axis and ϕ is the angle of the projection on the x-y plane from the positive x-axis) and the volume fraction of the nth fibre f_n (with $\sum f_n \leq 1$). The maximum number of fibres N per voxel is typically set to $N = 3$, but the appropriate N per voxel needs to be determined (model selection challenge).

The above model can be extended for multi-shell data Jbabdi et al. (2012) as:

$$f_Y(\mathbf{\Omega}) \rightarrow A_j = S_j/S_0 = \sum_{n=1}^{N} f_n \exp\left(\frac{d^2}{d_{std}} \log\left(\frac{d}{d + b_j(\mathbf{g}_j \cdot \mathbf{v}_n^T)^2 d_{std}^2}\right)\right) + \left(1 - \sum_{n=1}^{N} f_n\right) \exp\left(\frac{d^2}{d_{std}} \log\left(\frac{d}{d + b_j d_{std}^2}\right)\right), \tag{4}$$

where d and d_{std} denote the mode and spread of a Gamma distribution of diffusivities that can account for non-monoexponential signal decay. In summary, the vector of model parameters when the maximum number of fibres is n, $1 \leq n \leq N$. is given by $\mathbf{\Omega} = \{d, d_{std}, \{f_1, \ldots, f_n, \theta_1, \ldots, \theta_n, \phi_1, \ldots, \phi_n\}\}$.

2.3 SBI design and implementation

We designed and implemented a number of NPE networks for the forward models described in the previous section (setting N up to 3), using the Python SBI toolbox (v.0.22) Tejero-Cantero et al. (2020). We set the depth of the Neural Spline Flows to $K = 10$. For training and inference, we used the attenuation A_j as input for the networks, obtained after data pre-processing by dividing the diffusion-weighted signal by the mean of the $b = 0$ volumes (as normalised inputs generally provide more stable behaviour on neural networks). Depending on model complexity and number of compartments considered, different training sizes were used for different SBI architectures, but in general training sample sizes ranged from 3M to 5M samples. Once the networks were trained, we drew $M = 50$ samples from $q_\Phi(\Omega|Y_{obs})$ to characterise the posterior given unseen observations. Code for the implemented architectures can be found on `https://github.com/orgs/SPMIC-UoN/SBI_dMRI`.

We performed extensive comparisons between posterior estimates obtained from SBI and from MCMC (as available in FSL-bedpostX implementation (v.6.0.7) Hernandez-Fernandez et al. (2019)) on both single and multi-shell models. These included comparisons using both synthetic and in-vivo data, at a voxel (e.g., fibre orientations) and whole brain scale (tractography reconstructions). The equations and parameters underlying the MCMC implementation are provided in Supplementary Material for completeness, as well as the metrics we used for assessing accuracy and precision of parameter estimates.

We also investigated how SBI design choices affect accuracy and precision, and agreement with MCMC. In the following sections we present novel explorations on the training design for handling priors, treating noise and addressing model selection challenges. We finally present the data used and data processing aspects all the way to probabilistic tractography.

2.3.1 Uniform vs restricted priors $p(\mathbf{\Omega})$

Non-informative priors have been used in recent dMRI SBI implementations Jallais et al. (2022); Jallais and Palombo (2024); Karimi et al. (2024); Consagra et al. (2024), as they provide the simplest choice for training and offer uniform coverage over the whole parameter space. However, certain combinations of parameters will never be found in real-world cases or won't be feasible to detect due to limitations in the measurement process. Furthermore, certain model parameters have non-linear boundary conditions that cannot be easily defined by parametric prior distributions, requiring ad-hoc reparametrisation that often are not trivial Jallais and Palombo (2024). For instance, the sum of volume fractions in a multi-compartment model should be between 0 and 1. Using non-informative priors and forcing NPE to learn from implausible regions in parameter space could make training highly inefficient, potentially reduce the predictive performance, and increase estimation uncertainty.

We tested two SBI designs, one with uniform priors and another with restricted priors. While the former is straightforward to implement, the latter requires a more advanced design Deistler et al. (2022b,a). For the uniform prior implementation we used: $p(d) \sim \mathcal{U}(10^{-5}, 5 \cdot 10^{-3})$, $p(d_{std}) \sim \mathcal{U}(0, 5 \cdot 10^{-3})$, and $p(f_n) \sim \mathcal{U}(0, 1)$. The fibre orientations were sampled in the positive hemisphere (i.e. $\theta \in [0, \pi/2], \phi \in [0, \pi]$) to avoid degenerate solutions due to antipodal symmetry. To ensure uniform sampling on the sphere, the following correction was applied: $\theta = \arccos(1 - 2u)$ and $\phi = \pi \cdot u$, where $u \sim U(0, 1)$ (see Supplementary Material).

Subsequently, we designed and trained a joint Restricted Prior distribution that learns feasible regions of the parameter space and implements parameter prior constraints, as summarised in Table 1. To achieve that, we trained a classifier to learn what is the signal stemming from "valid" parameter combinations, i.e. for which restrictions and boundary conditions on individual parameters are respected (see Fig. S2). To do so, we 1) sampled each parameter independently from uniform priors $p(\mathbf{\Omega})$ specified above, 2) used these samples to generate forward predictions using $f_Y(\mathbf{\Omega})$, 3) tagged the cases that do not meet the boundary conditions into a different class of "non-valid", and 4) trained a classifier accordingly to separate valid vs non-valid cases. This classifier was wrapped into a density distribution object to produce samples only from the ("valid") restricted regions of the parameter space. We trained separate Restricted Priors for the cases of 1, 2 and 3 fibres, and compared the efficiency and performance of the inference against the network using the default uniform priors.

<table>
<tr><th></th><th colspan="4">Parameter constraints</th></tr>
<tr><th></th><th>No Fibres</th><th>1 Fibre</th><th>2 Fibres</th><th>3 Fibres</th></tr>
<tr><td>f_1</td><td>$\sim U[0, 0.0001]$</td><td colspan="3">$\sim U[0.05, 1]$</td></tr>
<tr><td>f_2</td><td colspan="2">$\sim U[0, 0.0001]$</td><td colspan="2">$\sim U[0.05, 1]$</td></tr>
<tr><td>f_3</td><td colspan="3">$\sim U[0, 0.0001]$</td><td>$\sim U[0.05, 1]$</td></tr>
<tr><td>f_{sum}</td><td>$\sim 3 \times U[0, 0.0001]$</td><td>$f_1 \leq 1$</td><td>$(f_1 + f_2) \leq 1$</td><td>$(f_1 + f_2 + f_3) \leq 1$</td></tr>
<tr><td>f_n</td><td colspan="2">None</td><td>$f_1 \geq f_2$</td><td>$f_1 \geq f_2 \geq f_3$</td></tr>
<tr><td>$f_n - f_m \ (m > n)$</td><td colspan="2">None</td><td colspan="2">> 0.05</td></tr>
<tr><td>angle($\mathbf{v}_n, \mathbf{v}_m$)</td><td colspan="2">None</td><td>$> 10°$</td><td>$> 30°$</td></tr>
<tr><td>d_{std}/d</td><td colspan="4">≤ 2</td></tr>
</table>

Table 1: **Parameter constraints and restricted priors** in the Ball&Sticks model. These ensure that i) the sum of the volume fractions ranges within $[0, 1]$, ii) there is some ordering in the volume fractions $f_1 > f_2 > f_3$, iii) detectable crossing angles cannot be lower than a minimum, reflecting limitations in angular contrast/resolution in the data.

2.3.2 Dealing with noise during training

Noisy realisations of the signal were considered during training. For consistency with the MCMC implementation, we assumed zero-mean Gaussian additive noise (i.e. $f_Y(\mathbf{\Omega}) \rightarrow f_Y(\mathbf{\Omega}) + \varepsilon$) with $\varepsilon \sim \mathcal{N}(0, \sigma)$, $\sigma = S_0/SNR$ and SNR the signal-to-noise ratio of the $b = 0$ signal.

Matching the noise level and features between training and test data improves robustness and accuracy of SBI estimates Masutani (2019). However, as such matching cannot be performed a-priori for a generalisable implementation, we evaluated the performance of SBI under two strategies for considering noise during training. For a given set of parameter values $\mathbf{\Omega_i}$ we explored two strategies: a) Single-level noise: A single noisy realisation Y_i was generated with an SNR value chosen randomly from a prior $p(SNR) \sim \mathcal{U}(3, 80)$, b) Multi-level noise: The same noise-free signal was corrupted with different noise levels chosen from multiple predefined SNR intervals (here we used 8 SNR intervals), i.e. $p(SNR_1) \sim \mathcal{U}(3, 10)$, $p(SNR_2) \sim \mathcal{U}(10, 20)$, ... $p(SNR_8) \sim \mathcal{U}(70, 80)$. I.e. we produce 8 noisy realisations of each Y_i, each at a different noise level. We kept the total number of training points equal for both strategies. Hence, the former approach provided for training more combinations of parameters sampled from the priors, each with a single noisy realisation; while the latter approach provided less combinations of parameters for training, but with more examples of noise contamination for each combination.

2.3.3 Model selection using SBI

In the multi-compartment model considered, there is an inherent challenge of deciding how many fibre compartments (i.e. orientations) must be estimated in each voxel. This is a model selection problem that the MCMC implementation addresses using sparsifying, Automatic Relevance Determination (ARD) priors Behrens et al. (2007) (see Suppl. Material). These are improper priors that cannot be sampled from hence not well suited and easily adaptable to simulations from the prior, as needed in SBI. To address the model selection problem, we have implemented and evaluated two strategies:

ClassiFiber + Individual NPEs: Similar to Karimi et al. (2024), we subdivided the problem into 2 stages: an initial classification step to select how many fibre compartments correspond to the signal of each voxel, and a second step

that fits the corresponding model based on the classifier output. For the first step, we designed the *ClassiFiber*, a feed-forward deep neural net that can take the dMRI signal (or a basis signal representation projection) as input and provide the number of fibres as output. The neural network is formed by 6 layers (input, 256, 128, 64, 32, output), with ReLu activation functions, 1% dropout, adaptive learning rate, Adam optimisation, and a Cross-Entropy loss function. We generated 6 million parameter combinations (2 million for N =1, 2 and 3 crossing-fibres configurations sampled from the restricted priors) and generated signal predictions using the multi-shell model. For the second step, we trained 3 independent NPEs for models with N=1, 2, or 3 fibre compartments, respectively, using the basis structure of sec.2.1. Hence, a network e.g. with $N = 2$ compartments was fed with only 2-way crossing patterns during training, while a network with $N = 1$ compartment was not trained with any fibre crossing examples. We refer to this setup as *SBI_ClassiFiber*.

Joint NPE: Contrary to the independent NPEs trained above, we trained a network using a training super-set that included examples of 1,2 and 3 crossing-fibre cases (equal number of 1, 2, and 3 fibre cases). In order to keep a fixed parameter vector size, $N = 3$ compartments were set into the model. For cases where $N < 3$ fibre compartments were simulated (i.e. for single fibres and 2-way crossings), the respective volume fractions of the nth degenerate compartments in the training data were set to $f_n \sim \mathcal{U}(0, 1e-3)$, while the values for the orientations (θ_n, ϕ_n) were randomly sampled from the priors (e.g. when two fibres were simulated, the volume fraction of the third compartment was effectively set to zero in the training data). This joint dataset was used to train a single NPE, which we evaluated on whether it can inherently learn the correct number of compartments, without a prior classification as above. We refer to this setup as *SBI_joint*.

2.4 Datasets, processing and tractography

Data: We evaluated the different architectures using both synthetic and in-vivo data. Preprocessed in-vivo dMRI data were obtained from a previous study Manzano Patron et al. (2024), consisting of 6 re-scans of the same subject, each acquired following a UKBiobank-like multi-shell protocol (2mm isotropic resolution, TR=3s, TE=92ms, MB=3, Siemens Prisma 3T scanner) Miller et al. (2016). Fifty gradient directions at b=1000 s/mm^2 and another fifty gradient directions at b=2000 s/mm^2 were obtained, with a mean SNR of 30 for the b=0 s/mm^2 image. Synthetic data allowed toy examples for testing the inference frameworks under scenarios where the true parameter values were known. We used the Ball&Sticks model to generate such synthetic datasets, following the same protocol (b-values and vectors) as in the in-vivo datasets.

Acquisition-specific vs Acquisition-agnostic training: As the above approach returns trained networks for a specific set of bvals and bvecs, we also explored using an acquisition-agnostic basis set representation of the data as input to the networks, instead of using directly the dMRI signal attenuation. The signal attenuations were expressed using Spherical Harmonics (SH) Descoteaux et al. (2007) (for single-shell) and the MAP-MRI basis (for multi-shell) Özarslan et al. (2013). Given the number of volumes in the data, we retained the even SH coefficients up to order $L = 6$ (28 coefficients), and for the MAP-MRI up to radial order $r = 6$ (45 coefficients) with Laplacian regularization (no positivity constraints nor anisotropic scaling). Both signal representations were obtained using Dipy implementations Garyfallidis et al. (2014).

Probabilistic tractography: Using the trained architectures, we drew samples from the posterior distribution of parameters given the observed data. For the same data and model, we drew posterior samples using MCMC as well. Each provided voxel-wise orientation distributions that mapped uncertainty for the estimated parameters. We subsequently used the estimated posterior orientation samples in probabilistic tractography. We ran the standardised FSL-XTRACT tractography protocols Warrington et al. (2020) using the estimates of each method. These protocols included a range of white matter tracts (commissural, projection, association and limbic). The SBI-derived and MCMC-derived tracts were then directly compared. As both of them reflect stochastic estimation processes, we further explored how the agreement in tract reconstruction between SBI and MCMC compares with scan-rescan reproducibility in MCMC-based tract reconstruction, across the six available repeats.

3 Results

3.1 Single-fibre model

We performed initial evaluations in the simple case of a single-fibre model ($N = 1$ in Eq. 3 and Eq. 4). Fig. 2 shows example comparisons for the posterior mean and uncertainty estimated by MCMC and SBI, in both single-shell (B) and multi-shell (C) models. A default SBI training paradigm was used in these first comparisons (sampling from uniform priors and single-level noise strategy), and the same priors were used as in MCMC. Representative examples of the

marginal posteriors for f_1 in different areas of the brain are shown in Fig. 2A, with very good agreement between MCMC and SBI in the means of these posteriors, and some small variation in their coverage. Overall, as can be seen in the whole-brain maps, the agreement between these two very different inference approaches is extremely high, with correlations of the posterior means above 0.95 for both diffusivity and volume fraction, and median differences of 2° for fibre orientations across white matter. Similarly, uncertainty maps show correlations of 0.92, i.e. a similar uncertainty mapping across the brain (lower uncertainty in WM and higher in CSF and GM for the fibre orientation). A very similar picture is depicted in Fig. 2C for the multi-shell model. These first findings suggest high agreement in both accuracy and precision of SBI estimates against MCMC, in the case of a simple model with a default training paradigm. In the next sections, we increase the model complexity and evaluate how different SBI training and designs affect performance.

3.2 Multi-fibre models - Synthetic data

3.2.1 Optimal training strategies

We subsequently increased the complexity of the models to include crossing fibres ($2 \leq N \leq 3$). First, we compared different training strategies using synthetic data (where ground truth values for model parameters are known). We investigated 1) increase in estimation accuracy due to usage of restricted priors compared to uniform priors, 2) the optimal strategy for considering noise during training, 3) how alternative signal representations work (e.g., using the raw signal vs spherical harmonics), and 4) how different SBI designs perform (independent vs joint training). Fig. 3 shows a collection of comparisons using synthetic data with $N = 2$ fibre compartments, where the anisotropic volume fraction f_1 and the crossing angle are varied. In each case, the mean of the estimated posterior distribution for volume fractions and crossings angles are shown. Given that the ground truth values are known, these results directly inform on the accuracy of the estimates for different inference approaches.

