# Generative sampling in bundle tractography using autoencoders (GESTA)

Jon Haitz Legarreta, Laurent Petit, Pierre-Marc Jodoin, Maxime Descoteaux

# GENERATIVE SAMPLING IN TRACTOGRAPHY USING AUTOENCODERS (GESTA)

PREPRINT. PAPER UNDER REVIEW

**Jon Haitz Legarreta**
Sherbrooke Connectivity Imaging Laboratory (SCIL)
Videos & Images Theory and Analytics Laboratory (VITAL)
Department of Computer Science, Université de Sherbrooke, Canada

**Laurent Petit**
Université Bordeaux, CNRS, CEA, IMN, GIN, UMR 5293, F-33000 Bordeaux, France

**Pierre-Marc Jodoin***
Videos & Images Theory and Analytics Laboratory (VITAL)
Department of Computer Science, Université de Sherbrooke, Canada
Imeka Solutions inc., 195 rue Belvédère Nord, Sherbrooke (Québec) J1H 4A7, Canada

**Maxime Descoteaux***
Sherbrooke Connectivity Imaging Laboratory (SCIL)
Department of Computer Science, Université de Sherbrooke, Canada
Imeka Solutions inc., 195 rue Belvédère Nord, Sherbrooke (Québec) J1H 4A7, Canada

*Co-senior author. These authors contributed equally.

April 26, 2022

## ABSTRACT

Current tractography methods use the local orientation information to propagate streamlines from seed locations. Many such seeds provide streamlines that stop prematurely or fail to map the true pathways because some white matter bundles are "harder-to-track" than others. This results in tractography reconstructions with poor white and gray matter spatial coverage. In this work, we propose a generative, autoencoder-based method, named GESTA (*Generative Sampling in Tractography using Autoencoders*), that produces streamlines with better spatial coverage. Compared to other deep learning methods, our autoencoder-based framework is not constrained by any prior or a fixed set of bundles. GESTA produces new and complete streamlines for any white matter bundle. GESTA is shown to be effective on both synthetic and human brain *in vivo* data. Our streamline evaluation framework ensures that the streamlines produced by GESTA are anatomically plausible and fit well to the local diffusion signal. The streamline evaluation criteria assess anatomy (white matter coverage), local orientation alignment (direction), geometry features of streamlines, and optionally, gray matter connectivity. The GESTA framework offers considerable gains in bundle coverage using a reduced set of seeding streamlines with a 1.5x improvement for the "Fiber Cup", and 6x for the ISMRM 2015 Tractography Challenge datasets. Similarly, it provides a 4x white matter volume increase on the BIL&GIN callosal homotopic dataset. It also successfully generates new streamlines in poorly populated bundles, such as the fornix and other hard-to-track bundles, on *in vivo* data. GESTA is thus the first deep tractography generative method that can improve white matter reconstruction of hard-to-track bundles.

***Keywords*** Representation Learning · Autoencoder · diffusion MRI · Tractography · Generative Networks · Anatomical Reliability

# 1 Introduction

White matter (WM) brain tractography has become an essential tool to study structural connectivity and track-specific properties in a wide range of applications. Tractography is the computational process of integrating local fiber orientation reconstructions of the white matter into long-range pathways reaching the gray matter (GM). Most commonly, such mapping is done through streamline propagation methods: given a map of local orientations, such as a voxel-wise fiber Orientation Distribution Function (fODF) estimated from diffusion Magnetic Resonance Imaging (dMRI) data, a map of initial locations (seeds), and possibly some anatomical constraints, continuous fiber trajectories are reconstructed using a numerical integration method (Jeurissen et al., 2019; O'Donnell et al., 2019). The result of this process is a "tractogram", composed of a set of three-dimensional curves, called "streamlines", estimating the white matter fiber pathways.

Conventional streamline propagation methods can either be deterministic or probabilistic, depending on whether they assume a unique fiber orientation in each voxel, or else, whether a distribution of possible trajectories are allowed at each location (Descoteaux et al., 2009). A number of other methods, including global optimization approaches (Reisert et al., 2011), particle filtering methods (Girard et al., 2014), or surface-enhanced (St-Onge et al., 2018, 2021) have emerged to try to adequately reconstruct the white matter pathways. A summary of these methods can be found in Jeurissen et al. (2019) and O'Donnell et al. (2019). More recently, deep learning-based tractography methods have been proposed as an alternative to conventional methods (Poulin et al., 2019).

Tractography methods have been shown to suffer from a number of shortcomings that become apparent as a trade-off between sensitivity and specificity (Maier-Hein et al., 2017). Many of the current tracking pipelines still miss to extract streamlines for some fiber pathways, or provide solutions that are later discarded because they do not represent plausible streamlines. Several works Rheault et al. (2020); Schilling et al. (2019, 2021b); Zhang et al. (2022) have studied the challenges faced by tractography methods, such as propagating streamlines through hard-to-track regions, which results in streamline groups (bundles) that show a poor volume occupancy and are relatively under-represented, among others (Jeurissen et al., 2019). Incorporating bundle-specific priors (anatomical, orientational, amongst others) to tractography was proposed in (Chamberland et al., 2017; Rheault et al., 2019) as a way to increase the likelihood of reconstructing streamlines in hard-to-track regions.

Following our previous work using autoencoders to filter implausible streamlines from a whole-brain tractogram, we show that the autoencoder can be used to generate new and complete streamlines to better fill-up the white matter, especially the hard-to-track regions and bundles. Our contributions are twofold: (i) a generative, autoencoder-based method for extracting new streamline candidates from the latent space; and (ii) an evaluation framework to asses the anatomical plausibility and dMRI signal fit of the generated streamline data. We show that our approach can be successfully applied to both synthetic and human brain *in vivo* data with extensive experiments. Our generative tractography method yields anatomically plausible streamlines that can be used to improve the spatial coverage of hard-to-track bundles or hard-to-extract clusters of streamlines. The proposed anatomical plausibility framework ensures that the provided streamlines comply with the required geometrical, brain tissue occupancy, diffusion signal fit, and connectivity features. To the best of our knowledge, GESTA (*Generative Sampling in Tractography using Autoencoders*) is the first deep generative tractography method.

## 1.1 Related work

A number of works have proposed ways to improve the results of tractography methods to provide a better white matter spatial coverage and/or cortical coverage, either using bundle-specific or whole-brain approaches.

Chamberland et al. (2014) proposed to iteratively improve the output of a tracking method by interactively changing the tractography parameters. By selectively seeding regions of interest, such as those with a poor spatial coverage, and adjusting the streamline propagation parameters, authors showed that their method produces tractograms with an improved white matter occupancy. The same authors later introduced a method (Chamberland et al., 2017) to improve tracking of the optic radiation by modifying the streamline propagation equation according to orientational priors within some given regions of interest. Bundle-specific tractography (Rheault et al., 2018, 2019) requires a set of bundle templates to scale the fiber orientation distributions accordingly, and employs a multi-parametric approach to extract streamlines in a bundle-wise fashion. In Poulin et al. (2018) authors used a different deep recurrent neural network to reconstruct streamlines for each of the considered bundles.

Whole-brain tractography strategies have been introduced to offer generalized solutions to enhance the reconstruction of hard-to-track bundles. In Battocchio et al. (2020, 2021), the authors presented a method to improve the WM spatial coverage of a tractogram by re-parameterizing the existing set of streamlines using B-spline functions and computing new streamline trajectories using Markov chain Monte Carlo (MCMC) methods. St-Onge et al. (2021) proposed

to seed the WM/GM interface in an adaptive manner, depending on the GM local features. They also proposed to optionally dynamically (iteratively) add seeds in regions presenting a low streamline endpoint density to provide a new tractogram that features an improved spatial and cortical surface coverage. Although this strategy excels in improving the surface coverage, it relies on the surface data availability, and is still hindered by the limitations in the underlying local tracking procedure, which might still track preferably along given orientations in the deep white matter.

Numerous methods using neural networks, based on optimizing the propagation direction predictions according to a loss function, have been proposed for whole-brain tractography in recent years (Poulin et al., 2019). Similarly, deep reinforcement learning has been applied to whole-brain tractography (Théberge et al., 2021) as another choice to avoid detrimental local detours in streamline propagation. In this case, choosing the most appropriate streamline propagation step is learned according to the values of a reward function.

In the broader context of diffusion MRI, deep generative models have emerged as a method capable of successfully performing image synthesis and super-resolution. Adversarial methods have been proposed to synthesize diffusion data (or its derivatives) using structural data (e.g. (Anctil-Robitaille et al., 2020)), or to provide high-resolution diffusion MRI from low-resolution images (e.g. (Luo et al., 2022)). These methods employ an *adversarial* discriminator to iteratively allow the network to improve the generated data so as to increase its anatomical reliability. Instead of using a discriminator network, autoencoder-based generative methods with explicit anatomical constraints have also been used successfully in cardiac image segmentation tasks (e.g. (Painchaud et al., 2020)). However, to the best of our knowledge, deep generative models have not been employed to generate a tractography product yet.