Among the different strategies explored, using a multi-level noise scheme combined with restricted priors provided the best performance in retrieving the ground-truth. These improved both accuracy (see Fig. 3A) and precision (see Fig. 3 A). With restricted priors, the sampling efficiency was increased as all of the proposed parameter combinations met the conditions in Table 1. This led to speed-ups of 62% compared to sampling from uniform priors, where 39% of proposed parameter combinations were rejected (see Fig. 2 for an example of how the parameter space gets constrained). Hence, multi-level noise and restricted-priors are used for any SBI model presented onwards.

Fig. 3B shows a comparison of the obtained estimates when SBI is trained against signal attenuation (for a given set of bvals and bvecs) or acquisition-agnostic signal representations, for both single-shell (using spherical harmonics) and multi-shell (using the MAP-MRI projections) scenarios. Some loss of accuracy can be observed in the signal representation cases, particularly at small crossing angles, suggesting that acquisition-agnostic training induces a small penalty in performance compared to training for data specific to a particular acquisition. However, overall good performance is preserved. For the rest of the paper we kept SBI trained against the full signal, but we also present some in-vivo results using the MAP-MRI basis.

Finally, Fig. 3C shows the performance of MCMC (with $N = 2$ and no ARD priors) compared against NPEs trained only with 2-fibres examples (Independent) or with a joint superset that contains both 1- and 2-fibres cases (Joint). Both MCMC and *Independent* are always set to fit $N = 2$ fibre compartments (i.e. the right model for these data), while *Joint* inherently performs indirect model selection (it can fit either a $N = 1$ or $N = 2$ fibres model). Hence, the former corresponds to SBI_ClassiFiber when the classifier identifies the correct number of compartments (N=2), while the latter corresponds to SBI_joint with $N \leq 2$. All methods recovered appropriately the ground-truth values, with *Joint* having slightly worse performance than the other two, indicative of its higher complexity and its ability to indirectly perform model selection. Nevertheless, for all simulated cases, it could identify the two fibre compartments supported by the data.

The above trends in terms of accuracy for the parameter estimates also hold in case of synthetic data where $N = 3$ fibres have been simulated (see Fig. S4. Taken together the findings suggest that SBI with appropriate training can perform very well in terms of posterior accuracy and precision (and similarly to MCMC) in synthetic data, when the model complexity (i.e. number of compartments is a-priori known).

3.2.2 Dealing with model selection

We subsequently explored more comprehensively the performance of different inference approaches with respect to model selection in 3-way crossing patterns. One of the main challenges when performing inference for a multi-compartment model is the identification of the number of compartments that are relevant to the observed signal. We explored two alternatives to deal with this challenge, SBI_ClassiFiber and SBI_joint. SBI_ClassiFiber performs an initial classification step to decide the appropriate model complexity N and then fits the respective NPE with $N = 1$,

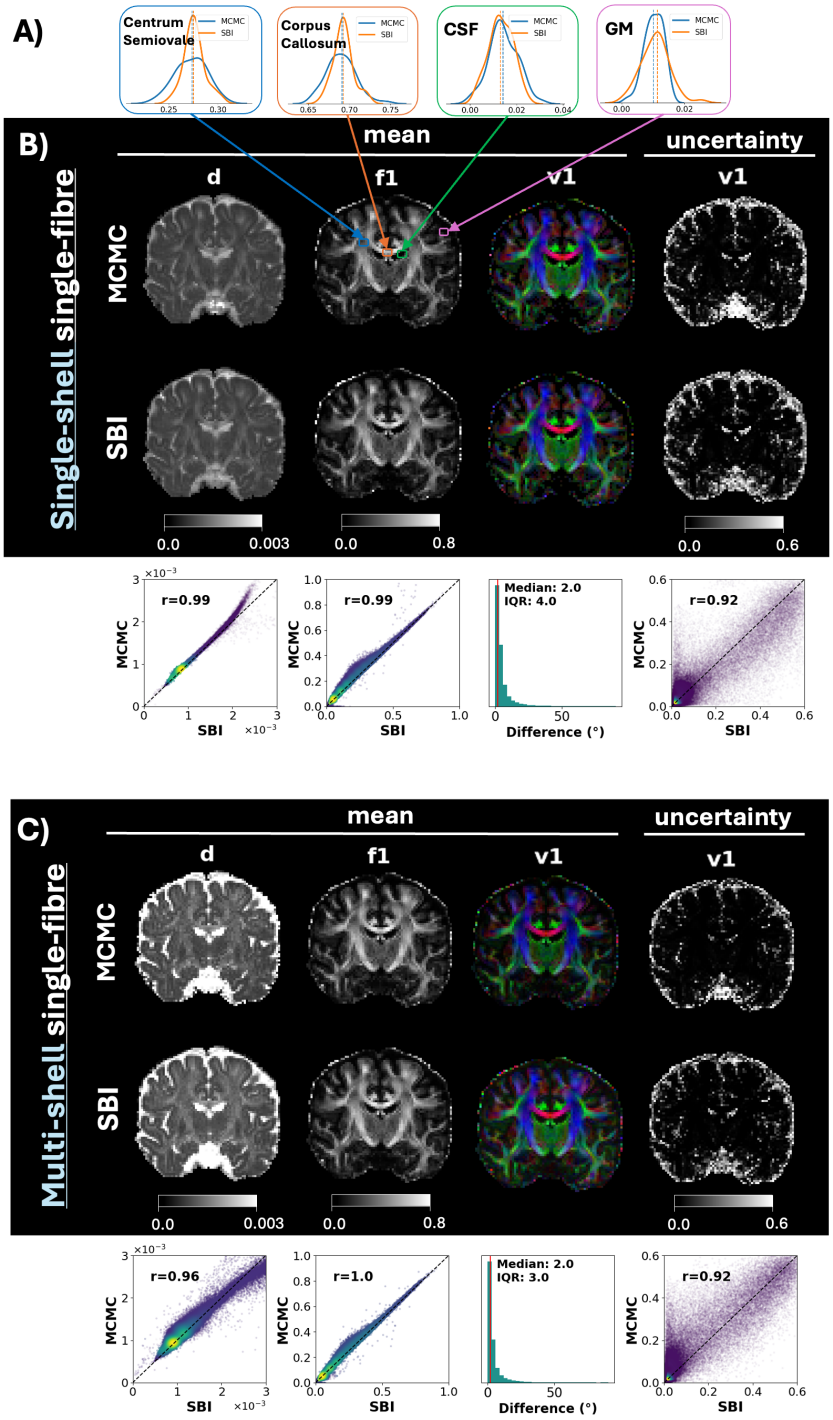

Figure 2: **Comparison of SBI and MCMC estimates for a single-fibre model** ($N = 1$) **in case of single-shell (A,B) or multi-shell (C) data.** A) Example posterior density plots of the anisotropic volume fraction f_1, as estimated by MCMC (blue) and SBI (orange), in four representative voxels: Centrum Semiovale (where fibre crossings are expected), highly anisotropic Corpus Callosum, CSF-filled ventricles, and gray matter. B) Whole-brain maps of estimated parameters using single-shell data (from left to right): Mean of d posterior, mean of f_1 posterior, mean of $\boldsymbol{v}_1$ posterior, orientation uncertainty (width of the $\boldsymbol{v}_1$ posterior). C) Similar as in (B) for multi-shell data. For (B, C), density scatter plots represent the agreement of estimates throughout the brain; the correlation r is calculated for scalar parameters, and the angular difference (in degrees) is calculated for the mean fibre orientations. NPEs were trained with 1 million samples from uniform priors with random SNR values (within the range of 3 to 80).

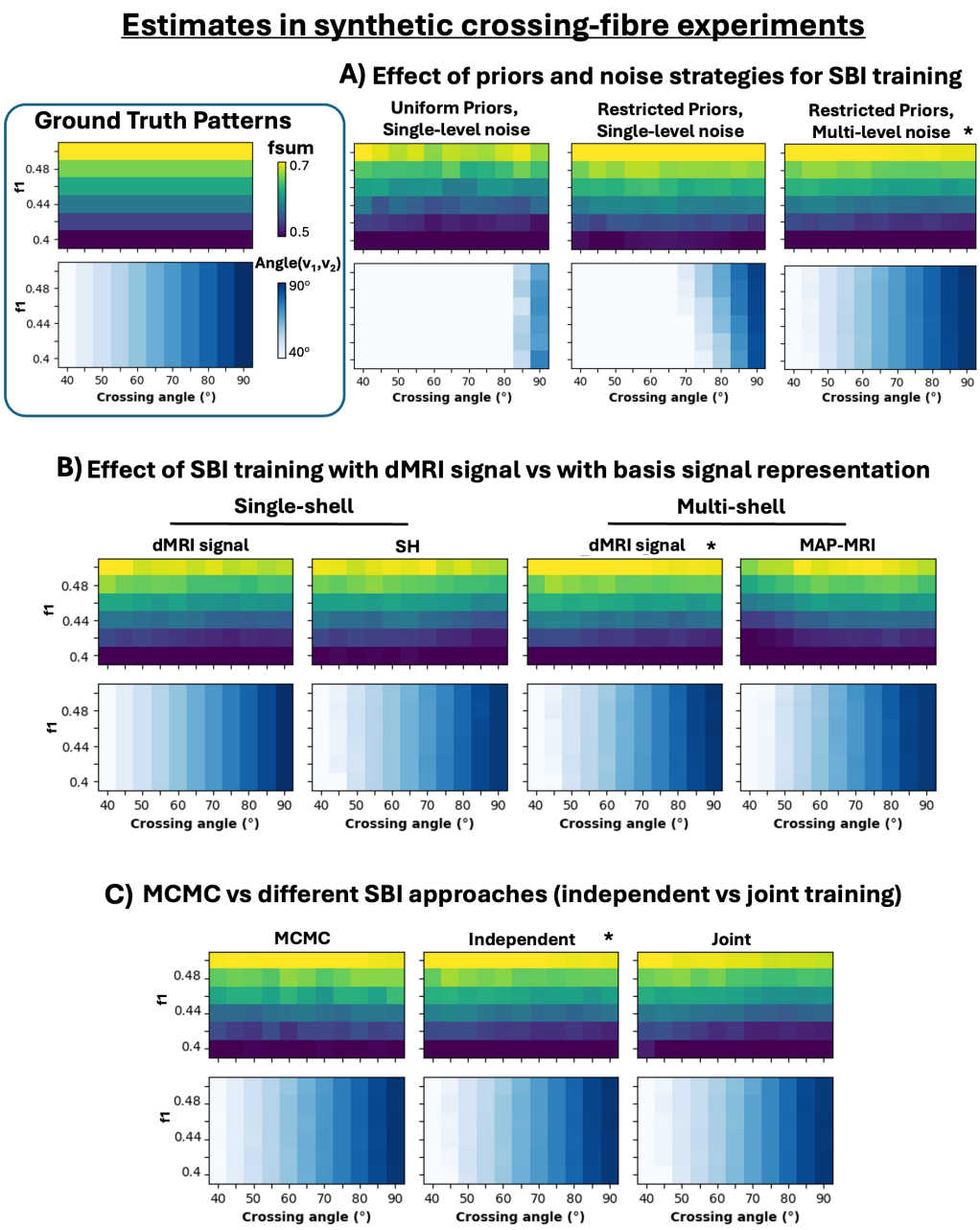

Figure 3: **Evaluating performance of different SBI designs in estimating model parameters in synthetic fibre crossings ($N = 2$), where ground truth is known.** dMRI was synthetically generated for patterns with two crossing fibres, varying the anisotropic volume fraction f_1 and their crossing angle (rest of parameters were $d = 0.0012mm^2/s$, $f_2 = 0.2$, $SNR = 25$ for all experiments). The ground-truth parameter values employed in these experiments are colour-coded (top-left corner), with each entry of these matrices corresponding to a different combination of f_1 and crossing angle. The ability to accurately estimate these values is explored with different SBI training designs. Estimates (median values over 100 noisy realisations) for the mean posterior of f_{sum} (top row) and crossing angle(v_1, v_2) (bottom row) are shown for the following cases: A) Different prior definition and noise strategies for SBI training (from left to right): (i) Uniform priors and single-level noise, (ii) Restricted priors and single-level noise, (iii) Restricted priors and multi-level noise. B) SBI trained with raw dMRI signal (acquisition-specific) vs a basis set signal representation (acquisition-agnostic). Spherical harmonics were used for single-shell data, and the MAP-MRI projection for multi-shell data. C) Comparing MCMC fitting a model with $N = 2$ fibres against NPEs trained only with 2-fibres examples (Independent) or with a joint superset that contains both 1- and 2-fibres cases (Joint). All SBI models were trained with 5 million samples, apart from *Joint* approach which was trained with 6 million samples including 1- and 2-fibre cases (3 million each). Results denoted with (*) correspond to the same SBI design across A, B, C (i.e. trained with only $N = 2$ examples, multi-shell signal, restricted priors and multi-level noise).

$N = 2$ or $N = 3$ compartments, each of which are independently trained. SBI_joint aims to learn the inherent structure by being trained jointly against $N = 1, 2, 3$ cases, avoiding the classification step and the need of training different NPEs. We compared these architectures in synthetic data against MCMC, for which we fit the most complex model ($N = 3$) and used ARD shrinkage priors for the volume fractions of each compartment Behrens et al. (2007).

Fig. 4A shows the performance for similar synthetic examples as before, where 2-fibre crossings were simulated, but up to 3 compartments could be estimated by all approaches. This allowed us to explore whether underfitting/overfitting occurred, contrary to the previous tests, where the number of compartments in the model was pre-fixed to the actual number of compartments simulated. All methods correctly identified the correct number of fibre compartments and showed good overall performance, with MCMC providing the best mean estimates. SBI architectures were close in estimation accuracy, with SBI_joint, as the most complex NPE, having slightly lower performance.

We then tested the model selection performance in a larger set of 10,000 synthetic scenarios sampled from the restricted priors, with $N = 0 - 3$ (i.e. from zero to three fibre compartments, with various combinations of volume fractions and crossing angles). Fig. 4B shows in matrix format, comparisons of the number of fibre compartments detected (matrix rows) when different number of fibres are simulated (matrix columns). These detection matrices are presented for different noise levels and for different inference approaches; diagonal matrices correspond to ideal model selection, while elements above/below the diagonal correspond to over/under-fitting. At very high SNR, MCMC had a correct detection rate (diagonal of the true vs predicted matrices) above 90% for all cases. SBI methods also did well in detecting the correct number of compartments, albeit with a slightly lower performance (particularly for SBI_joint). However, at noisier (and more realistic) scenarios, the performance of SBI methods was more stable than of MCMC, capturing many crossings even at very low SNR. Overall, SBI_ClassiFiber presented the best performance at mid- and low-SNR, identifying half of the 3-way crossings and avoiding overfitting in the non-fibre cases. SBI_joint and MCMC returned similar results at mid-SNR although the former presented a considerably better identification of crossings at low-SNR. A general behaviour observed across all approaches was the tendency to underfit the number of compartments (e.g. return single-fibre when 2 fibres exist) with lower SNR.

3.3 Multi-fibre models - In-vivo data

After evaluating and optimising the performance of SBI with respect to MCMC in terms of accuracy, precision, and model selection in synthetic data, we compared them using in-vivo dMRI multi-shell brain data. As there is no ground-truth available in this case, we used MCMC estimates as the reference to compare against. Fig. 5A shows a comparison of whole-brain mean and uncertainty maps obtained by the different methods for a model with $N = 2$ fibre compartments. Both SBI models provide very high correlations with MCMC (above $r = 0.9$ for all scalar parameters, see Fig. 5B). This inherently implies also a good agreement in the model selection results. Estimated crossing fibres show a good agreement as well (Fig. 5.C), with the median angular difference between SBI and MCMC being around 6° for $\mathbf{v}_1$ and around 10° for $\mathbf{v}_2$ (see Fig. S5A). The example through the centrum semiovale shown in Fig. 5C, highlights all approaches resolving crossings between callosal and corona radiata projections, with the SBI architectures returning less of fibre fanning within the corpus callosum. Regarding orientation uncertainty, MCMC patterns are overall reproduced through SBI although some broader posteriors were observed, especially in the second fibre orientation (see last two columns of Fig. 5A and B).

Similar trends can be observed in Fig. 6 for a $N = 3$ fibres multi-shell model. Panel A) shows mean maps of the diffusivity and volume fractions. As in Fig. 4, SBI_joint showed a more similar number of complex fibre patterns detected to MCMC, with 22.5% and 25.4% of the voxels in the white matter identified as 3-way crossings by each method, respectively, while SBI_ClassiFiber returned 33.3%. High agreement with MCMC was found for the first two compartments by both SBI methods, with correlations of $r_d = 0.97$, $r_{d_std} = 0.88$, $r_{f_1} = 0.97$, and $r_{f_2} \sim 0.88$. Larger differences were found in the third compartment; the correlation in the mean f_3 reduced to 0.7 in SBI_ClassiFiber, and 0.58 in the SBI_joint.

Panel B) of figure 6 shows an example case of the orientations in the centrum semiovale, where 3-way crossings can be expected (corpus callosum fibres running left-to-right, corona radiata running superior-to-inferior and association fibres anterior-to-posterior). All methods depicted these, with SBI_joint having slightly lower coherence and crossings overall. Whole-brain orientation estimates are shown in Fig. S5B).

Uncertainty maps for all three fibre orientations are shown in panel Fig. 6C, where SBI methods exhibited in general higher uncertainty compared to MCMC. Even if the general patterns were similar in all methods (i.e. highest uncertainty in CSF and GM, lower uncertainty in WM), SBI methods returned higher uncertainty in WM with increased model complexity.

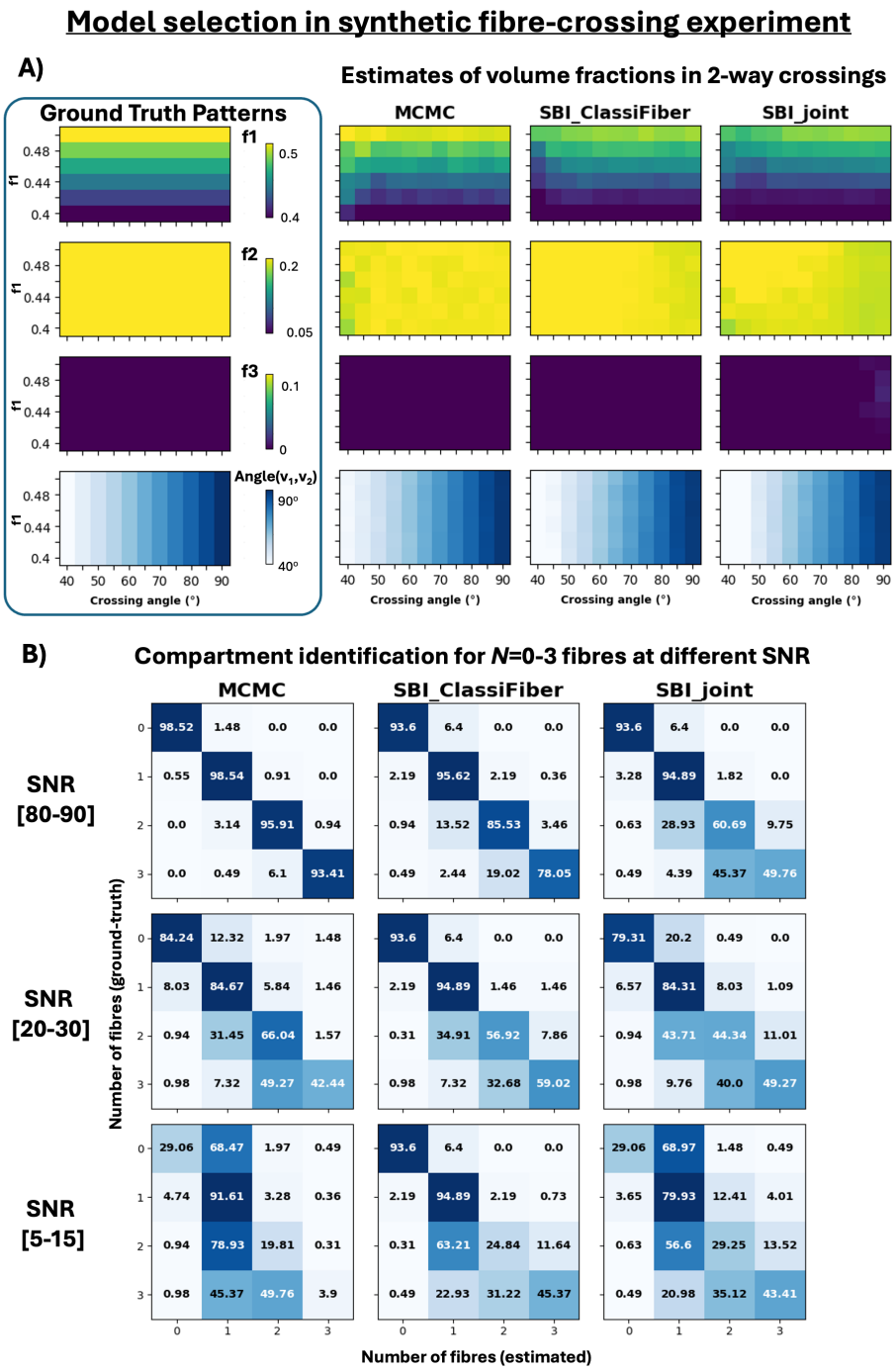

Figure 4: **Evaluating performance in detecting the right number of fibre compartments in synthetic experiments, of MCMC, SBI_ClassiFiber and SBI_joint, all fitting up to $N = 3$ fibre compartments**. A) Recovering a two-fibre pattern (same synthetic data as in Fig.3, where $f_1, f_2 \neq 0$ and $f_3 = 0$), when fitting 3-fibre models to detect over/under-fitting. Estimates (median values over 100 noisy realisations) for the mean posterior of f_1, f_2 and f_3 and crossing angle(v_1, v_2) are shown. Both MCMC and SBI return zero f_3 in all cases, detecting two fibre compartments with the correct volume fractions. B) Performance in compartment identification as a function of SNR for the three different approaches and for 10,000 synthetic patterns containing *N*=0-3 fibres compartments. For each SNR, matrices show for each ground-truth number of fibres (matrix row), the percentage of cases that were estimated as supporting $N = 0 - 3$ patterns (matrix column), i.e. each row of each matrix sums up to 100%. The ideal detection rate corresponds to a diagonal matrix; off-diagonal elements indicate over/under-fitting (above and below diagonal, respectively). In all cases, mean posterior f_n >5% was used to indicate support for the existence of a compartment. Results for SBI_ClassiFiber refers to the number of fibres supported by the final NPE estimates (i.e. f_n >5%), and not to the initial output of the classifier, which can be refined by NPE.

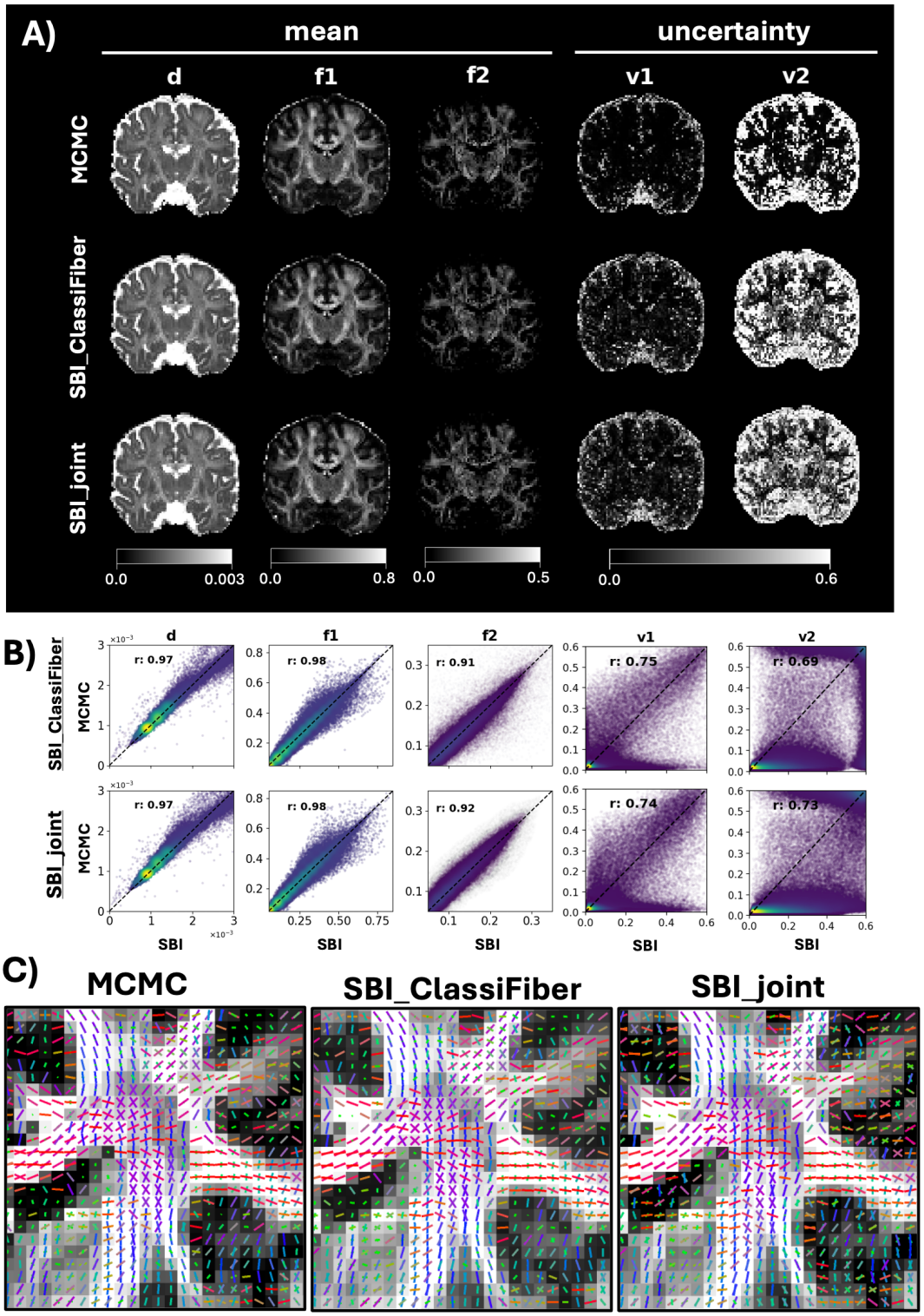

Figure 5: **Inference for in-vivo brain data using a $N = 2$ fibres multi-shell model and different inference approaches (MCMC, SBI_ClassiFiber, SBI_joint)**. A) Exemplar coronal slice with the mean and uncertainty maps estimated by the different methods for different model parameters. B) Density plots representing the agreement between MCMC and SBI in the white matter; correlation r is calculated between scalar maps. C) RGB colour-coded crossing-fibres estimated by each method in the centrum semiovale. The mean posterior orientation is plotted for each compartment.

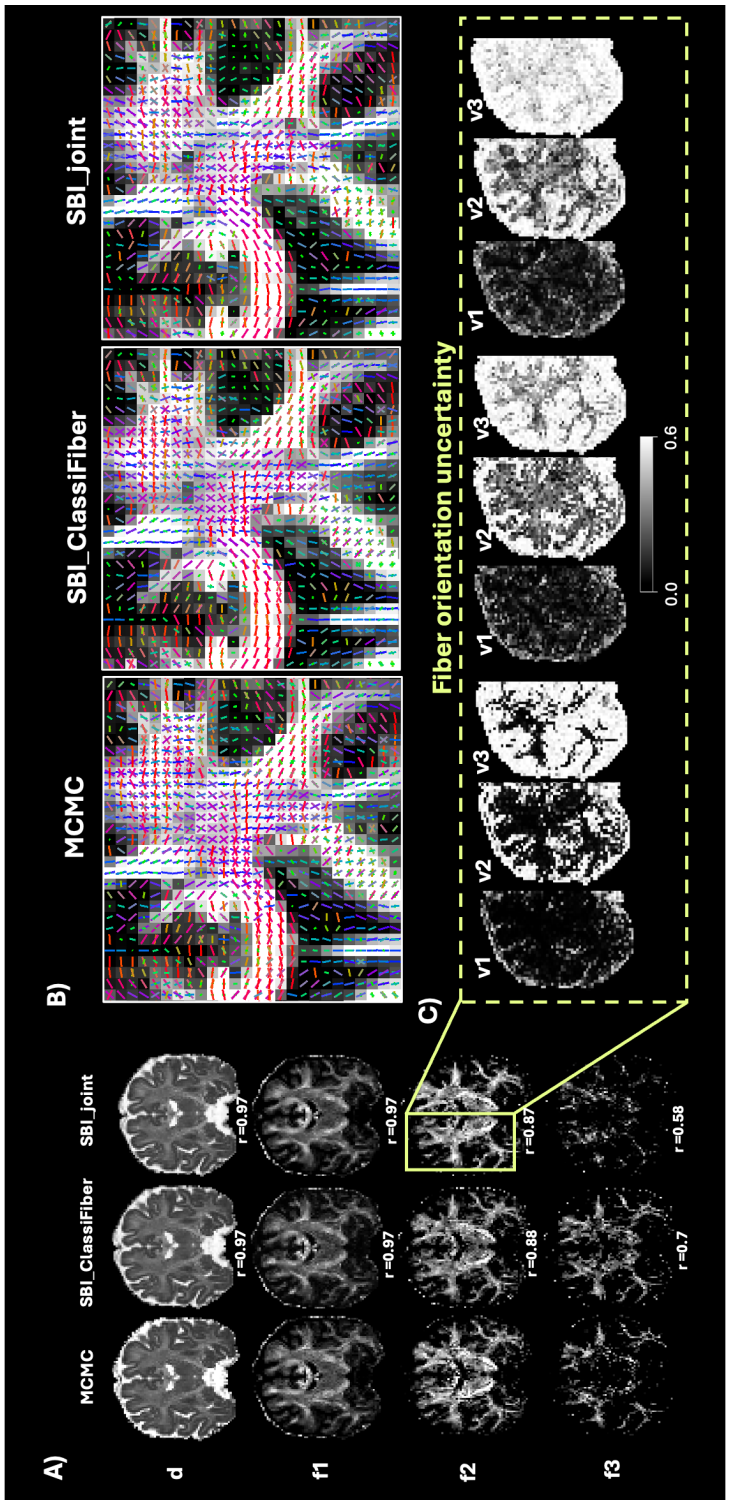

Figure 6: **Inference for in-vivo brain data using a $N = 3$ fibres multi-shell model and different inference approaches (MCMC, SBI_ClassiFiber, and SBI_joint)**. A) Exemplar coronal slice with the mean scalar maps estimated by the different methods for different model parameters (correlation with MCMC is indicated at the bottom of each map). B) RGB colour-coded crossing-fibres estimated by each method in the centrum semiovale. The mean posterior orientation is plotted for each compartment. C) Orientation uncertainty maps for each fiber compartment and for each inference approach.