In this work, we propose a generative, autoencoder-based tractography approach that is not restricted by the underlying tracking procedure, and that can be readily applied to both synthetic and human brain *in vivo* data. Compared to other solutions, our method (i) uses a single model to yield new and complete streamlines; (ii) it does not involve iterative optimizations of the seeding strategy or streamline trajectories; and (iii) it does not require locally propagating an orientational field. Our procedure works by generating new streamline candidates for a bundle of choice in the representation space of an autoencoder, and evaluating the anatomical plausibility of the streamlines to accept or to discard them. Our anatomical plausibility evaluation framework includes structural and connectivity constraints, streamline geometry properties, and fixel-based features (Raffelt et al., 2015).

## 2 Material and methods

We leverage the FINTA autoencoder architecture presented in Legarreta et al. (2021) to introduce a generative tractography method, GESTA (*Generative Sampling in Tractography using Autoencoders*). In our former FINTA work, the proposed convolutional deep neural autoencoder network was aimed at filtering streamlines in tractography. In GESTA, the autoencoder does not need to be retrained to accomplish the generative task. The learned latent space is re-used as a "streamline yard" to produce new streamlines through a bundle-wise sampling process in GESTA. The sampled latent vectors are then decoded into streamline space, and once their anatomical plausibility established, a new, anatomically plausible (generative) tractogram is provided.

The autoencoder is trained to minimize the mean squared-error loss between the input streamlines and their reconstructions at the output. It is assumed that the training process allows the autoencoder's latent space to hold a (statistically) meaningful representation of the input data points (Bengio et al., 2013). GESTA takes advantage of such representational space to estimate new points within the data distribution.

GESTA uses a sampling strategy and a streamline acceptance/rejection procedure. A sampling strategy of choice probes new streamline candidates in the latent space using a set of available streamlines as reference, hereinafter named *seed streamlines*. Due to the relatively high dimensionality of the latent space and the uncertainty of the sampling procedure, anatomically implausible streamlines can be generated when probing the latent space. Thus, the newly produced latent samples are decoded into the streamline space, and their features are evaluated according to a number of anatomical, geometrical, diffusion signal fit, and connectivity criteria to determine their plausibility. Given the streamlines of a tractogram as the input, the procedure works as follows:

1. Train once an autoencoder using raw, unlabeled streamlines.
2. Select a subset of seed streamlines of a poorly covered bundle that will serve as the reference for the sampling procedure.
3. Project the set of seed streamlines to the latent space of the trained autoencoder.
4. Use a sampling procedure to yield new samples in the latent space.
5. Decode the new streamline samples.

6. Compute the streamlines' features, and keep the anatomically plausible streamlines.

Note that the autoencoder needs to be trained only once and remains unaltered to perform all considered tasks (including those performed in Legarreta et al. (2021)). Input streamlines are resampled to contain 256 equidistant vertices, and the same 32-dimensional representational latent space dimensionality and network parameters are used as previously done in FINTA (Legarreta et al., 2021).

Figure 1 depicts the GESTA pipeline applied to a subject at test time. At first, streamlines are brought to a common, standard space (the same space used to train the autoencoder). The learned latent space of the already trained autoencoder is used to generate new streamlines for each fascicle of interest (e.g. to fill the spatial coverage of bundles). The plausibility of the generated streamlines is evaluated using different criteria, including WM occupancy (anatomy), local orientation alignment (direction), geometry, and GM occupancy (connectivity) features.

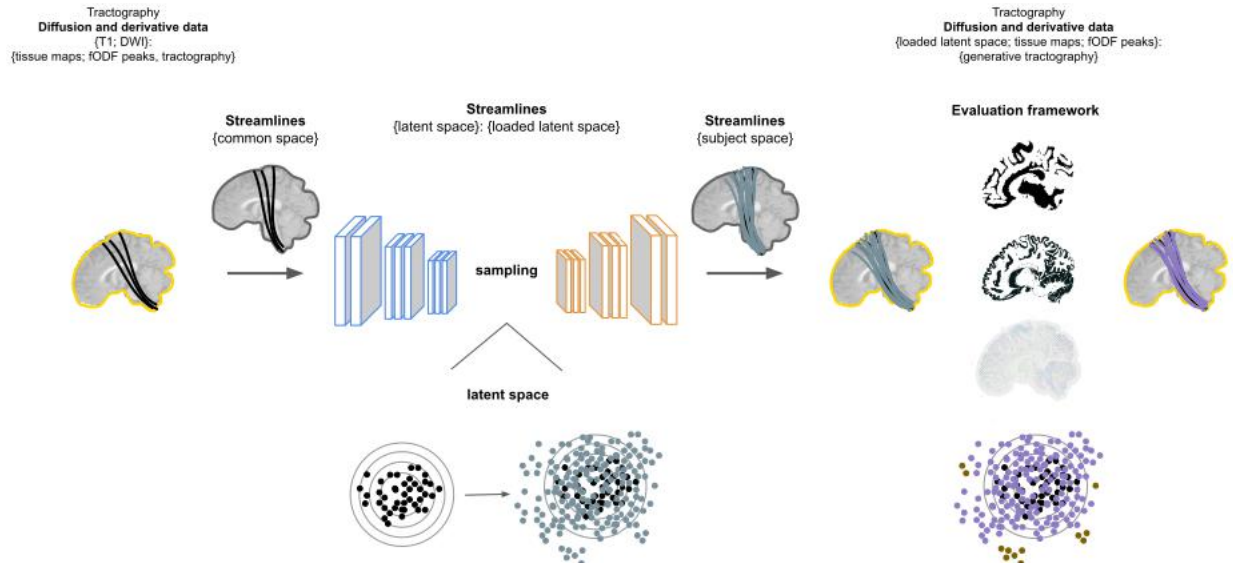



Figure 1: Illustration of the GESTA (*Generative Sampling in Tractography using Autoencoders*) generative tractography pipeline using autoencoders. The streamlines of a given subject are put into a common space, such as the MNI-space for example. Given a set of seed streamlines of some bundle (black color streamline and dots) at the input, the learned latent space of a trained autoencoder holds a meaningful representation of the input streamlines. Such data points are assumed to embody some unknown distribution. A sampling method applied to the subject- and bundle-wise estimated distribution is used to generate new streamline candidates (gray color streamlines and dots). The latent samples, decoded and brought back to the subject space, are accepted (purple streamlines and dots) or rejected (dark gold dots) depending on some plausibility criteria involving anatomical and diffusion-related constraints. Note that the autoencoder training procedure is not shown in the figure.

### 2.1 Latent space sampling

GESTA uses the rejection sampling method to generate new observations (i.e. latent vectors) from the unknown distribution of the (encoded) seed streamlines of a bundle. Given a hard-to-sample distribution $p(z)$, rejection sampling uses a simpler proposal distribution $q(z)$, which is easier to sample from, in order to generate values from the original distribution. It assumes that a constant $k$ exists such that $kq(z) \geq p(z)$, $\forall z$. First, a $z_0$ value is obtained from the distribution $q(z)$. A number $u_0$ is then generated from the uniform distribution over $[0, kq(z_0)]$. If $u_0 > p(z_0)$, the sample is rejected; otherwise, $u_0$ is accepted. In our case, we do not know the $p(z)$ distribution of our seed streamlines in the latent space, and thus, we estimate it using a Parzen estimator. In our experiments, $q(z)$ is set to be a multivariate Gaussian distribution with mean and variance values estimated from the seed data.

Figure 2 shows the streamlines sampled using the autoencoder-based generative tractography method as their t-SNE projection (van der Maaten and Hinton, 2008) in the latent space for the "Fiber Cup" dataset (see section 2.3.1) bundles 1 and 7 (following the numbering in Côté et al. (2013)). Using a few seed streamlines for each bundle (shown in black), the latent space sampling retrieves streamlines that are issued by the same distribution. Once their fit to anatomical and diffusion features has been determined, these generative samples are considered as reliable streamlines for all purposes and intents.

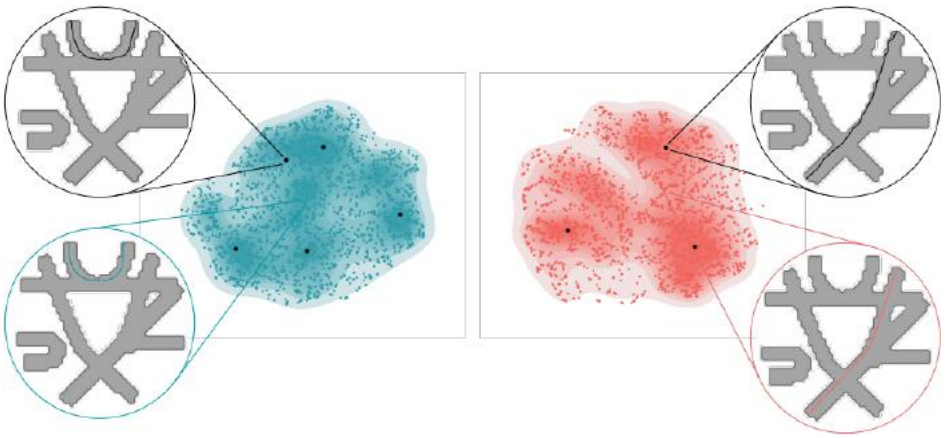


Figure 2: t-SNE projection of the autoencoder-based generative streamlines corresponding to the "Fiber Cup" dataset bundles 1 (turquoise) and 7 (red). 3% of the test set streamlines' were used to seed each bundle, and the data distribution was estimated using a bandwidth of 1.0 (see section 3.1). 2000 samples were generated for each bundle. The generated streamlines are shown as colored dots; black dots represent the seed streamlines.

### 2.2 Streamline evaluation framework

The latent space sampling procedure yields a number of latent vectors $N$. These latent vectors constitute streamline candidates whose anatomical plausibility needs to be established to yield the final generative tractogram. The streamline evaluation framework accepts or discards the latent sampled streamlines depending on a number of features and some given thresholds. The following anatomical reliability criteria are established for streamline acceptance:

- *ADG* (*Anatomy, Direction, Geometry*): constrains the streamlines' geometry-, fixel-, and WM occupancy-related features. The geometry features are determined in terms of the streamlines' length and winding. The streamline-to-fixel compliance is measured in terms of the streamline's local orientation to fODF peak alignment cone. The local orientation is computed as the orientation vector between two consecutive streamline vertices, and the corresponding fODF peak values are interpolated at each of the vertices. Additionally, streamlines are required to be located within the WM tissue. Alternatively, streamlines' WM occupancy compliance can be softened to be set as the ratio of the streamline vertices that lie in the WM tissue over the total number of streamline vertices.
- *ADGC* (*Anatomy, Direction, Geometry, Connectivity*): in addition to the above criteria, evaluates the degree at which streamlines reach the GM (or the voxels that simulate the expected terminations in the case of synthetic datasets).

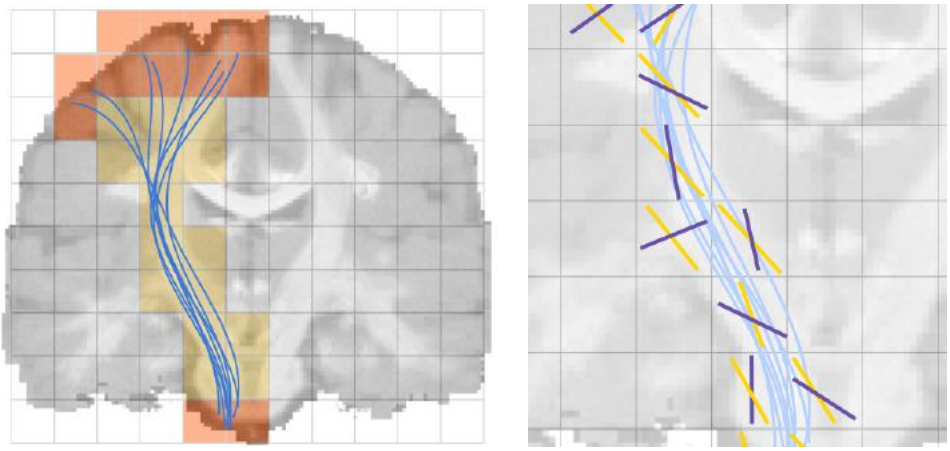
Figure 3 depicts the features involved in the streamline evaluation framework.

### 2.3 Data

GESTA is tested across a variety of datasets and target bundles. Our use cases include bundles from (i) synthetic tractography data ("Fiber Cup" dataset); (ii) clinical-style realistic data (ISMRM 2015 Tractography Challenge dataset); (iii) whole-brain *in vivo* human tractography data (HCP dataset); and (iv) multi-subject, partial tractography data (BIL&GIN callosal homotopic data). Particularly, the BIL&GIN callosal homotopic data is chosen to showcase (i) how the generative tracking framework can sample the white matter pathways using a different bundling organization without modification; and (ii) how the generative framework can improve the spatial coverage of bundles that might be affected after a tractography processing step.

#### 2.3.1 "Fiber Cup"

A synthetic "Fiber Cup" dataset generated using Fiberfox (Neher et al., 2014) is used to mimic the original "Fiber Cup" phantom (Côté et al., 2013; Fillard et al., 2011). This "Fiber Cup" phantom contains 7 bundles, and the ground truth consists of a total of 7833 streamlines. The raw diffusion data were generated using 30 gradient directions, a diffusion gradient strength of $1000\,\mathrm{s\,mm^{-2}}$, and a $3\,\mathrm{mm}$ isotropic spatial resolution for a $64 \times 64 \times 3$ volume.



**(a)** **(b)**

Figure 3: Streamline evaluation criteria. (a) Anatomy (WM occupancy) and connectivity (GM occupancy); (b) Direction (local orientation alignment). Yellow color depicts voxels belonging to the WM (anatomy), and orange designates voxels belonging to GM tissue (connectivity). The local orientation is depicted as yellow and purple sticks corresponding to different fODF peaks. Only 2 fODF peaks are shown in each voxel for illustrative purposes. The streamline geometry criteria (namely, the length and the winding) are not shown. The *ADG* criterion evaluates the streamlines' WM occupancy (anatomy), the local orientation alignment (direction), and geometry features; the *ADGC* adds to these the GM occupancy (connectivity) criterion.

#### 2.3.2 ISMRM 2015 Tractography Challenge

The ISMRM 2015 Tractography Challenge dataset of Maier-Hein et al. (2015) is used as a human brain model quantitative testbed. The dataset consists of a clinical-style, realistic single subject tractogram (approximately 200 000 ground truth streamlines) with 25 ground truth fiber bundles, raw diffusion-weighted data and a structural T1-weighted MRI volume generated using Fiberfox (Neher et al., 2014). The raw dMRI data were generated at a $2\,\mathrm{mm}$ isotropic spatial resolution, with 32 gradient directions, and a b-value of $1000\,\mathrm{s\,mm^{-2}}$. According to Maier-Hein et al. (2017), the dataset contains 18 bundles that are considered hard-to-track or very hard-to-track.

#### 2.3.3 Human Connectome Project (HCP) data

A subject from the Human Connectome Project (HCP) dataset (Glasser et al., 2016) is used to perform whole-brain generative tractography on an *in vivo* dataset. The data were acquired using an HCP scanner equipped with high-end hardware enabling diffusion encoding gradient strengths of $100\,\mathrm{mT\,m^{-1}}$ at $1.25\,\mathrm{mm}$ isotropic spatial resolution along 270 gradient directions. The available preprocessed data is used here.

#### 2.3.4 BIL&GIN callosal homotopic data

39 subjects of the BIL&GIN (*Brain Imaging of Lateralization* by the *Groupe d'Imagerie Neurofonctionnelle*) human brain dataset (Chenot et al., 2019; Mazoyer et al., 2016) are used to demonstrate the autoencoder-based generative tractography framework on clinical-like real data. Acquisitions were done with a Philips Achieva $3\,\mathrm{T}$ MR scanner using 21 non-colinear diffusion gradient directions and a diffusion gradient encoding strength of $1000\,\mathrm{s\,mm^{-2}}$ with four averages and an isotropic spatial resolution of $2\,\mathrm{mm}$. In this work, only the corpus callosum is considered, and within the bundle, homotopic streamlines lying within any of 26 pairs of gyral-based segments are considered as plausible streamlines. Readers are referred to Legarreta et al. (2021) for the details about the gyral-based callosal segments used for the extraction of the streamlines. The number of segments that have a mean streamline count larger than 5% of the mean homotopic streamline count across subjects is 3. Hence, most segments can be considered as poorly populated and very hard-to-track.

## 3 Experiments

Conventional streamline propagation methods were used on all datasets to obtain the tractography data for this work. Local probabilistic tracking was used for the "Fiber Cup" and ISMRM 2015 Tractography Challenge datasets; the

HCP tractogram was reconstructed using global tracking (Reisert et al., 2011); and the BIL&GIN callosal homotopic tractograms were obtained using the probabilistic setting of the Particle Filtering Tracking (PFT) (Girard et al., 2014) method. Further details relevant to the tracking parameters can be found in our previous work Legarreta et al. (2021).