Finally, as suggested in the previous section, using the MAP-MRI signal representations for a 3-fibres model led to similar parameter estimates (mean posterior) as using the raw dMRI signal (see Fig. S6).

3.3.1 Sources of SBI vs MCMC uncertainty differences

So far we have observed that MCMC and SBI return very similar mean parameters, but they can show discrepancies in the width of the posterior, particularly evident in in-vivo data. We further explored these differences in uncertainty quantification and identified scenarios where these occur.

A first identified case corresponds to the potential of fitting fewer compartments than supported by the data (e.g. due to inaccurate model selection). Fig. 8A shows the marginal and pairwise joint distribution for all parameters in an example synthetic case where a 3-way crossing pattern is simulated (crossing-angles of 90 degrees, $f_1 = 0.4$, $f_2 = 0.3$, $f_3 = 0.2$), but a 2-fibre model is fitted, i.e. wrong model identification. MCMC returns unimodal marginal distributions for all parameters; it finds one of the 3 possible orientation modes and keeps sampling from it (as it is very *expensive* to jump to a different mode), missing the other modes and providing very low uncertainty (overconfidence). On the other hand, SBI captures all three orientation modes, assigning their samples to (mainly) v_1 (see red arrows) which, therefore leads to increased uncertainty due to multiple modes arised from fitting a model with the wrong number of compartments (epistemic uncertainty).

Another source of differences in uncertainty is compartment swapping, which can occur when volume fractions of fibre compartments are very close to each other e.g. $(f_i - f_j) \approx 0$. In such cases, assigning a compartment to be first, second or third becomes arbitrary, as any combination could explain the signal in the same way. SBI is based on minimising the negative log-likelihood, so when the compartments have similar volume fractions, it will equally likely assign a sample from orientation $\mathbf{v}_i$ to parameter $\mathbf{v}_j$ and vice versa. Hence, drawing multiple samples can lead to a multi-modal marginal posterior for a specific fibre orientation parameter with P modes, with P the number of compartments with similar volume fraction. This compartment swapping is much more difficult to occur with a random-walk Metropolis Hastings MCMC algorithm, which needs very high energies to jump between multiple modes and therefore samples tend to inherently cluster together. Nevertheless, this issue in SBI can be mitigated with a simple re-ordering of the posterior samples based on the similarity of orientations. Fig. 8B shows the marginal and pairwise joint distribution for all parameters in an example case from in-vivo brain data, with compartment swapping, before and after the re-ordering to compensate for the swapping.

The above findings suggest that the calculated SBI uncertainty maps are potentially inflated by the way parameter samples are grouped into compartments and uncertainty is estimated. They also hint that individual samples are likely correct in all cases, even if summary uncertainty maps suggest otherwise when compared against summary maps from MCMC. We therefore performed probabilistic tractography, which uses the posterior orientation samples rather than summary uncertainty maps, to test this hypothesis.

3.4 Probabilistic tractography using SBI estimates

Probabilistic tractography spatially propagates orientation uncertainty (i.e. samples from the marginal posterior distribution of the fibre orientations) to build spatial distributions that estimate white matter pathways. We explored how the orientations estimated from the 3 different approaches (MCMC, SBI_ClassiFiber, SBI_joint) affect tractography performance. We used the posterior fibre orientation samples estimated by each method using a $N = 3$ fibre, multi-shell model. Fig. 7 shows example Maximum Intensity Projections (MIPs) of the path distributions for a range of 42 commissural, association, limbic and projection tracts Warrington et al. (2020). For reference, the population-average UK-Biobank atlas is also shown. All methods reconstructed all tracts successfully and showed a high qualititative agreement between the corresponding paths. SBI-reconstructed tracts had a higher correlation with the atlas (up to 15% more in SBI_ClassiFiber), compared to MCMC-based reconstructions.

We then looked into quantitative comparisons between the reconstructed tracts obtained using MCMC vs SBI estimates. Fig. 9A shows scan-rescan reproducibility of each method among 6 consecutive repeats of the same MRI acquisition protocol on the same participant. SBI methods reported higher reproducibility in tract reconstructions, with an overall median correlation r of $r = 0.96$ for SBI_joint and $r = 0.95$ for SBI_ClassiFiber (among tracts and repeats), which is consistently higher than in MCMC ($r = 0.87$). This provides evidence for the hypothesis that posterior orientation samples from SBI are appropriately precise and support the sources of the seemingly higher orientation uncertainty shown in the previous section. In addition, Fig. 9B demonstrates the median (among 6 repeats) correlation between MCMC-based and the corresponding set of SBI-based reconstructed tracts. This is overlayed on a shaded zone (in light blue), which depicts the interquantile range (across the 6 repeats) of the correlations between MCMC-based tract reconstructions themselves. In general, both SBI_ClassiFiber and SBI_joint perform similarly well to MCMC; certain

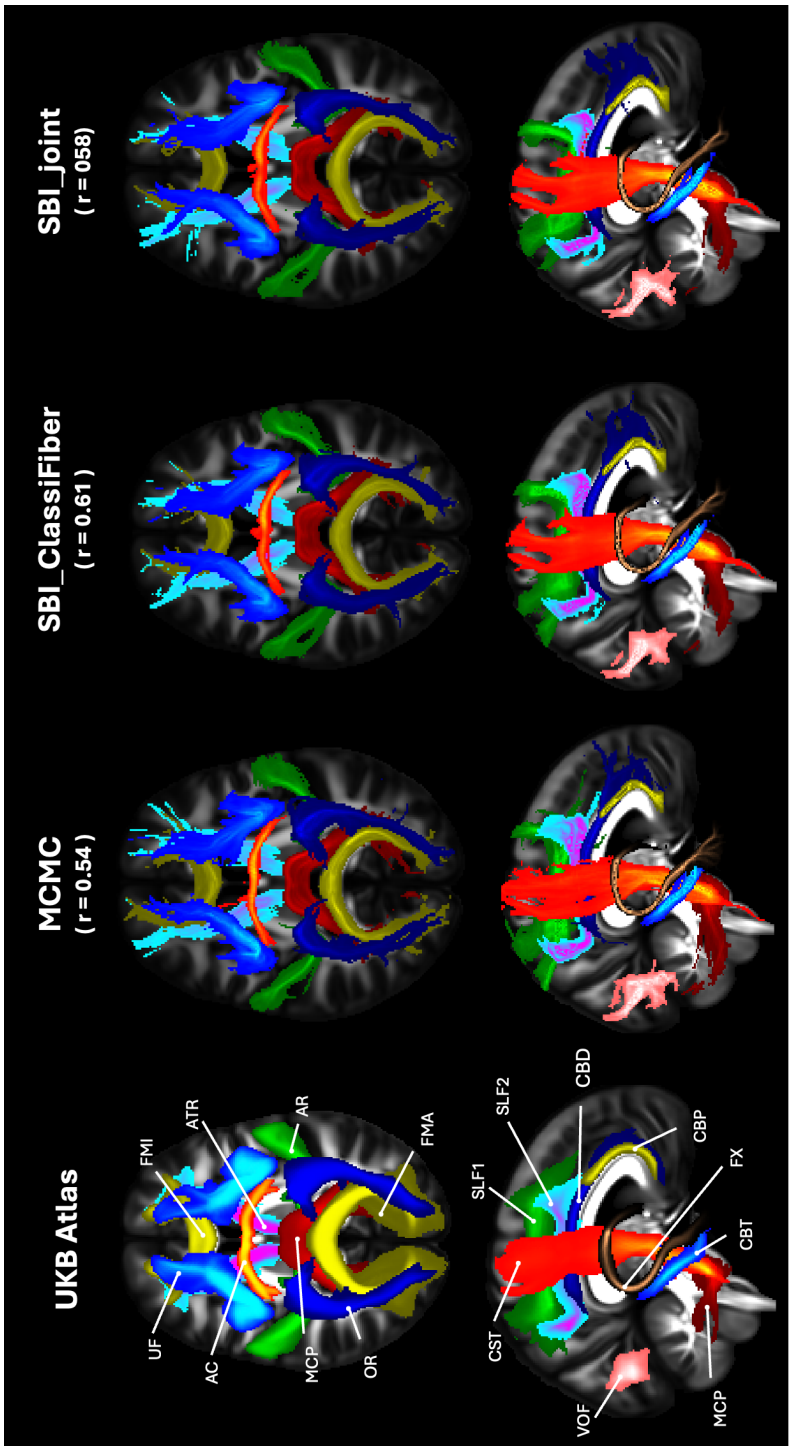

Figure 7: **Probabilistic tractography using MCMC and SBI posterior orientation distributions**. White matter tract reconstruction for 42 bilateral bundles were obtained using the orientation estimates of each method. The spatial distribution of each of the paths is shown as a maximum intensity projection on axial (top row) and saggital (bottom row) planes (thresholded at probabilities>0.1%). Tract reconstructions from a population-average atlas (using data from UK Biobank and MCMC-based orientation inference) is shown as a reference. Each tract was spatially correlated against the corresponding reconstruction from the atlas. The median of these correlation values across tracts is indicated under each method.

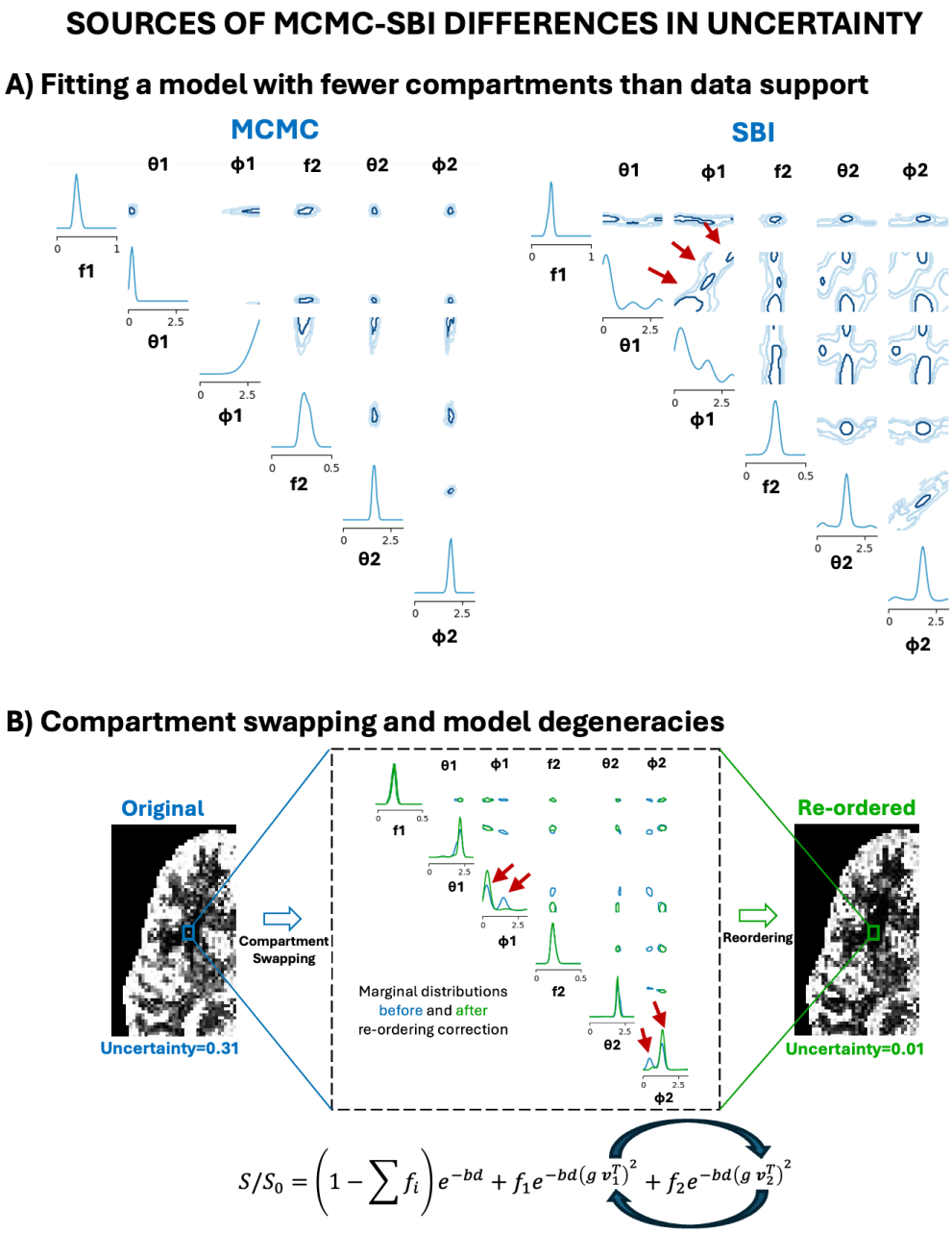

Figure 8: **Sources of differences in uncertainty mapped with MCMC and SBI**. A) Fitting fewer compartments than supported by data: Marginal and pairwise joint distributions of all parameters obtained from a 2-fibre model fitted to synthetic data from a 3-fibre pattern (ground-truth), using MCMC (left) and SBI (right). The marginal posterior distribution for every model parameter is shown first in each row, followed by the joint 2D posterior with every other parameter. When the number of compartments in the model is lower than the number of compartments supported by the data (i.e. effectively a wrong model is fitted), MCMC tends to return unimodal posteriors, missing the third compartment that is supported by the data. This leads to narrow (but wrong) distributions (overconfidence). SBI correctly captures all three modes within the compartments of the model (see red arrows, all captured in the $\mathbf{v}_1$ posterior in this example). However, this leads to broader posteriors (i.e. uncertainty). B) Compartment swapping: Marginal and pairwise joint distributions of all parameters obtained from SBI for a voxel from in-vivo data (square), where $f_1 \sim f_2$ (blue). In such cases, SBI can include samples from the first and second fibre orientation indistinctively to the corresponding marginal posteriors, making them multimodal (red arrows). A simple re-ordering (based on similarity of orientation samples) can mitigate this (green), leading to unimodal posteriors. Notice that there is nothing wrong with either approaches, as the joint orientation posterior samples stay the same, and what changes is how they are distributed to compartments. For reference, orientation uncertainty maps of $\mathbf{v}_2$ returned by SBI, before (left) and after the correction (right) are shown.

tracts that have multiple 3-way crossings along their route are an exception (AF, AR, SLF1, SFL2, SLF3) as for these SBI_ClassiFiber performs better than SBI_joint. Nevertheless, these plots suggest that differences between MCMC and SBI in tract reconstructions are within the range of (and most of the time smaller than) scan-rescan variability (i.e., the variability of MCMC-based tract reconstructions across the 6 repeats). For the exceptions observed, where correlations between MCMC and SBI methods were lower, SBI methods showed a more precise reconstruction among repeats compared to MCMC (see panel A-top), as also indicated in Fig. 9C, where the averaged reconstructed tracts over the 6 repeats are shown.

Taken together, these results indicate that for a range of representative white matter tracts, including those with complex fiber architecture and 3-way crossings along them, SBI-based reconstructions i) agreed better with a population-average white matter atlas, ii) were more precise and reproducible across MRI repeats, iii) had differences with their MCMC counterparts that were within or lower than the magnitude of differences expected due to scan-rescan variability.

3.5 Computational performance

We measured the average computational time required for inference across cases with 1, 2, and 3 fibre compartments (see Table 2). Performing inference with the trained SBI models achieved up to 70x speed-up over the CPU-based MCMC implementation. Note that SBI is a CPU-based Python implementation, while BedpostX is highly optimised in C. All in all, with a simple data-level parallelisation across e.g. 8 CPU-cores, inference on a standard in-vivo brain dataset, as the one used here, with the current SBI implementation takes only $\sim$ 4 minutes.