The HCP data were registered (Avants et al., 2011) to the ISMRM 2015 Tractography Challenge dataset space. The BIL&GIN callosal homotopic tractograms were registered to the MNI template common space using the MNI152 2009 standard-space T1-weighted average structural $1\,\mathrm{mm}$ isotropic resolution template image (Fonov et al., 2011).

The autoencoder trained on each of the mentioned data is maintained unmodified with respect to Legarreta et al. (2021) for the upcoming experiments.

Appropriate bundles were extracted from each dataset's test set of anatomically plausible streamlines to be seeded separately in GESTA. For the "Fiber Cup" dataset, the bundle assignments were obtained using the ground truth bundle endpoints. The bundles for the ISMRM 2015 Tractography Challenge dataset were obtained using the Tractometer tool (Côté et al., 2013; Maier-Hein et al., 2017). The Tractometer tool was similarly employed to obtain the bundles of interest for the HCP dataset, as the data had been registered to the ISMRM 2015 Tractography Challenge data space. Finally, the BIL&GIN callosal homotopic data were segmented from the FINTA-determined plausible streamlines according to twenty-six (26) gyral homotopic regions of interest of the JHU template (Zhang et al., 2010). Section A.2 summarizes the bundle and segment naming used in this work.

To demonstrate GESTA's ability to improve the white matter spatial coverage of a tractogram, we subsample the streamlines in the mentioned datasets' test sets at different rates, and measure the bundle overlap (or the white matter volume) provided by the generative tractogram. Starting from a *complete* ($100\%$ set) of plausible streamlines for each bundle, different subsets are generated ($P\%$ sets) by bundle-wise subsampling to serve as seeds. The generative tractography framework is then run to extract new, complete, plausible streamlines from such seeds without requiring any iterative propagation procedure to induce an improvement on the measured feature.

Given a trained autoencoder, and the aforementioned complete streamline data, the experiments were completed as follows:

1. For each bundle, a set of streamlines is selected randomly to serve as the seeds in the latent space. The ratio of the selected streamlines is varied in the set $P = \{3,5,10,100\}\%$ for the "Fiber Cup" and ISMRM 2015 Tractography Challenge dataset to test the generative tractography procedure under different hard-to-track conditions across all bundles. All available test set streamlines ($P = 100\%$) are used as seeds in the HCP and BIL&GIN callosal homotopic dataset experiments.
2. New streamline candidates are generated from the latent space using the rejection sampling algorithm.
3. The generated latent samples are decoded and are accepted/rejected using an anatomical plausibility evaluation framework.

The numerical measures used to grade the white matter spatial coverage improvement offered by GESTA are described in section 3.3.

### 3.1 Sampling parameters

The Parzen estimator used in the rejection sampling method was allowed a degree of freedom by using different bandwidth factor values to vary the span of the explored latent space to some degree. The values used ranged from $1.0$ to $10.0$: as a general rule, smaller values result in a weaker sampling performance (both in terms of the necessary time to accept a sample and the wealth of streamlines generated), and larger values result in a large proportion of streamlines that are not fitting the seed streamlines, and hence yield excessively low plausibility ratios. Larger values do not provide significantly better bundle overlap scores either, with only a slight improvement on the "Fiber Cup" dataset. Thus, the parameter is fixed to a value of $1.0$ in all experiments.

A total number of $N = 2000$ streamlines are sampled for each the "Fiber Cup" bundles; $N = 15\,000$ streamlines are generated for each bundle of the ISMRM 2015 Tractography Challenge and the HCP datasets; and the value is fixed to $N = 6000$ for the BIL&GIN callosal homotopic dataset, prior to the plausibility assessment.

### 3.2 Streamline plausibility thresholds

To consider as plausible any streamline generated from the latent space, a set of thresholds are established for the *ADG(C)* (Anatomy, Direction, Geometry, Connectivity) criteria introduced in section 2.2. The thresholds are set as sensible values in tractography practice. For the sake of simplicity, the constraints that required numerical thresholding

values are set to the same values across all bundles for the same dataset. The *ADG* constraints are considered as a fundamental set of features, and are also required to fully accept a streamline under the *ADGC* criterion.

Due to the nature of the "Fiber Cup" dataset (which does not suffer from any partial volume effect), the WM occupancy is set as a hard, binary requirement. The criterion is relaxed to a soft requirement for the rest of the datasets, where $95\%$ of the streamline vertices are required to be in the WM tissue. See section 5.2 for a more comprehensive discussion. Table 1 summarizes the values used for each dataset.

Table 1: Plausibility criteria values. LOA: local orientation angle; WM: white matter occupancy; T/F: binary requirement (true/false). Note that *ADGC* includes the *ADG* criteria.

| | | **"Fiber Cup"** | **ISMRM 2015** | **HCP** | **BIL&GIN callosal** |
|---|---|---|---|---|---|
| *ADG* | Length ($\mathrm{mm}$) | 20–220 | 20–220 | 20–220 | 20–220 |
| | Winding (°) | $<330$ | $<330$ | $<340$ | $<360$ |
| | LOA-to-fODF peak (°) | $<30$ | $<30$ | $<40$ | $<40$ |
| | WM | T/F | - | - | - |
| | WM ratio (%) | - | 95 | 95 | 95 |
| *ADGC* | GM | T/F | T/F | T/F | T/F |

The streamline evaluation framework using the WM occupancy binary constraint is designated as $ADG_B$ (and $ADGC_B$); $ADG_R$ (and $ADGC_R$) is used when using the WM occupancy ratio hereafter. Readers are referred to section A.1 for further details concerning the evaluation framework parameterization.

### 3.3 White matter coverage improvement quantification

The performance of the GESTA method is graded according to the bundle volume overlap (OL). The OL measures the ratio of voxels occupied within the volume of a ground truth bundle traversed by at least one plausible streamline associated with the bundle. The overlap allows to quantitatively measure the white matter spatial coverage improvement induced by the generative tractography framework. The overlap measure is computed for the "Fiber Cup" and ISMRM 2015 Tractography Challenge datasets, as ground truth bundle models are required. For each $P = \{3,5,10,100\}\%$ seed streamline ratio value, the bundle volume overlap of the reconstructed tractogram is compared to the generative tractogram for the corresponding ratio.

Similarly, the overlap is used for quantitative comparison purposes on the HCP dataset. Note that the HCP data had been registered to the ISMRM 2015 Tractography Challenge dataset space in Legarreta et al. (2021); as no re-training is done in this work, it naturally follows measuring the performance in terms of the overlap with the ISMRM 2015 Tractography Challenge dataset models.

For the BIL&GIN callosal homotopic dataset, the volume (in $\mathrm{mm}^3$) occupied by the streamlines (both the seed and the latent-generated ones) is measured and compared.

### 3.4 Corpus callosum considerations

The dissected corpus callosum in the ISMRM 2015 Tractography Challenge and HCP datasets are further split into $6$ regions (CC_Fr_1; CC_Fr_2; CC_Oc; CC_Pa; CC_Pr_Po; CC_Te) according to the atlas used by the RecobundlesX bundling method Rheault (2021) to test GESTA. Only the streamlines recognized as belonging to these regions are kept. In practice, this approach allows the sampling method to effectively generate observations in the corpus callosum system.

Due to both a poor quality under visual inspection and an insufficient amount of streamlines available for the sampling process (see section 5.4), the CC_Te streamlines were discarded. Thus, no generative streamline is extracted for the temporal section of the corpus callosum. Finally, note that the BIL&GIN callosal homotopic dataset is split into different callosal segments by definition, and uses another delineation atlas, as mentioned in section 2.3.4.

## 4 Results

First, we show how GESTA can generate streamlines on the "Fiber Cup" and ISMRM 2015 Tractography Challenge datasets close to completing their ground truth coverage using only a small proportion of seed streamlines. GESTA's performance using the full set of available seed streamlines (still relatively low for many bundles and segments) is analyzed thereafter for the HCP and BIL&GIN callosal homotopic datasets.

### 4.1 "Fiber Cup"

Table 2 shows the improvement in the white matter spatial coverage measured in terms of the bundle volume overlap (OL) using GESTA for the "Fiber Cup" dataset. The seed ratio represents the ratio of streamlines used as seeds in the latent space over the total number of available streamlines for a given bundle in the test set. Results are averaged across bundles. Results show that the proposed generative tractography method is able to raise the overlap measure using a very limited set of seed streamlines: for the $P = 3\%$ case, it is increased by $1.5$ times. As more seed streamlines become available, the overlap provided by the latent-generated streamlines increases. The bundle overlap measures are maintained even when the plausibility requirements become more demanding using the $ADGC_B$ criterion. This reveals that prematurely terminated streamlines might be avoided by GESTA.

Table 2: "Fiber Cup" dataset reconstructed seed streamlines' and generative streamlines' overlap. The overlap is the ratio of voxels occupied within the volume of a ground truth bundle traversed by at least one plausible streamline associated with the bundle. $N = 2000$ streamlines are generated for each bundle with a bandwidth factor of value $1.0$, and the plausibility is evaluated using the $ADG_B$ and $ADGC_B$ criteria. Mean and standard deviation values across bundles.

| | | $ADG_B$ | $ADGC_B$ |
|---|---|---|---|
| **Seed ratio** ($P\%$) | **OL** | **OL** ($\uparrow$) | **OL** ($\uparrow$) |
| 3 | 0.57 (0.11) | 0.83 (0.16) | 0.83 (0.16) |
| 5 | 0.68 (0.1) | 0.9 (0.08) | 0.9 (0.08) |
| 10 | 0.76 (0.09) | 0.91 (0.05) | 0.91 (0.05) |
| 100 | 0.90 (0.07) | 0.94 (0.05) | 0.94 (0.05) |

Plausible streamlines generated for the "Fiber Cup" dataset bundles 1, 3 and 7 are shown in figure 4 together with their corresponding seed streamlines. Due to the fact that seed streamlines are chosen randomly, the picked seeds might show an unfavorable spatial distribution for the generative process. This is especially noticeable when the ratio of streamlines used is low (e.g. $P = 3\%$ in bundle 7). The generative framework then suffers from a limited ability to explore the latent space (for a given Parzen estimator bandwidth factor value), and the generated streamlines improve the bundle overlap to a lesser extent.

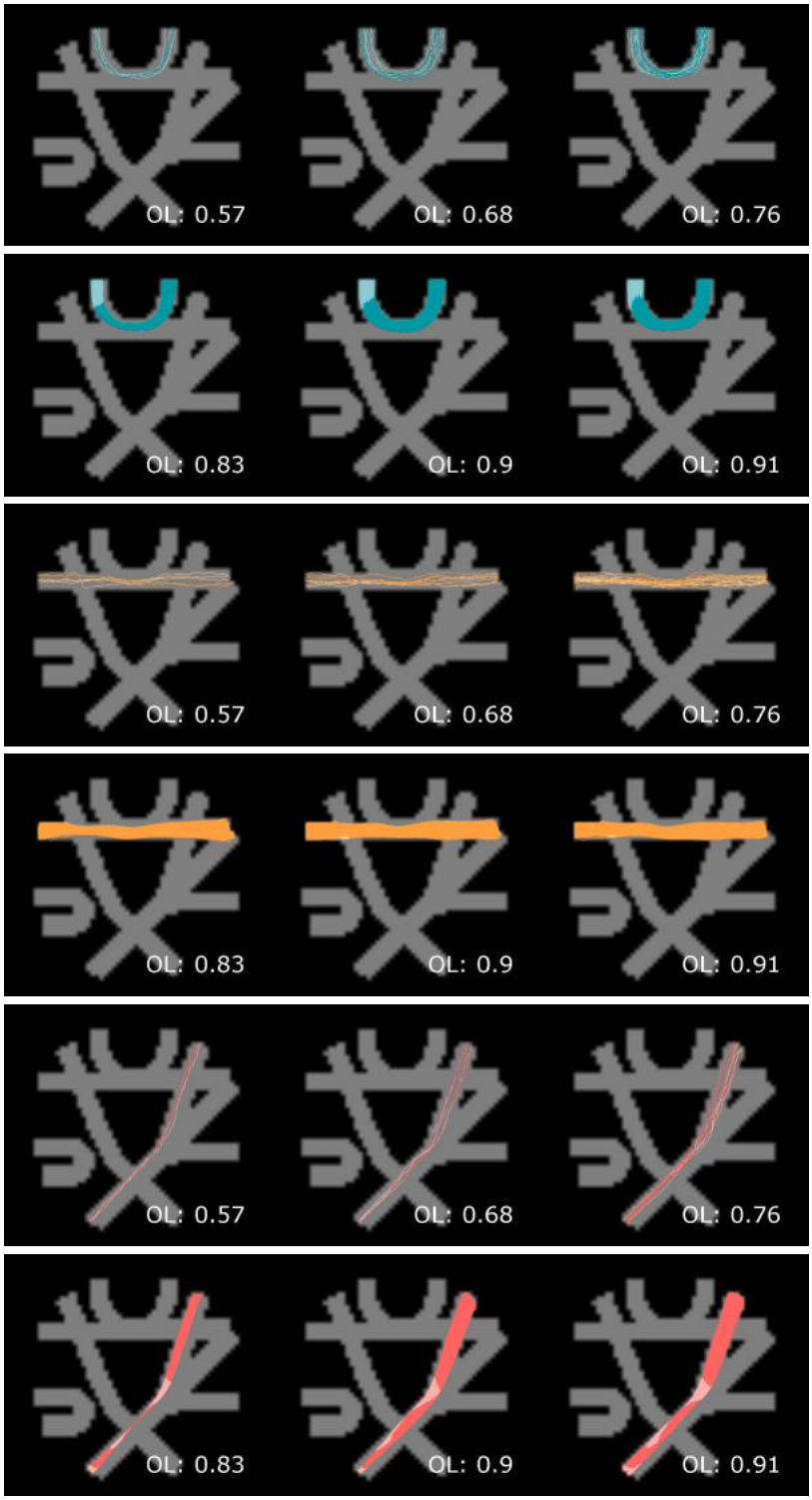



Figure 4: Latent space-generated streamlines for bundles 1 (turquoise), 3 (orange), and 7 (red) corresponding to "Fiber Cup" dataset. For each bundle: top: seed streamlines; bottom: latent-generated plausible streamlines. All axial superior views. The seed streamlines represent $P = 3\%$ (a), $P = 5\%$ (b), and $P = 10\%$ (c) of the test set streamlines for each bundle; and streamlines' plausibility is assessed following the $ADG_B$ criterion. Seed streamlines are depicted with a slightly larger diameter compared to the generative streamlines. The overlap legends refer to the mean figures across all bundles.

### 4.2 ISMRM 2015 Tractography Challenge

Table 3 shows the bundle overlap results (averaged across bundles) on the ISMRM 2015 Tractography Challenge dataset using GESTA. As it is the case with the "Fiber Cup" dataset, the generative tractography method improves remarkably the bundle spatial coverage: at the $P = 3\%$ lower-end seed ratio, the overlap is raised to $0.74$ from $0.13$ (a $5.7$x factor) for the $ADG_R$ criterion. The ISMRM 2015 Tractography Challenge dataset being a more complex one, as the plausibility requirements become more demanding using the $ADGC_R$ criterion, less streamlines qualify as plausibles, and thus a decrease in the bundle overlap is observed with respect to the $ADG_R$ criterion. Despite the streamline connectivity constraint limiting the gain, a value of $0.74$ is achieved for the same $P = 3\%$ seed ratio.

Table 3: ISMRM 2015 Tractography Challenge dataset reconstructed seed streamlines' and generative streamlines' overlap. $N = 15\,000$ streamlines are generated for each bundle with a bandwidth factor of value $1.0$, and the plausibility is evaluated using the $ADG_R$ and $ADGC_R$ criteria. Mean and standard deviation values across bundles.

| | | $ADG_R$ | $ADGC_R$ |
|---|---|---|---|
| **Seed ratio** ($P\%$) | **OL** | **OL** ($\uparrow$) | **OL** ($\uparrow$) |
| 3 | 0.13 (0.05) | 0.74 (0.15) | 0.65 (0.27) |
| 5 | 0.19 (0.07) | 0.78 (0.12) | 0.69 (0.23) |
| 10 | 0.29 (0.1) | 0.81 (0.09) | 0.7 (0.24) |
| 100 | 0.65 (0.12) | 0.84 (0.09) | 0.74 (0.24) |

Figure 5 shows the generative tractography process applied to the left CST, fornix and right SLF bundles of the ISMRM 2015 Tractography Challenge dataset for different seed streamline counts (from left to right, $P = \{3,5,10\}\%$). According to the ISMRM 2015 Tractography Challenge results (Maier-Hein et al., 2017), the right SLF bundle reconstruction was one whose difficulty was medium, whereas the left CST and fornix were considered to be hard. The recovered spatial coverage has considerably improved already using as little as $P = 3\%$ of the available streamlines in each bundle's test set. As more seed streamlines are available, the better coverage can be expected. However, as the seeds are added randomly in our experiments, the underlying data distribution might be shifted, and the sampling method might preferentially draw samples around some given seed streamlines.