	MCMC	**SBI_ClassiFiber**	**SBI_joint**
1 fibre	110.65 ms	2.86 ms	-
2 fibres	200.87 ms	3.74 ms	3.42 ms
3 fibres	311.27 ms	4.82 ms	4.38 ms

Table 2: **Computational performance - Time (in milliseconds) needed to perform inference for each method** and obtain 50 samples from the posterior (averaged in a dataset with 10,000 synthetic voxels). Times are all for a single CPU. Time for the classification in SBI_ClassiFiber is orders of magnitude smaller and not included. MCMC includes 1,000 burn-in iterations and a thinning period of 25 (i.e. 2,250 iterations in total).

4 Discussion

In this work, we have presented a Simulation-Based Inference framework for Bayesian fitting of multi-compartment fibre orientation models in diffusion MRI data, mapping parameter uncertainty and utilising the posterior distributions of model parameters given the data to perform probabilistic tractography. Importantly, we performed a direct and comprehensive comparison against classical inference approaches, based on MCMC, to identify optimal training configurations and SBI architectures, and confirm the similarity of voxel-wise estimates and whole-brain tractography reconstructions to an established and heavily used approach. Our findings suggest that SBI with in-silico training can match the performance of classical inference approaches, opening new avenues for amortised inference in challenging likelihood-free scenarios; while offering orders of magnitude of computational speed-up over MCMC.

We performed novel explorations for optimal SBI designs, in the context of dMRI modelling. Firstly, we explored how to incorporate noise into training. Previous studies have either not considered or added noise at random values Jallais and Palombo (2024); Karimi et al. (2024). We have found that following a multi-level noise scheme improves accuracy and precision of SBI, even if 8-times fewer parameter combinations are considered for training (Fig. 3). From the perspective of neural networks applied to conditional density estimation, this noise-regularisation strategy can be seen as a form of data augmentation and is in line with recent findings Rothfuss et al. (2020).

Secondly, classical sampling-based, such as MCMC, can be very flexible in handling parameter constraints as prior knowledge. This is more challenging in an SBI framework, where parametric prior distributions are typically expected. We explored how to incorporate boundaries and non-linear constraints in the parameters. We showed that it is possible to train restricted prior objects to learn highly complex and nonlinear constraints for the model parameters that translates into better performance (Fig. 3) and sampling efficiency (Fig. S2). Density estimators can struggle to generalize across the entire parameter space even for simple problems Hermans et al. (2021), so a trade-off between enough diversity to capture the high-dimensional parameter space (restricted priors) and having enough, slightly-perturbed, samples (noise strategy) can be beneficial during training, as shown by our results.

Thirdly, we explored two strategies for handling model selection in cases of nested multi-compartment models. Recent studies have proposed a classification step that chooses what model complexity to fit Karimi et al. (2024); Consagra

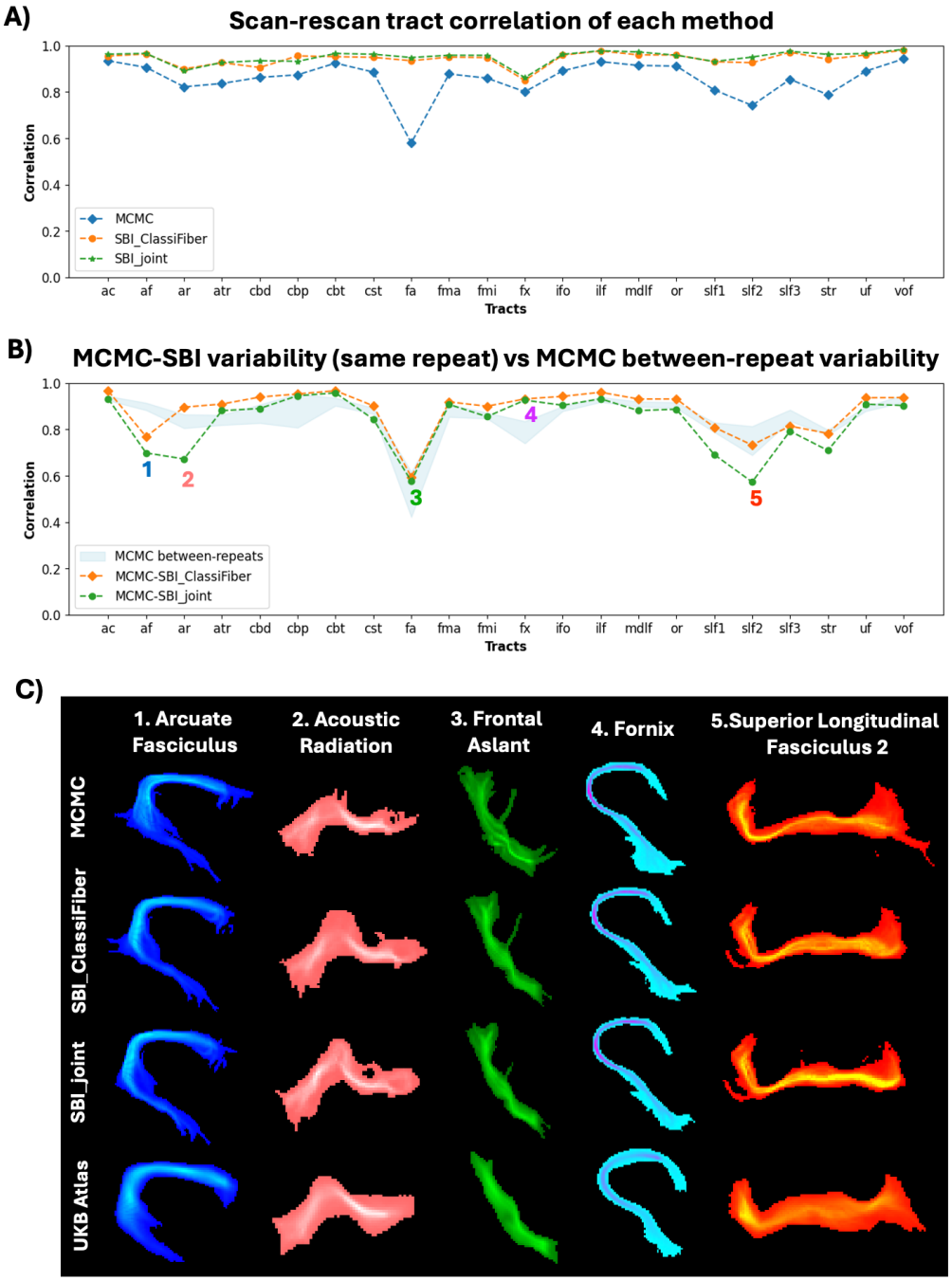

Figure 9: **Scan-rescan reproducibility of probabilistic tractography reconstructions for a range of white matter tracts**, across 6 MRI scan repeats of the same participant. Correlations of the spatial path distributions reconstructed for each tract and using orientation samples from MCMC, SBI_ClassiFiber and SBI_joint are shown. A) Median (across repeats) pairwise tract correlation for each method (i.e. median across rep1 vs rep2, rep1 vs rep3, rep1 vs rep4,... rep5 vs rep6). B) Correlation between SBI and MCMC-based reconstructed tracts (median across 6 repeats SB1 vs MCMC1, SB2 vs MCMC2,...) overlaid on the interquartile range of correlations of MCMC reconstructions across the 6 repeats. C) Averaged path distributions (across the 6 repeats) for a subselection of tracts that had lower agreement between MCMC and SBI.

et al. (2024) followed by an NPE trained independently on that model. We took a step further and evaluated our implementation of this approach (SBI_ClassiFiber) vs another one where NPE learns the appropriate model complexity by being trained on multiple potential models (SBI_joint). Results in both synthetic and in-vivo brain data suggest a comparable performance of both approaches, with SBI_joint providing slightly worse crossing fibre rate detection, accuracy and precision than SBI_ClassiFiber, but tractography end results suggesting similar or better performance than MCMC (which relied on ARD shrinkage priors Wipf and Nagarajan (2007)). I.e. training NPE with enough examples from different configurations can be sufficient to learn the model selection at the cost of some precision loss. This is important, as in SBI_ClassiFiber, the classifier and NPEs are trained independently so a misclassification by the former will directly affect the latter without any feedback. However, by learning the complexity indirectly as part of training, model selection and parameter fitting can interact during inference. Such implementations can lead to more advanced SBI network architectures. For instance, in a recent approach Schröder and Macke (2024), a classifier conditions the density estimator, but one can envisage feedback between the two in a subsequent design.

We evaluated the inference performance of SBI against an MCMC-based framework for the same models, which we have developed in the past Behrens et al. (2007); Hernández et al. (2013); Jbabdi et al. (2012); Hernandez-Fernandez et al. (2019), and is widely and heavily used as part of FSL Woolrich et al. (2009). For both synthetic and in-vivo brain data, we found that SBI closely follows the patterns of MCMC. In terms of agreement in the means of the posterior distributions (as a proxy for accuracy), correlations above 0.9 between SBI and MCMC architectures were found across the brain for most of the parameters in the different models. In terms of uncertainty mapping, correlations above 0.9 were found for simple models (with a single fibre compartment, Fig. 2). As the number of compartments increased, the whole-brain patterns remained (i.e. maximum orientation uncertainty in GM and CSF, lower in WM) but orientation uncertainty estimated by SBI in the WM was generally higher than by MCMC (Fig. 6). Longer training did not change these trends considerably (results not shown). Despite these seemingly large differences in uncertainty mapping, probabilistic tractography (that spatially propagates uncertainty in orientation estimates) using MCMC and SBI approaches returned very similar results. In fact, SBI-based tract reconstructions agreed better with a population-level white matter atlas, were more reproducible across repeats of the same MRI experiment, and their differences with MCMC-based reconstructions had a magnitude smaller than expected from scan-rescan variability (Figs. 7, 9).

We explored what could drive this discrepancy between uncertainty mapping differences at the voxel vs whole-brain level. One clear difference is how uncertainty is depicted at the two levels: a) At the voxel-level, we used the approach implemented in FSL Behrens et al. (2007); Hernández et al. (2013), which groups orientation samples into a finite set of N fibre compartments and subsequently calculates their mean orientation and spread of samples around it using vector dyadic products Jones (2003). This provides a mean orientation and uncertainty around it for each fibre compartment supported per voxel. b) At the whole-brain level probabilistic tractography propagates using directly the orientation samples, i.e. from the joint posterior of orientations. Hence, if the grouping of orientation samples into finite sets is not consistent between MCMC and SBI, large differences can be observed at the voxel level and not at the tractography level. In Fig. 8 we presented scenarios and confirmed with examples from the data where this occurred.

Fitting the wrong model to data can be one such scenario. This adds epistemic uncertainty and the challenge of what the mapped uncertainty represents. In cases of multi-modal orientation distributions, we observed how SBI could capture all the modes (even with a model with fewer compartments than modes), spreading the samples across modes and adding into orientation uncertainty. While the MCMC implementation sampled only from certain modes, providing samples with high precision, but completely ignoring an existing mode of the target distribution (Fig. 8A). Multiple modes can also arise from non-injective models, where different combinations of parameter values produce similarly likely solutions (i.e., degeneracies). This is a known challenge for certain microstructural models and the multi-modal posterior can be used to evaluate the different likely solutions, as illustrated in Jallais et al Jallais and Palombo (2024). Here, by similar mechanisms, we have identified other type of degeneracies using SBI due to compartment swapping (Section 3.3.1).

4.1 Limitations and Future work

Even if we aimed to match as closely as possible MCMC to SBI implementations in order to compare their performance, a perfect matching is not feasible as they represent quite different approaches. A number of algorithmic features can be contributing to observed differences. Firstly, the MCMC implementation relies on ARD shrinkage priors, which introduce large a-priori penalties on model complexity and have considerable effects on fibre crossing detection and on uncertainty mapping. Such improper priors cannot be implemented for SBI in a straightforward manner. Secondly, SBI networks use a Kullback-Leibler divergence loss function to train NPEs. Due to the mode-covering property of the KL divergence, the approximate posterior will always cover the true posterior Minka (2005) with deviations leading to broader posteriors. In density estimators, this is known as *leakage* and is particularly problematic in neural spline

flows because they do not have analytical corrections pap (2016); Lueckmann et al. (2017); **?**); Greenberg et al. (2019). Longer training or refinements, such as importance sampling have been proposed to compensate to some extent to converge to the true posterior Midgley et al. (2023); Dax et al. (2023) but it is still an open problem that requires more research. Thirdly, the Metropolis-Hastings MCMC implementation used for comparisons is based on local sampling techniques. Hence, the energy needed to jump from one mode of a posterior to another can be very high. In this way, MCMC is more likely to converge to one of the modes and "stick to it" than SBI, missing other likely modes. This also suggests that SBI can perform a better global search than MCMC methods, which highly depends on the initialisation and the progression of the chain. Despite all these factors, the level of similarity between MCMC and SBI in local and global estimates was remarkable and considerably higher than the differences.

A limitation specific to SBI is that training against the dMRI signal involves model predictions for a particular acquisition scheme, i.e. a particular set of b values and diffusion-sensitising gradient orientations in the case of dMRI. Even if re-training for new data and acquisition schemes can be done once and is relatively cheap, another approach would be to represent the dMRI signal by a number of orthonormal sets of basis functions. This would allow acquisition-agnostic inference, as NPEs are trained against generalisable basis sets rather than signal sampled with a specific scheme. We explored training against such representations, including spherical harmonics (for single-shell) and MAP-MRI (for multi-shell data). Results in both synthetic and in-vivo brain data suggested that NPEs trained against both representations could provide similar accuracy performance compared to NPEs trained against the signal, at the cost of some precision loss (Fig. 3B and Fig. S3). This agrees with results in Jallais and Palombo (2024) where signal feature selection was done in a data-driven fashion.

In summary, we argue that SBI is a powerful approach for performing Bayesian inference and has shown comparable performance to an established and heavily-used MCMC-based framework. Although further research and validation is needed, SBI could pave the way for tackling unaddressed problems in neuroimaging, given its reliance on in-silico, likelihood-free forward predictions and its scalability to multi-dimensional problems, the two main limitations of classical Bayesian inference approaches.

Acknowledgements

JP and SS are supported by an ERC Consolidator Grant (101000969), and JP is also supported by a Wellcome Trust bioimaging technology award (313367/Z/24/Z). MD, CS and JHM by the German Research Foundation (DFG) through Germany's Excellence Strategy (EXC-Number 2064/1, PN 390727645) and SFB1233 (PN 276693517), the German Federal Ministry of Education and Research (Tübingen AI Center, FKZ: 01IS18039) the Carl Zeiss Foundation and the Else Kröner Fresenius Stiftung (Project ClinbrAIn).

References

(2016). *Fast ϵ-Free Inference of Simulation Models with Bayesian Conditional Density Estimation*. Curran Associates, Inc.

Alexander, D. C. (2008). A general framework for experiment design in diffusion MRI and its application in measuring direct tissue-microstructure features. *Magn. Reson. Med.*, 60(2):439–448.

Alexander, D. C., Dyrby, T. B., Nilsson, M., and Zhang, H. (2019). Imaging brain microstructure with diffusion MRI: practicality and applications. *NMR in Biomedicine*, 32(4).

Beaumont, M. A. (2010). Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41(Volume 41, 2010):379–406.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian Computation in Population Genetics. *Genetics*, 162(4):2025–2035.

Behrens, T., Berg, H. J., Jbabdi, S., Rushworth, M., and Woolrich, M. (2007). Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? *NeuroImage*, 34(1):144–155.