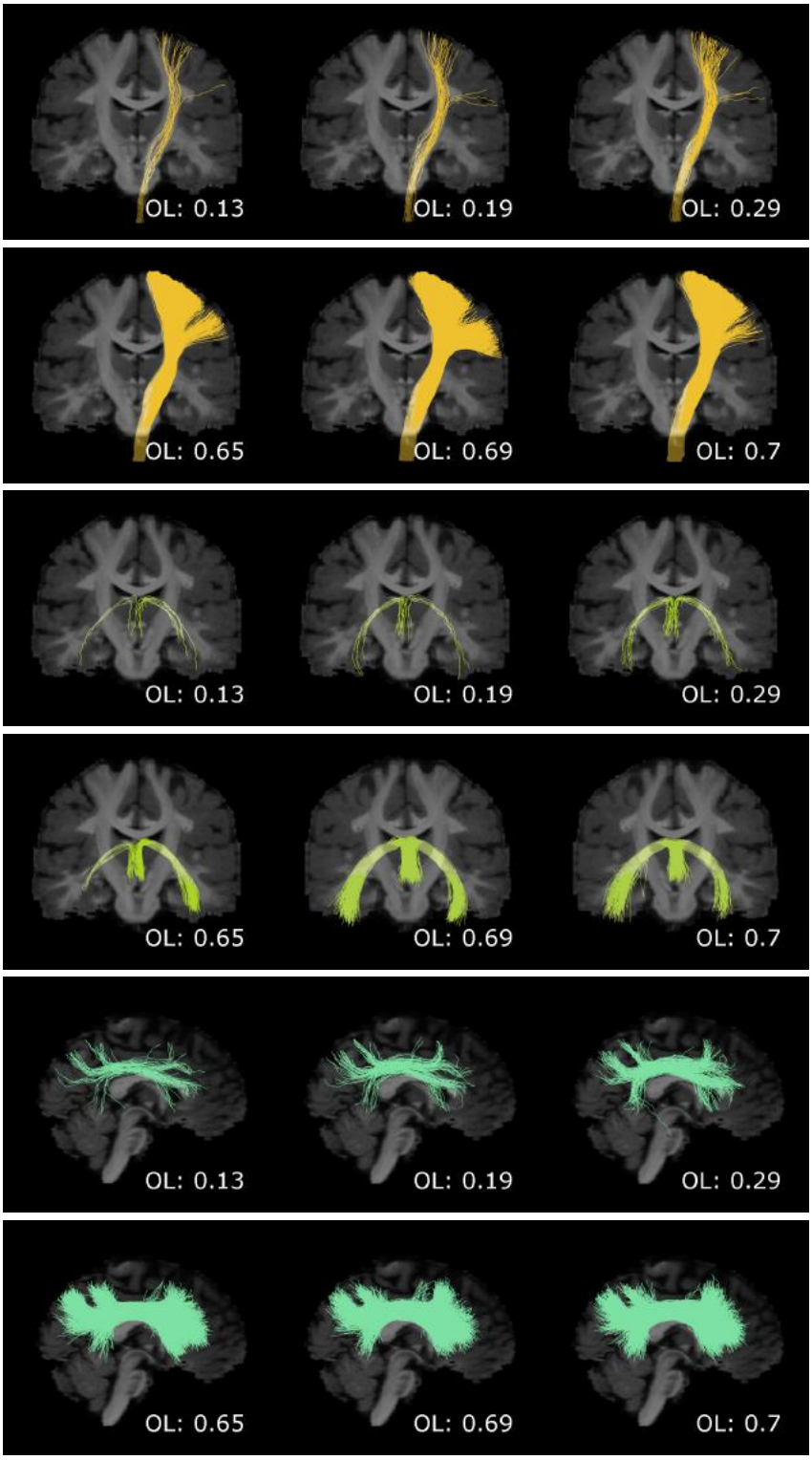
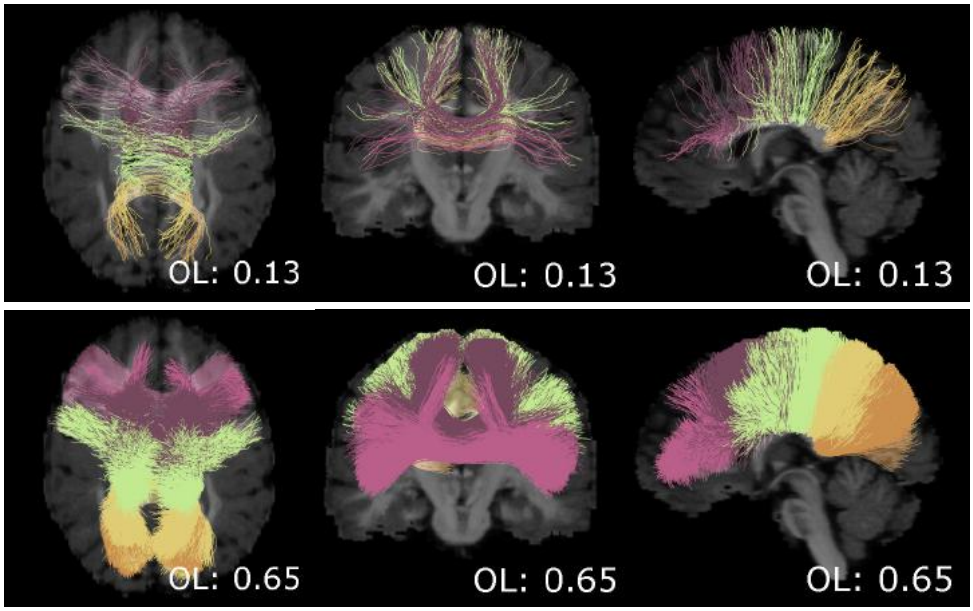



Figure 5: Generative tractography applied to the ISMRM 2015 Tractography Challenge dataset left CST (yellow orange color), fornix (lime), and right SLF (sea green) bundles. For each bundle: top: seed streamlines; bottom: latent-generated plausible streamlines. Coronal anterior (left CST and fornix) and sagittal right (right SLF) views. The seed streamlines represent $P = 3\%$ (a), $P = 5\%$ (b), and $P = 10\%$ (c) of the test set streamlines for each bundle; and streamlines are evaluated according to the $ADGC_R$ criterion. The overlap legends refer to the mean figures across all bundles.

Figure 6 shows the result of generating streamlines in the latent space for the ISMRM 2015 Tractography Challenge dataset corpus callosum. Using $P = 3\%$ of the available seeds for each segment, GESTA is able to successfully recover the potential white matter pathways of interest.




**(a)** **(b)** **(c)**

Figure 6: Generative tractography applied to the ISMRM 2015 Tractography Challenge dataset corpus callosum. Top row: seed streamlines; bottom row: latent-generated plausible streamlines. (a) Axial superior; (b) coronal anterior; (c) sagittal left views. The seed streamlines represent $P = 3\%$ of the corpus callosum test set streamlines; and streamlines are evaluated using the $ADGC_R$ criterion. Atlas region to color code: CC_Fr_1: pink; CC_Fr_2: purple; CC_Pr_Po: yellow green; CC_Pa: sand; CC_Oc: orange. The overlap legends refer to the mean figures across all bundles.

### 4.3 HCP

Table 4 shows the bundle overlap of the latent-generated streamlines for both streamline plausibility criteria compared to the overlap of the seed streamlines on the HCP dataset. On average, GESTA increases the overlap significantly by proposing new, anatomically plausible streamlines using the available set of seed streamlines.

Table 4: HCP dataset reconstructed seed streamlines' and generative streamlines' overlap. $N = 15\,000$ streamlines are generated for each bundle with a bandwidth factor of value 1.0, and the plausibility is evaluated using the $ADG_R$ and $ADGC_R$ criteria. Mean and standard deviation values across bundles.

| | | $ADG_R$ | $ADGC_R$ |
|---|---|---|---|
| **Seed ratio** ($P\%$) | **OL** | **OL** (↑) | **OL** (↑) |
| 100 | 0.29 (0.17) | 0.74 (0.15) | 0.62 (0.15) |

Ten (10) bundles obtained using the proposed generative tractography method for the HCP dataset are shown in figure 7. As it is the case with the "Fiber Cup" and ISMRM 2015 Tractography Challenge datasets, GESTA populates each of the bundles with anatomically plausible streamlines using a limited set of seed streamlines: it increases by 2.6 times the average overlap across bundles for the $ADG_R$ criterion. Although the number of seeds varies notably across bundles (see the difference between the MCP and both POPT bundles), the rejection sampling procedure is still able to produce new samples using an extremely modest set of seeding streamline. See section 5.3 for a discussion about the seeding strategy.