Behrens, T., Woolrich, M., Jenkinson, M., Johansen-Berg, H., Nunes, R., Clare, S., Matthews, P., Brady, J., and Smith, S. (2003). Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magn. Reson. Med.*, 50(5):1077–1088.

Bernstein, A. S., Chen, N.-k., and Trouard, T. P. (2019). Bootstrap analysis of diffusion tensor and mean apparent propagator parameters derived from multiband diffusion MRI. *Magn Reson Med*, 82(5):1796–1803.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, (518):859–877. arXiv: 1601.00670.

Blum, M. G. B. and François, O. (2009). Non-linear regression models for Approximate Bayesian Computation. *Stat Comput*, 20(1):63–73.

Chappell, M. A., Groves, A. R., Whitcher, B., and Woolrich, M. W. (2009). Variational Bayesian Inference for a Nonlinear Forward Model. *IEEE Transactions on Signal Processing*, 57(1):223–236.

Consagra, W., Ning, L., and Rathi, Y. (2024). A Deep Learning Approach to Multi-Fiber Parameter Estimation and Uncertainty Quantification in Diffusion MRI.

Cranmer, K., Brehmer, J., and Louppe, G. (2020). The frontier of simulation-based inference. *Proc Natl Acad Sci USA*, 117(48):30055–30062.

Dax, M., Green, S. R., Gair, J., Pürrer, M., Wildberger, J., Macke, J. H., Buonanno, A., and Schölkopf, B. (2023). Neural Importance Sampling for Rapid and Reliable Gravitational-Wave Inference. *Phys. Rev. Lett.*, 130(17):171403.

Deistler, M., Goncalves, P. J., and Macke, J. H. (2022a). Truncated proposals for scalable and hassle-free simulation-based inference. In *Thirty-Sixth Conference on Neural Information Processing Systems*.

Deistler, M., Macke, J. H., and Gonçalves, P. J. (2022b). Energy-efficient network activity from disparate circuit parameters. *Proceedings of the National Academy of Sciences*, 119(44):e2207632119.

Descoteaux, M., Angelino, E., Fitzgibbons, S., and Deriche, R. (2007). Regularized, fast, and robust analytical Q-ball imaging. *Magnetic Resonance in Medicine*, 58(3):497–510.

Dockhorn, T., Ritchie, J. A., Yu, Y., and Murray, I. (2020). Density Deconvolution with Normalizing Flows. *arXiv:2006.09396 [cs, stat]*.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. (2019). Neural Spline Flows. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Durkan, C., Papamakarios, G., and Murray, I. (2018). Sequential Neural Methods for Likelihood-free Inference.

Eggl, M. F. and Santis, S. D. (2024). More with less: Simulation-based inference enables accurate diffusion-weighted MRI with minimal acquisition time.

Garyfallidis, E., Brett, M., Amirbekian, B., Rokem, A., van der Walt, S., Descoteaux, M., Nimmo-Smith, I., and Dipy Contributors (2014). Dipy, a library for the analysis of diffusion MRI data. *Front. Neuroinform.*, 8.

Geffner, T., Papamakarios, G., and Mnih, A. (2023). Compositional score modeling for simulation-based inference. In *International Conference on Machine Learning*, pages 11098–11116. PMLR.

Girard, G., Aydogan, D. B., Dell'Acqua, F., Leemans, A., Descoteaux, M., and Sotiropoulos, S. N. (2024). 14. probabilistic tractography. In Dell'Acqua, F., Leemans, A., and Descoteaux, M., editors, *Handbook of Diffusion MRI Tractography*. Elsevier.

Gloeckler, M., Deistler, M., Weilbach, C. D., Wood, F., and Macke, J. H. (2024). All-in-one simulation-based inference. In *Forty-first International Conference on Machine Learning*.

Greenberg, D., Nonnenmacher, M., and Macke, J. (2019). Automatic Posterior Transformation for Likelihood-Free Inference. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2404–2414. PMLR. ISSN: 2640-3498.

Gu, X., Eklund, A., Özarslan, E., and Knutsson, H. (2019). Using the Wild Bootstrap to Quantify Uncertainty in Mean Apparent Propagator MRI. *Frontiers in Neuroinformatics*, 13.

Harms, R. L., Fritz, F. J., Schoenmakers, S., and Roebroeck, A. (2024). Fast and robust quantification of uncertainty in non-linear diffusion MRI models. *NeuroImage*, 285:120496.

Hermans, J., Delaunoy, A., Rozet, F., Wehenkel, A., and Louppe, G. (2021). Averting A Crisis In Simulation-Based Inference.

Hernandez-Fernandez, M., Reguly, I., Jbabdi, S., Giles, M., Smith, S., and Sotiropoulos, S. N. (2019). Using GPUs to accelerate computational diffusion MRI: From microstructure estimation to tractography and connectomes. *NeuroImage*, 188:598–615.

Hernández, M., Guerrero, G. D., Cecilia, J. M., García, J. M., Inuggi, A., Jbabdi, S., Behrens, T. E. J., and Sotiropoulos, S. N. (2013). Accelerating Fibre Orientation Estimation from Diffusion Weighted Magnetic Resonance Imaging Using GPUs. *PLOS ONE*, 8(4):e61892.

Jallais, M. and Palombo, M. (2024). mu-GUIDE: a framework for quantitative imaging via generalized uncertainty-driven inference using deep learning. *eLife*, 13.

Jallais, M., Rodrigues, P. L. C., Gramfort, A., and Wassermann, D. (2022). Inverting brain grey matter models with likelihood-free inference: a tool for trustable cytoarchitecture measurements. *Melba*, 1(IPMI 2021):1–28.

Jbabdi, S., Sotiropoulos, S. N., Haber, S. N., Van Essen, D. C., and Behrens, T. E. (2015). Measuring macroscopic brain connections in vivo. *Nat Neurosci*, 18(11):1546–1555.

Jbabdi, S., Sotiropoulos, S. N., Savio, A. M., Graña, M., and Behrens, T. E. J. (2012). Model-based analysis of multishell diffusion MR data for tractography: How to get over fitting problems. *Magnetic Resonance in Medicine*, 68(6):1846–1855.

Jones, D. K. (2003). Determining and visualizing uncertainty in estimates of fiber orientation from diffusion tensor MRI. *Magn. Reson. Med.*, 49(1):7–12.

Jones, D. K. (2008). Tractography Gone Wild: Probabilistic Fibre Tracking Using the Wild Bootstrap With Diffusion Tensor MRI. *IEEE Transactions on Medical Imaging*, 27(9):1268–1274.

Kaden, E., Anwander, A., and Knösche, T. R. (2008). Variational inference of the fiber orientation density using diffusion MR imaging. *NeuroImage*, 42(4):1366–1380.

Kaden, E., Knösche, T. R., and Anwander, A. (2007). Parametric spherical deconvolution: Inferring anatomical connectivity using diffusion MR imaging. *NeuroImage*, 37(2):474–488.

Kaden, E. and Kruggel, F. (2012). Nonparametric Bayesian inference of the fiber orientation distribution from diffusion-weighted MR images. *Medical Image Analysis*, 16(4):876–888.

Karimi, D. and Warfield, S. K. (2024). Diffusion mri with machine learning. *Imaging Neuroscience*, 2:1–55.

Karimi, H. S., Pal, A., Ning, L., and Rathi, Y. (2024). Likelihood-free posterior estimation and uncertainty quantification for diffusion MRI models. *Imaging Neuroscience*, 2:1–22.

Kauermann, G., Claeskens, G., and Opsomer, J. D. (2009). Bootstrapping for Penalized Spline Regression. *Journal of Computational and Graphical Statistics*, 18(1):126–146.

Lueckmann, J.-M., Boelts, J., Greenberg, D., Goncalves, P., and Macke, J. (2021). Benchmarking Simulation-Based Inference. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 343–351. PMLR.

Lueckmann, J.-M., Goncalves, P. J., Bassetto, G., Öcal, K., Nonnenmacher, M., and Macke, J. H. (2017). Flexible statistical inference for mechanistic models of neural dynamics. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Magdon-Ismail, M. and Atiya, A. (1998). Neural Networks for Density Estimation. In *Advances in Neural Information Processing Systems*, volume 11. MIT Press.

Manzano Patron, J. P. (2023). *On noise, uncertainty and inference for computational diffusion MRI*. PhD Thesis, University of Nottingham.

Manzano-Patron, J. P., Kypraios, T., and Sotiropoulos, S. N. (2022). Amortised inference in diffusion mri biophysical models using artificial neural networks and simulation-based frameworks. In *ISMRM (Proc. Intl. Soc. Mag. Reson. Med. 30)*.

Manzano Patron, J. P., Moeller, S., Andersson, J. L., Ugurbil, K., Yacoub, E., and Sotiropoulos, S. N. (2024). Denoising diffusion MRI: Considerations and implications for analysis. *Imaging Neuroscience*, 2:1–29.

Masutani, Y. (2019). Noise Level Matching Improves Robustness of Diffusion Mri Parameter Inference by Synthetic Q-Space Learning. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 139–142. ISSN: 1945-8452.

Midgley, L. I., Stimper, V., Simm, G. N. C., Schölkopf, B., and Hernández-Lobato, J. M. (2023). Flow Annealed Importance Sampling Bootstrap.

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L. R., Griffanti, L., Douaud, G., Okell, T. W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., Matthews, P. M., and Smith, S. M. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci*, 19(11):1523–1536.

Minka, T. (2005). Divergence measures and message passing.

Ning, L., Szczepankiewicz, F., Nilsson, M., Rathi, Y., and Westin, C.-F. (2021). Probing tissue microstructure by diffusion skewness tensor imaging. *Sci Rep*, 11(1):135.

Novikov, D. S. (2021). The present and the future of microstructure MRI: From a paradigm shift to normal science. *Journal of Neuroscience Methods*, 351.

Novikov, D. S., Veraart, J., Jelescu, I. O., and Fieremans, E. (2018). Rotationally-invariant mapping of scalar and orientational metrics of neuronal microstructure with diffusion MRI. *NeuroImage*, 174:518–538.

Pajevic, S. and Basser, P. J. (2003). Parametric and non-parametric statistical analysis of DT-MRI data. *Journal of Magnetic Resonance*, 161(1):1–14.

Palombo, M., Ianus, A., Guerreri, M., Nunes, D., Alexander, D. C., Shemesh, N., and Zhang, H. (2020). SANDI: A compartment-based model for non-invasive apparent soma and neurite imaging by diffusion MRI. *Neuroimage*, 215:116835.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021). Normalizing Flows for Probabilistic Modeling and Inference. *Journal of Machine Learning Research*, 22(57):1–64.

Papamakarios, G., Pavlakou, T., and Murray, I. (2017). Masked autoregressive flow for density estimation. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Papamakarios, G., Sterratt, D., and Murray, I. (2019). Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 837–848. PMLR.

Rothfuss, J., Ferreira, F., Boehm, S., Walther, S., Ulrich, M., Asfour, T., and Krause, A. (2020). Noise Regularization for Conditional Density Estimation.

Salvador, R., Peña, A., Menon, D. K., Carpenter, T. A., Pickard, J. D., and Bullmore, E. T. (2005). Formal characterization and extension of the linearized diffusion tensor model: Linearized Diffusion Tensor Model. *Hum. Brain Mapp.*, 24(2):144–155.

Schröder, C. and Macke, J. H. (2024). Simultaneous identification of models and parameters of scientific simulators.

Sharrock, L., Simons, J., Liu, S., and Beaumont, M. (2022). Sequential neural score estimation: Likelihood-free inference with conditional score based diffusion models. *arXiv preprint*.

Sisson, S. A., Fan, Y., and Beaumont, M. A. (2018). Overview of Approximate Bayesian Computation. *arXiv:1802.09720 [stat]*.

Sotiropoulos, S. N., Jbabdi, S., Andersson, J. L., Woolrich, M. W., Ugurbil, K., and Behrens, T. E. J. (2013). RubiX: Combining Spatial Resolutions for Bayesian Inference of Crossing Fibers in Diffusion MRI. *IEEE Trans. Med. Imaging*, 32(6):969–982.

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518.

Tejero-Cantero, A., Boelts, J., Deistler, M., Lueckmann, J.-M., Durkan, C., Gonçalves, P., Greenberg, D., and Macke, J. (2020). sbi: A toolkit for simulation-based inference. *JOSS*, 5(52):2505.

Warrington, S., Bryant, K. L., Khrapitchev, A. A., Sallet, J., Charquero-Ballester, M., Douaud, G., Jbabdi, S., Mars, R. B., and Sotiropoulos, S. N. (2020). XTRACT - Standardised protocols for automated tractography in the human and macaque brain. *NeuroImage*, 217:116923.

Whitcher, B., Tuch, D. S., Wisco, J. J., Sorensen, A. G., and Wang, L. (2008). Using the wild bootstrap to quantify uncertainty in diffusion tensor imaging. *Hum Brain Mapp*, 29(3):346–362.

Wildberger, J., Dax, M., Buchholz, S., Green, S., Macke, J. H., and Schölkopf, B. (2024). Flow matching for scalable simulation-based inference. *Advances in Neural Information Processing Systems*, 36.

Wipf, D. P. and Nagarajan, S. S. (2007). A New View of Automatic Relevance Determination. page 8.

Woolrich, M. W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., Beckmann, C., Jenkinson, M., and Smith, S. M. (2009). Bayesian analysis of neuroimaging data in FSL. *NeuroImage*, 45(1):S173–S186.

Yendiki, A., Panneck, P., Srinivasan, P., Stevens, A., Zöllei, L., Augustinack, J., Wang, R., Salat, D., Ehrlich, S., Behrens, T., Jbabdi, S., Gollub, R., and Fischl, B. (2011). Automated Probabilistic Reconstruction of White-Matter Pathways in Health and Disease Using an Atlas of the Underlying Anatomy. *Frontiers in Neuroinformatics*, 5:23.

Özarslan, E., Koay, C. G., Shepherd, T. M., Komlosh, M. E., İrfanoğlu, M. O., Pierpaoli, C., and Basser, P. J. (2013). Mean apparent propagator (MAP) MRI: a novel diffusion imaging method for mapping tissue microstructure. *Neuroimage*, 78:16–32.