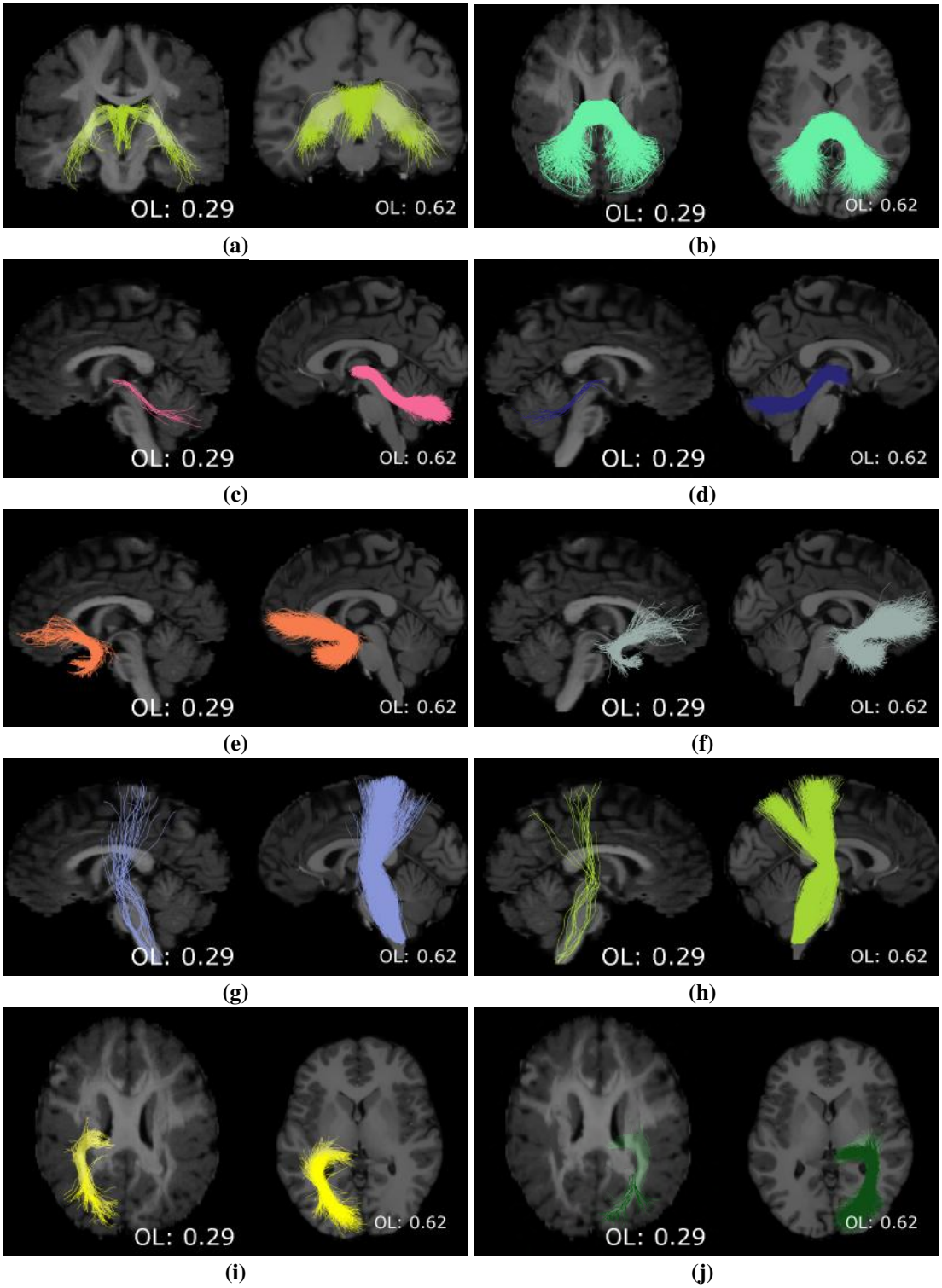



Figure 7: GESTA applied to the (a) fornix; (b) MCP; (c) left SCP; (d) right SCP; (e) left UF; (f) right UF; (g) left POPT; (h) right POPT; (i) left OR; and (j) right OR bundles of the HCP dataset. For each bundle: left: seed streamlines; right: GESTA-generated plausible streamlines. All available streamlines are used for seeding in the latent space; and streamlines are evaluated using the $ADGC_R$ criterion. Seed streamlines are in the ISMRM 2015 Tractography Challenge dataset space; latent-generated streamlines are in subject space. The overlap legends refer to the mean figures across all bundles.

### 4.4 BIL&GIN callosal homotopic

Table 5 shows the volume of the seed streamlines and the generative streamlines for the BIL&GIN callosal homotopic dataset. GESTA produces new streamlines that successfully increase the volume coverage using a limited set of streamlines (3.9 times on average across all segments for the $ADG_R$ criterion; 3.7 times for the $ADGC_R$ criterion).

Table 5: BIL&GIN callosal homotopic dataset seed streamlines' and generative streamlines' volume measures (mean and standard deviation). Measures are averaged over all available subjects and segments. All segments having at least one streamline are included in the computation of the seed streamline volume measure; however, as mentioned in section 5.4 segments that have only one streamline cannot be effectively sampled, and hence, are not considered in the generative framework. All measures are in $\mathrm{mm}^3$.

| | | $ADG_R$ | $ADGC_R$ |
|---|---|---|---|
| **Seed ratio** ($P\%$) | **Volume** | **Volume** | **Volume** |
| 100 | 9349 (1973) | 36736 (4143) | 34843 (3991) |

Figure 8 shows the result of the generative tractography performed on 6 of the gyral-based segments, namely the AG, Cu, IFG, LG, MOG, and the STG, of the BIL&GIN callosal homotopic dataset. The FINTA autoencoder-based tractography filtering method presented in Legarreta et al. (2021) showed a poor performance on these segments. The figure shows that the GESTA generative tractography method can be used to recover the lost white matter spatial coverage.

Figure 9 shows the seed streamlines and the latent-generated streamlines for a given BIL&GIN callosal homotopic dataset test subject. As it can be seen, GESTA improves the white matter spatial occupancy across all segments. Particularly, the segments where the available seed streamline count is extremely low, such as the LG, SMG, or the SPG, there is a notable increase in the number of streamlines that are incorporated by the generative framework.

The proposed generative tractography framework reliably reconstructs streamlines in regions where multiple fibers exist. The method does not involve propagating streamlines in a local orientation field, and thus, the reconstruction process is not influenced by potentially wrong local decisions. The close-up coronal view in figure 10 demonstrates that GESTA successfully creates multiple streamline populations in a sector of the corpus callosum where both the SFG and MFG homotopic segments contain fibers.