Supplementary Material

Mapping uncertainty for probabilistic tractography using Simulation-based Inference: A comparison with classical Bayes in diffusion MRI

Jose-Pedro Manzano-Patron[1], Michael Deistler[2], Cornelius Schröder[2],Theodore Kypraios[3]
Pedro Gonçalves[4], Jacob H. Macke[2,5], Stamatios N. Sotiropoulos[1]

[1]*Sir Peter Mansfield Imaging Centre (SPMIC), School of Medicine, University of Nottingham*
[2]*Machine Learning in Science, Excellence Cluster Machine Learning, University of of Tübingen & Tübingen AI Center*
[3]*School of Mathematical Sciences, University of Nottingham*
[4]*VIB-Neuroelectronics Research Flanders (NERF)*
[5]*Department Empirical Inference, Max Planck Institute for Intelligent Systems, Tübingen*

MCMC implementation

We used MCMC implementation of FSL-bedpostx Behrens et al. (2007); Jbabdi et al. (2012); Hernández et al. (2013). The likelihood is using the Ball&Sticks forward model and is defined assuming: a) independence between diffusion-sensitizing measurements $j = 1 : J$, b) additive zero-mean Gaussian noise with precision $\tau = 1/\sigma^2$. Gaussian noise is a good approximation for Rician MRI noise when the SNR is larger than 3 Salvador et al. (2005). The likelihood function, $p(Y|\mathbf{\Omega}, \boldsymbol{\tau})$ is therefore:

$$p(Y|\mathbf{\Omega}, \tau) = \mathcal{N}(S, \sigma^2) \tag{5}$$

$$p(Y|\mathbf{\Omega}, \tau) = \prod_{j=1}^{J} \left(\frac{\sqrt{\tau}}{\sqrt{2\pi}} \right) \cdot \exp\left(-\frac{\tau}{2} \sum_{j=1}^{J} (Y_j - S_j)^2 \right), \tag{6}$$

where Y_j is the measured signal with gradient j, Ω are model parameters of interest, and S_j is the predicted signal by the forward model, given by the Ball&Sticks model (see eq.3 and eq.4 in the main text). Noise variance is marginalised out using reference priors (so that it is not inferred), leading to a likelihood dependent only on model parameters Ω Behrens et al. (2003):

$$p(Y|\mathbf{\Omega}) \propto \left(\sum_{j=1}^{J} \frac{(Y_j - S_j)^2}{2} \right)^{-\frac{J}{2}} \tag{7}$$

Regarding the prior distributions of the parameters, the prior on spherical angles ensures a uniform prior on the sphere, while positivity constraints are used on S_0, d, and f_n (Behrens et al., 2003).

$$\begin{gathered} p(\theta_n, \phi_n) \sim \mathcal{U}(0, 2\pi)|_{cart} \rightarrow \frac{|\sin\theta_n|}{2}|_{sph} \\ p(f_1) \sim \mathcal{U}(0,\, 1) \\ p(d) \sim \mathcal{U}(0,\, 1) \\ p(S_0) \sim \mathcal{U}(0,\, \infty) \end{gathered}$$

In case of multiple fibre compartments, the sum of all anisotropic volume fractions is constrained to be in [0,1]. In addition, Automatic Relevance Determination (ARD) priors are used for the corresponding volume fractions f_n, $n \geq 2$. ARDs are shrinkage priors that a-priori heavily penalise model complexity and the existence of extra compartments, providing high support for near to zero values for volume fractions f_n, $n \geq 2$. They therefore allow indirect model selection to be performed, when the most complex model is fitted, with low prior support for multiple compartments. As indicated in (**?**), ARD priors are $p(f_n) \sim 1/f_n \quad (n \geq 2)$.

A univariate Random-walk adaptive Metropolis-Hastings algorithm is used for MCMC. Model parameters are initialised by a non-linear deterministic fit (using Levenberg-Marquadt non-linear optimisation). Proposals for MCMC are univariate Gaussians, centred in the last accepted sample and with adaptive variance to ensure a balanced acceptance/rejection ratio of 50%. MCMC runs for 1,000 burn-in iterations that allow the chain to converge and subsequently 1,250 samples are obtained for each parameter. To avoid autocorrelation in the Markov chain, one of every 25 samples is retained (thinning=25). Hence, after thinning and burn-in, the approximated posterior distribution of each model parameter contains 50 samples.

Uniform sampling on the sphere

Linear and spherical geometries obey different principles. A uniform sampling in the plane does not return a uniform sampling in the sphere, as there is a different surface area for the same number of points. Close to the poles of the sphere, the differential surface area element gets smaller (it shrinks) and the wrapping of the uniform planar distribution produces an oversampled area. In contrast, close to the equator, the differential surface area increases (it stretches)

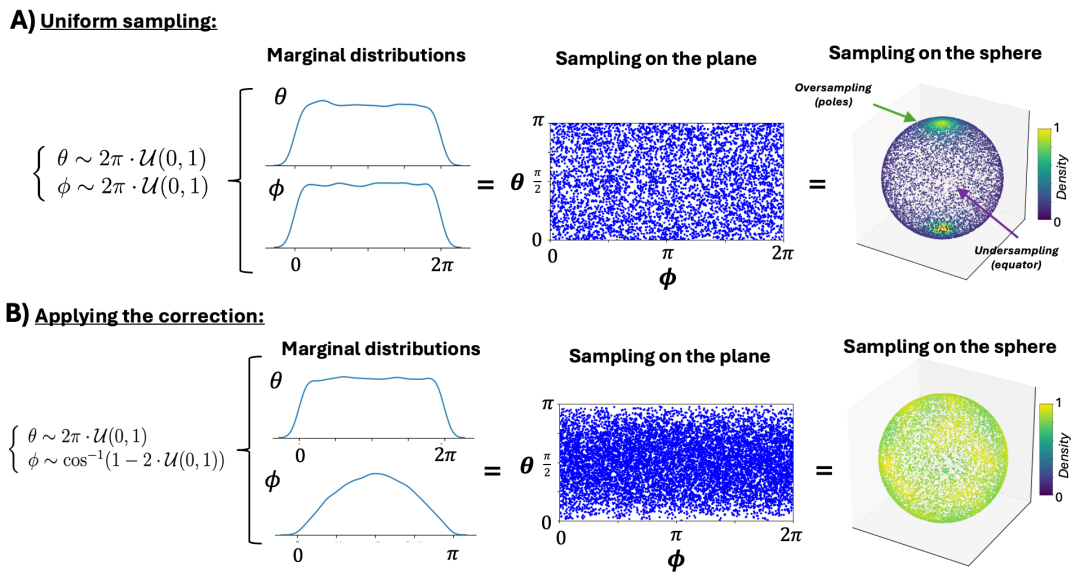

Figure S 1: **Defining uniform priors on the sphere for fibre orientations**. A) Uniform sampling of orientations in the interval $(0, 2\pi)$ leads to uniform coverage on the plane, but not on the sphere, as it produces oversampling at the poles and undersampling at the equator. B) Applying the correction, uniform sampling on the sphere can be achieved, even if this does not look uniform on the plane $\theta - \phi$. In both A) and B), from left to right: (i) Definition of the prior distributions of the fibre orientations, (ii) Marginal distributions produced when sampling from such priors, (iii) joint distribution represented on the $\theta - \phi$ plane, (iv) joint distribution represented on the sphere, colour-coded by the density of the samples.

and the density of points is lower (see Fig.1.A). Thus, we should include fewer points near the poles and more points near the equator to achieve a uniform distribution on the sphere. We apply the following correction to achieve uniform sampling in the sphere when training and evaluating the SBI framework (see Fig.1.B):

$$\theta \sim 2\pi \cdot \mathcal{U}(0,1) \tag{8}$$

$$\phi \sim \cos^{-1}(1 - 2 \cdot \mathcal{U}(0,1)) \tag{9}$$

Estimating mean fibre orientations

When having multiple samples from spherical angles (θ, ϕ), averaging them to get the mean orientation will provide wrong results in many cases. E.g., consider two orientations represented by angles $\alpha_1 = 1°$ and $\alpha_2 = 359°$. The arithmetic mean would be $\alpha_{\text{mean}} = 180°$, which points in a completely different direction; the true mean should be around $0°$ or $360°$. As indicated in section 3.2.2, angles or orientations in spherical geometry follow non-linear geometry and can wrap around (e.g.,$0°$ and $360°$ point actually to the same direction). The arithmetic mean assumes linearity and hence does not correctly account for the curvature of the sphere.

As each fibre of these multi-compartment models is represented by an independent tensor, the mean orientation of the fibre orientation samples $\hat{\mathbf{v}}_n$ (dyadic vector) can be obtained from the dyadic tensor $\mathbf{D}$. They dyadic tensor is the outer product of the vector $\hat{\mathbf{v}}_n$ with itself, averaged over the M samples:

$$\mathbf{D} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{v}_m \mathbf{v}_m^\top \tag{10}$$

The principal eigenvector (the one corresponding to the largest eigenvalue) represents the mean orientation of the fiber:

$$\mathbf{D}\mathbf{e}_1 = \lambda_1 \mathbf{e}_1 \tag{11}$$

, where $\mathbf{v}_1$ is the principal eigenvector and λ_1 is the largest eigenvalue. This can be rewritten as:

$$(\mathbf{D} - \lambda_1 \mathbf{I})\mathbf{e}_1 = \mathbf{0} \tag{12}$$

, where $\mathbf{I}$ is the identity matrix of the same size as $\mathbf{D}$. To solve for λ_1, we need to solve the characteristic equation:

$$\det(\mathbf{D} - \lambda \mathbf{I}) = 0 \tag{13}$$

The determinant of $\mathbf{D} - \lambda\mathbf{I}$ is a polynomial in λ. The roots of this polynomial are the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. The largest root is typically chosen as λ_1 because it corresponds to the direction of greatest variance. Substituting λ_1 back into the characteristic equation 13 provides a system of linear equations that can be solved to find the eigenvector $\mathbf{e}_1 = \mathbf{v}_n$.

The eigenvalue λ_n provides a measure of how concentrated the orientations are around that direction n. Hence, the orientation uncertainty can be computed as:

$$uncertainty(\hat{\mathbf{v}}_n) = 1 - \text{EigMax}(\mathbf{D}) \tag{14}$$

Accuracy and Precision Metrics

To provide quantitative performance metrics, a reference value is generally needed. For comparison of the mean estimates in synthetic data, these reference values are the ground-truth parameter values used to simulated the dMRI signals. However, there is no ground-truth for uncertainty, nor for estimates in in-vivo brain data. In those cases, we compared the SBI estimates against the estimates from the established MCMC implementation in FSL-bedpostX, i.e. the latter are used as the reference values.

Accuracy The agreement between scalar parameters is generally indicated as the linear Pearson's correlation r between the reference values and the mean of the estimated posterior distribution samples, while errors are calculated as the $\%$ of their difference:

$$e_{\omega_p}[\%] = 100 \times \frac{\omega_p - \hat{\omega}_p}{\omega_p} \tag{15}$$

For fibre orientations, the angular difference $\angle e$ was defined as the crossing-angle between the reference value $\mathbf{v}_n$ and the mean estimated $\hat{\mathbf{v}}_n$ from the posterior samples:

$$\angle e_n[^\circ] = cos^{-1}\langle \mathbf{v}_n \cdot \hat{\mathbf{v}}_n \rangle), \tag{16}$$

where $\hat{\mathbf{v}}_n$ is the dyadic vector, i.e. the principle eigenvector of the dyadic tensor (see section 3.2.2). In the case of multiple fibre orientations, we matched the closest pairs before calculating the errors per orientation.

For model selection, we calculated the confusion Matrix (CM) of the actual number of fibres N vs predicted number of fibres $\hat{N}$ for each approach. This offered not only the detection rates for each fibre but also potential trends and patterns in the mis-classified cases.

For tractography evaluations, for each estimated tract, the corresponding spatial distribution was thresholded ($\geq$ 0.001) to remove noisy tails. To compare a pair of corresponding tract reconstructions, we calculated the Pearson's correlation between the respective spatial distributions (after vectorising each of them). To avoid bias from any of the tract reconstructions in comparison, the correlation was performed within a population-based atlas-derived tract mask, available from **?**.

Precision For scalar parameters, the standard deviation (width) of the posterior distribution was used to obtain a metric of uncertainty. For fibre orientations, the maximum eigenvalue λ_n of the dyadic tensor of the posterior orientation

samples was used to provide a measure of uncertainty, i.e. how concentrated the orientations were around the mean direction of the samples. In synthetic data, we evaluated how such orientation uncertainty changed as a function of noise level and training data size, as uncertainty is expected to reduce as the SNR and/or the training data size increase (see also Manzano-Patron et al. (2022)). In real data, we used MCMC uncertainty maps as reference for direct comparisons, and the probabilistic tractography results as indirect indications of how consistent spatially-propagated uncertainty were.

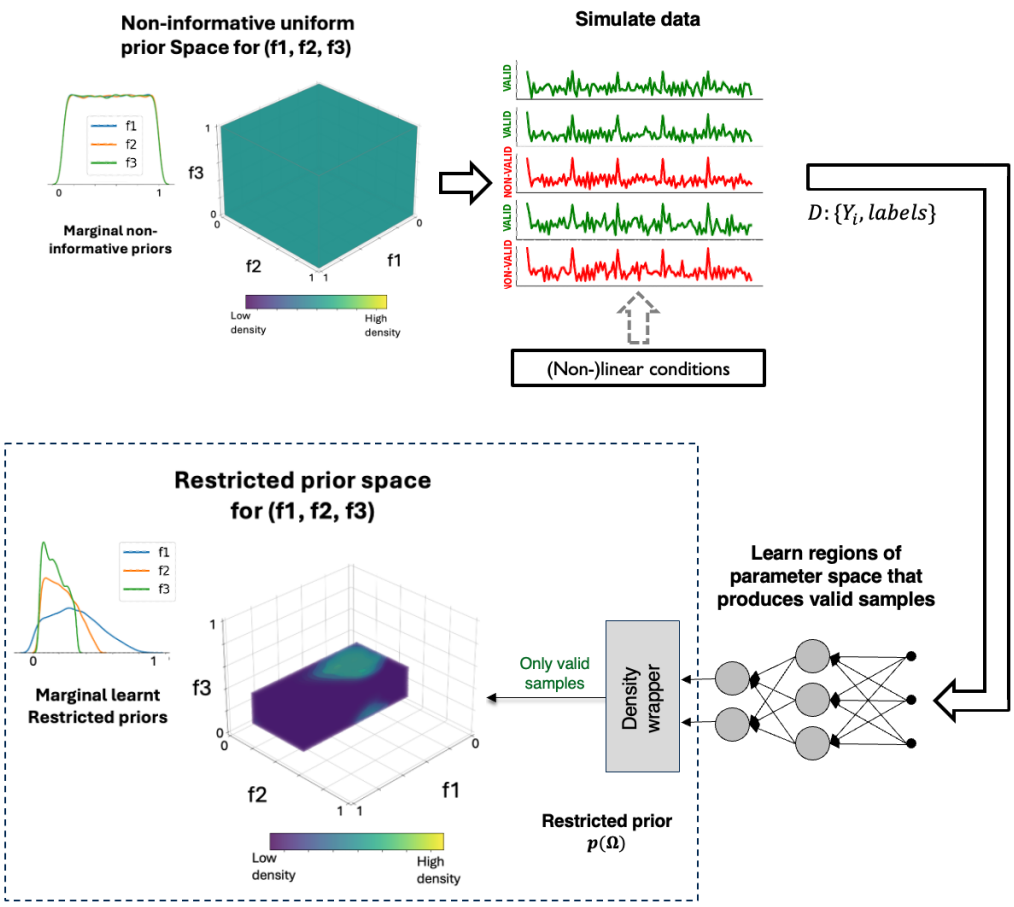

Figure S 2: **Restricted priors constraining the parameter space** compared to non-informative uniform prior sampling. We simulated data Y using samples from the non-informative uniform priors and tagged these as "valid" or "non-valid" (according to the constraints we wanted to enforce) to create our training dataset D. A classifier was then trained to learn what is the signal stemming from "valid" parameter combinations and was used to produce samples only from the regions of the parameter space of interest i.e. a restricted prior. The figure shows an example of the learnt restricted priors used in this paper, which reflect the conditions and non-linear boundaries specified in Table 1 and target the high density areas; in this case, the marginal (top 2-D histograms) and the 3-D joint space of (f_1, f_2, f_3) is shown. For a 3-fibre model, the sampling efficiency of prior samples is also increased with speed-ups of 105% compared to using a uniform sampling scheme and reject using the constraints as a post-sampling step.

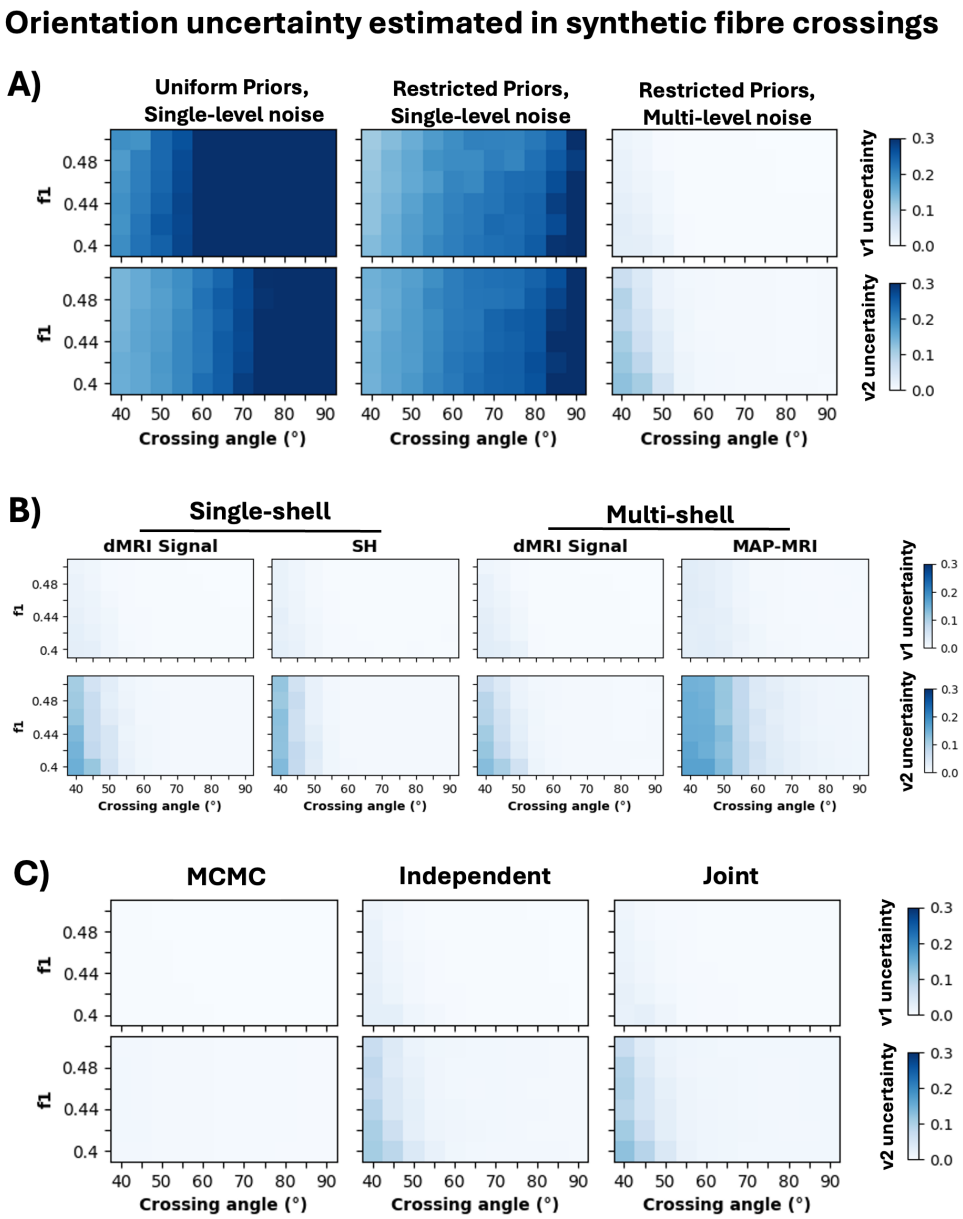

Figure S 3: **Evaluating performance of different SBI designs in mapping orientation uncertainty in synthetic fibre crossings** ($N = 2$)**.** dMRI was synthetically generated for patterns with two crossing fibres, varying the anisotropic volume fraction f_1 and their crossing angle (rest of parameters were $d = 0.0012\ mm^2/s$, $f_2 = 0.2$, $SNR = 25$ for all experiments). Uncertainty in $\mathbf{v}_1$ and $\mathbf{v}_2$ obtained from the width of the respective posteriors (median over 100 noisy realisations) are shown for the same cases as in Fig. 3: A) Different prior definition and noise strategies for SBI training. B) SBI trained with raw dMRI signal (acquisition-specific) vs a basis set signal representation (acquisition-agnostic). C) Comparing MCMC fitting a model with $N = 2$ fibres against NPEs trained only with 2-fibres examples (Independent) or with a joint superset that contains both 1- and 2-fibres cases (Joint).

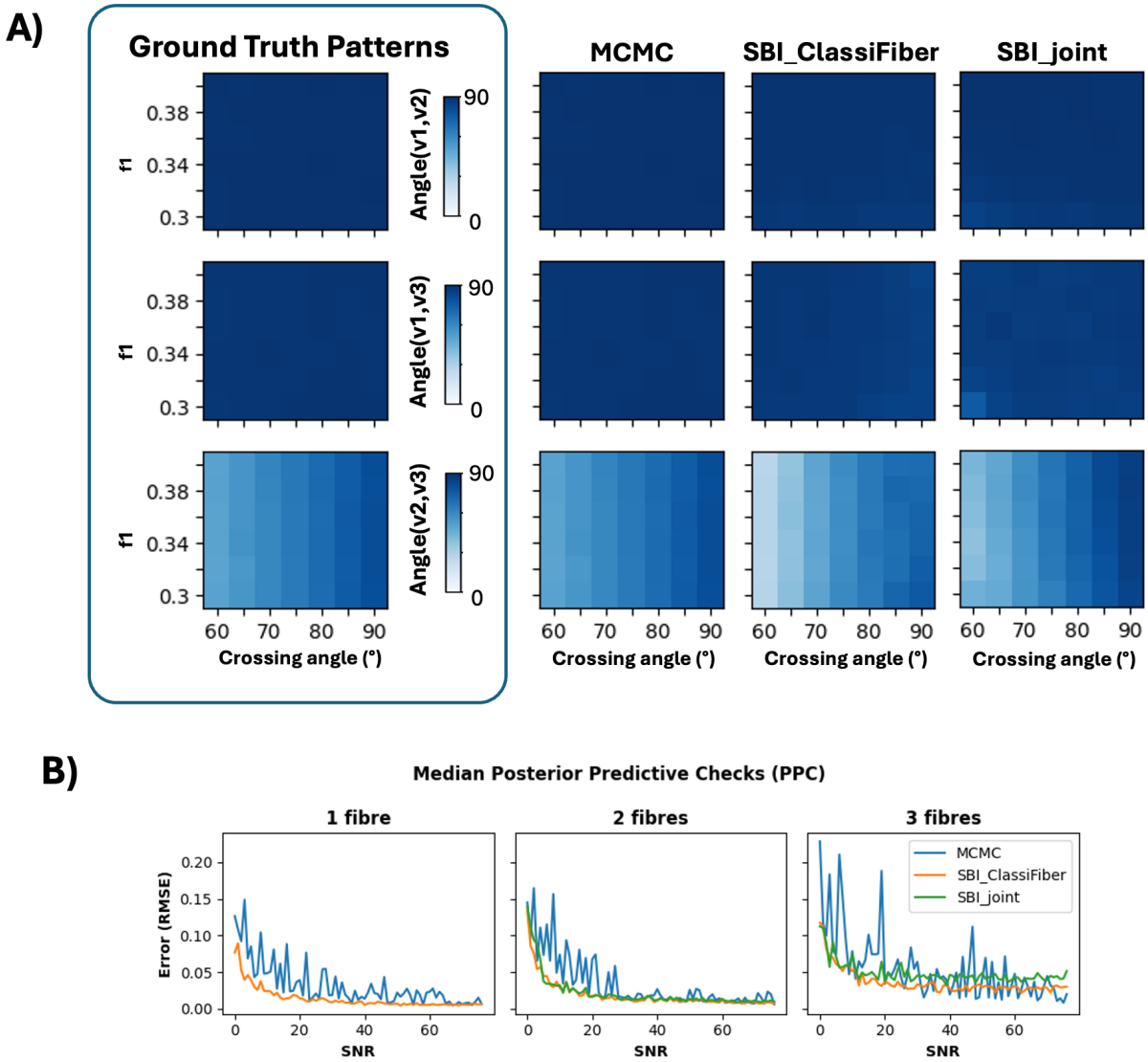

Figure S 4: **Evaluating performance of different SBI designs and MCMC in estimating crossing angles in synthetic 3-way fibre crossings** ($N = 3$)**.** A) Estimates from three approaches for synthetic data, where the first fibre orientation was perpendicular to the rest of compartments, and f_1 and the crossing angle between the second and third fiber were varied. The rest of the parameters were kept to $d = 0.0012$, $f_2 = 0.25$, $f_3 = 0.15$, $SNR = 25$. B) Posterior Predictive Checks (PPC) as a function of SNR for synthetic experiments with different fiber complexities ($N = 1, 2, 3$) and parameter combination. Each plot shows the relative error (median across 1,000 experiments) of the model predicted signal (using the mean posterior value for each parameter) with respect to the true signal.

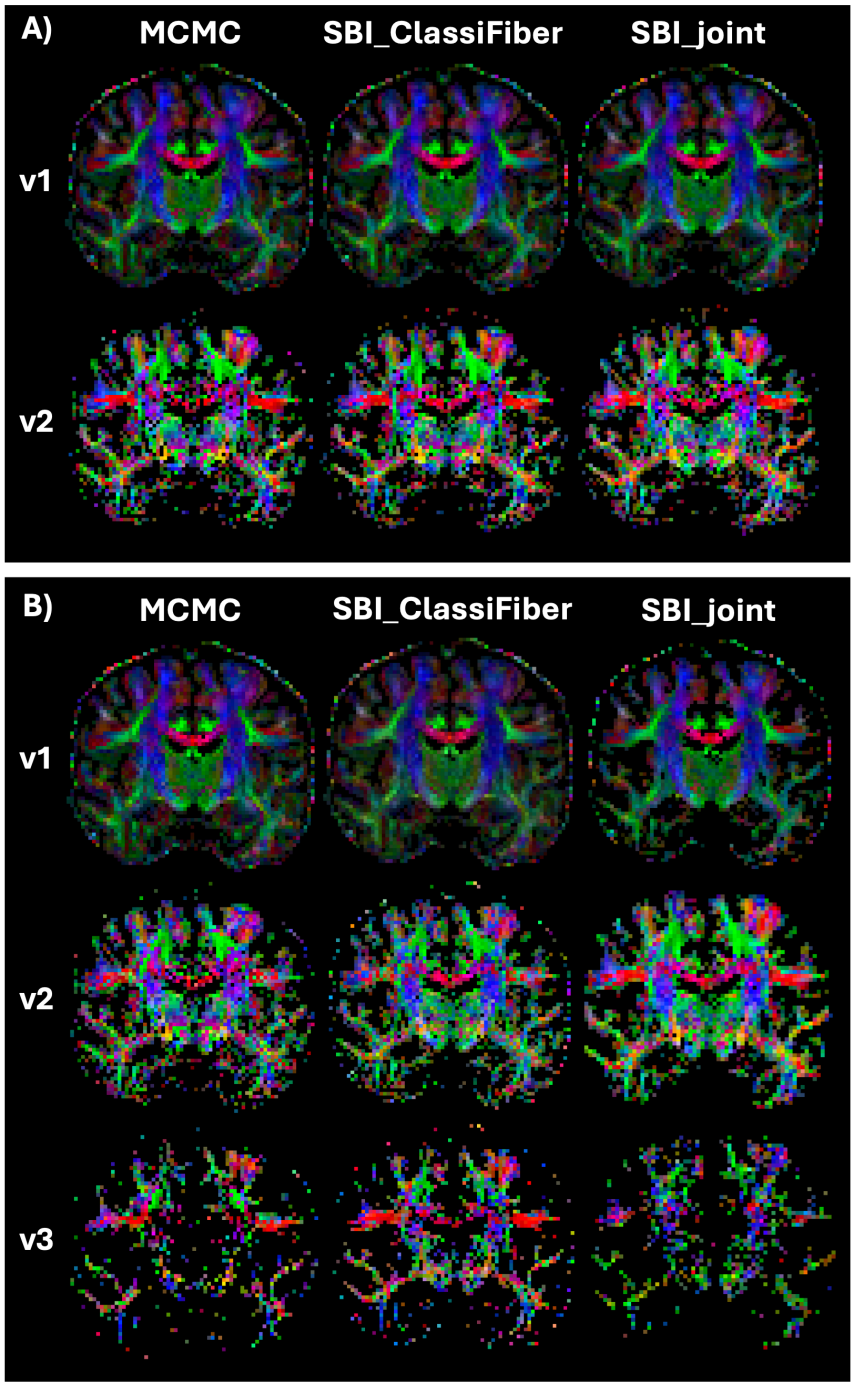

Figure S 5: **Maps of fiber orientation estimates (mean orientations of the posterior distributions) in in-vivo brain data obtained using MCMC, SBI_ClassiFiber and SBI_joint.** A) Multi-shell 2-fibres model. B) Multi-shell 3-fibres model. In SBI, all samples have been corrected to avoid compartment swapping, as detailed in Figure 8.

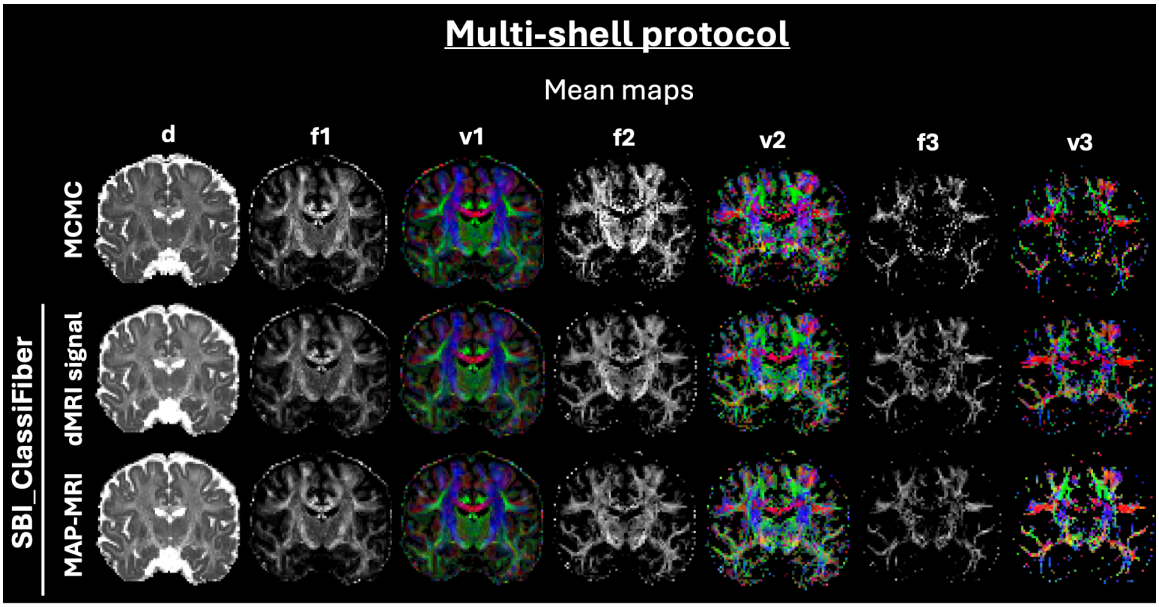

Figure S 6: **Maps of mean posterior estimates in in-vivo data from SBI trained against the raw dMRI signal (acquisition-specific) or a basis representation of the signal (acquisition=agnostic)**, for $N = 3$ compartments. MCMC estimates are shown for reference at the top row. Multi-shell data have been used and a MAP-MRI basis. NPEs are trained with 4 million samples and with the SBI_ClassiFiber architecture.