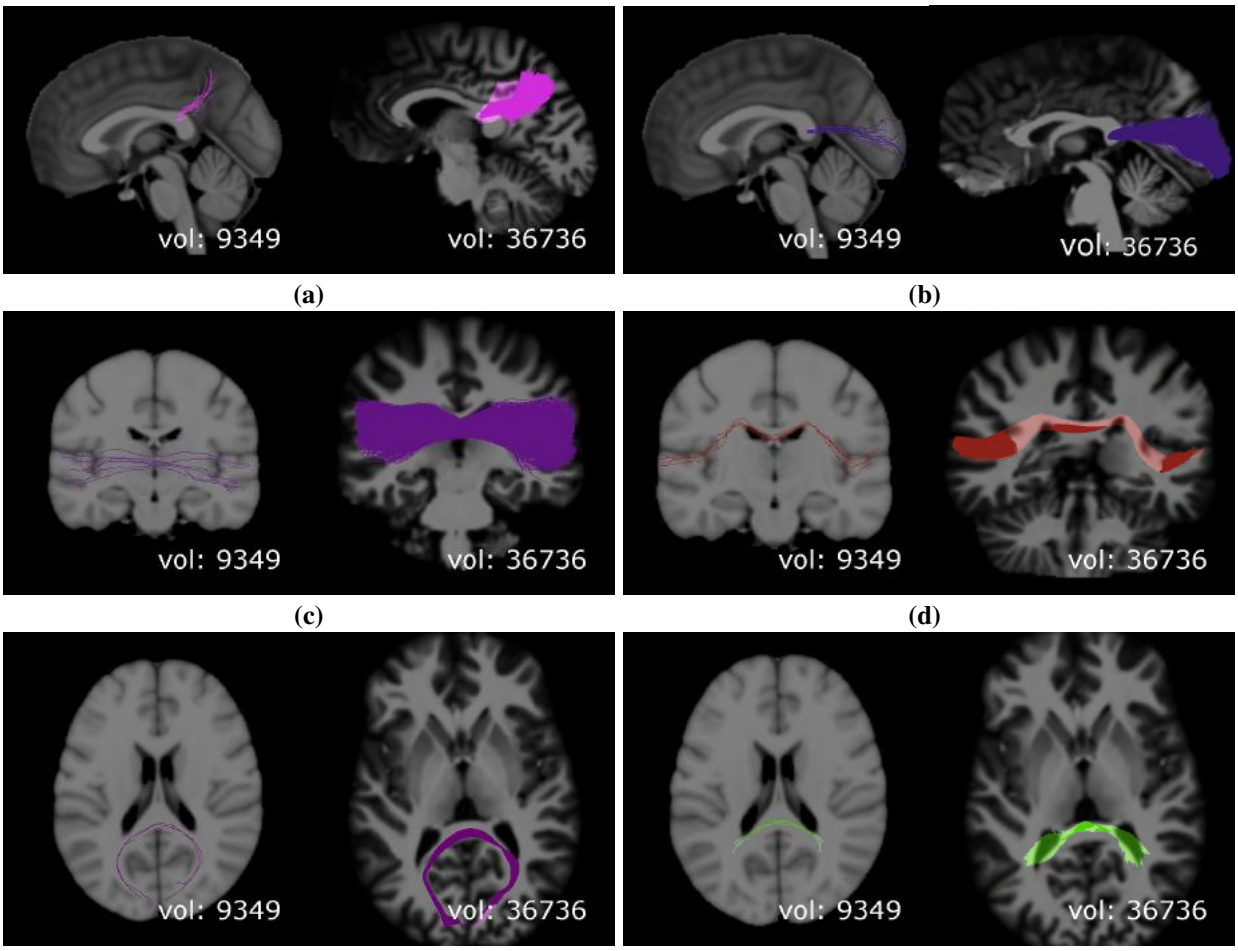
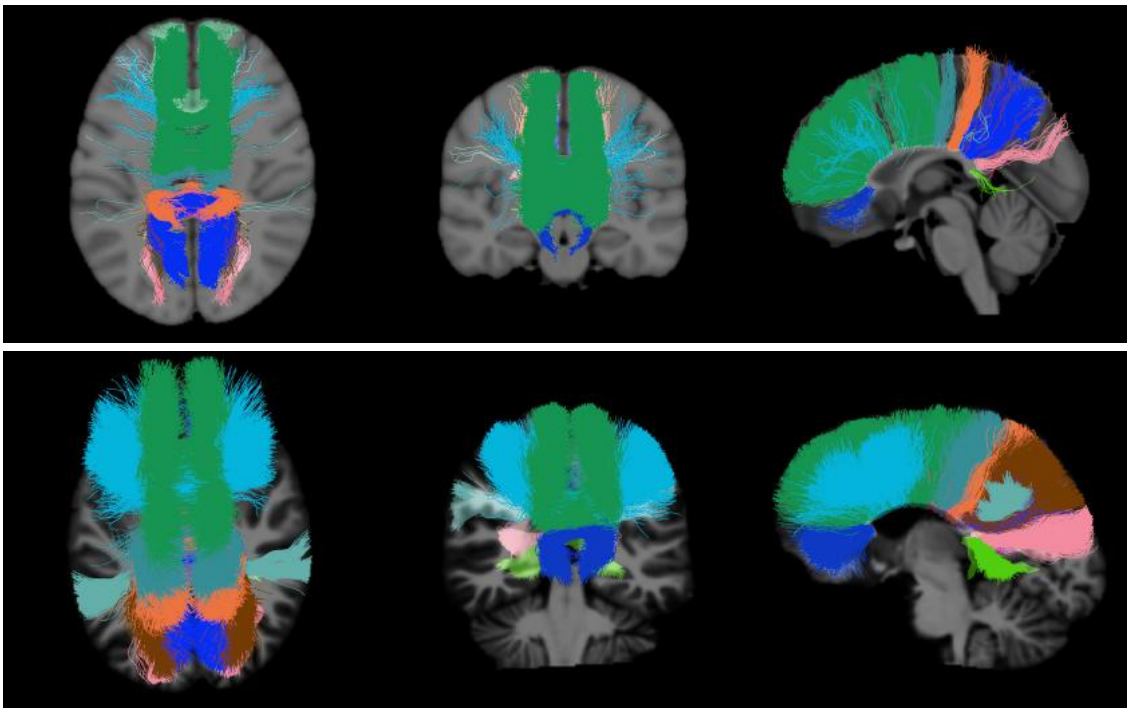



Figure 8: GESTA applied to the BIL&GIN callosal homotopic dataset. For each segment: left: seed streamlines; right: GESTA-generated plausible streamlines. Segments: (a) AG; (b) MOG; (c) IFG; (d) STG; (e) Cu; (f) LG. All available streamlines filtered with FINTA are used for seeding in the latent space; and streamlines are evaluated using the $ADG_R$ criterion. Data corresponds to different test subjects (note that not all subjects contain streamlines in all segments). Views have been chosen to best visualize the bundles. The legends refer to the mean bundle volume figures (in $\mathrm{mm}^3$) across all subjects and segments.



**(a)** **(b)** **(c)**

Figure 9: Seed streamlines and latent-generated plausible streamlines for a given BIL&GIN callosal homotopic dataset test subject. All available streamlines filtered with FINTA are used for seeding in the latent space; and streamlines are evaluated using the $ADG_R$ criterion. (a) Axial superior; (b) coronal anterior; (c) sagittal left views. Segment color code (approximate): Cing: pink (not visible); LG: lime; MFG: cyan; PoCG: orange; PrCG: turquoise; PrCu: blue; RG: cobalt blue; SFG: emerald; SMG: aero blue; SOG: cotton candy; SPG: brown.

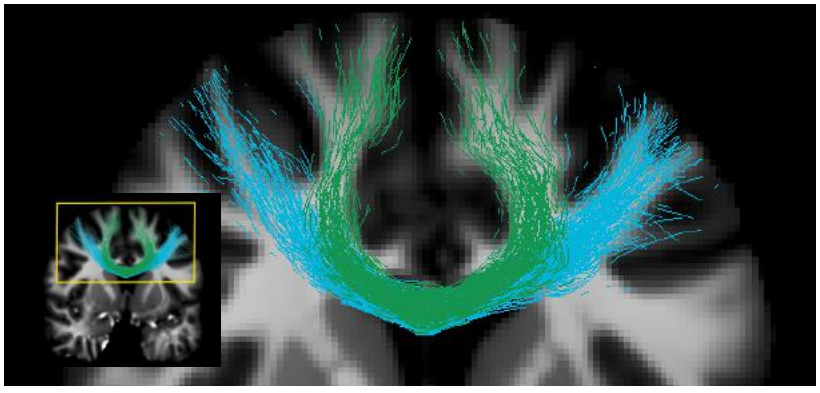


Figure 10: Generative streamlines for a given BIL&GIN callosal homotopic dataset subject. Streamlines shown correspond to SFG (emerald) and MFG (cyan) segments. Coronal anterior view.

# 5 Discussion

We presented GESTA (*Generative Sampling in Tractography using Autoencoders*), a generative method to provide new, anatomically plausible streamlines that enhances the spatial coverage of an existing tractogram. We demonstrate that the method can be specially useful to reliably fill bundles that are poorly populated in tractograms. The method allows to circumvent the difficulties of a given tracking method on regions where a low local orientation signal prevents a streamline to be propagated or privileges a given orientation. Conventional local orientation integration methods, global optimization methods, microstructure-informed methods, or yet deep learning-based tracking methods (where the training data might be implicitly impacted by the mentioned effect) might equally benefit from the proposed approach.

By its *generative* nature, our approach is fundamentally different from other retrospective tractogram improvement strategies, such as the SIFT (Smith et al., 2015) or COMMIT frameworks (Daducci et al., 2015). These approaches are *filtering* methods seeking to improve the fit of the streamlines present in a tractogram to the diffusion orientation signal or the tissue microstructure by typically removing streamlines with a poor fit.

Sampling from the latent space has the advantage of by-passing difficulties imposed by the (i) diffusion image quality (streamlines could be generated from very low resolution images); and (ii) hard-to-track regions. Streamlines would still be proposed for regions that are hard-to-track provided that reliable bundle models are available at learning and to sample from. Meanwhile, other methods such as Bundle-Specific Tractography (BST) (Rheault et al., 2019) or dynamically seeded Surface Enhanced Tractography (SET) (St-Onge et al., 2021), might find it difficult to traverse some hard-to-track regions given that they are still guided by some streamline-, diffusion- or anatomically-based signal.

Compared to conventional streamline propagation methods, the proposed generative streamline sampling method is independent of the diffusion acquisition sequence, or the choice of a particular local orientation reconstruction method. Similarly, our generative tractography process does not require computing a response function, any compartment volume fraction, or any other local construct.

## 5.1 Latent space sampling aspects

Our experiments show that some latent space regions benefit from using a more localized scope in order to be appropriately sampled. These regions likely correspond to bundles with a large spatial extent or having a diversity of streamline configurations, such as the corpus callosum. We hypothesize that such bundles are encoded by a more complex distribution in the learned latent space. An unconstrained autoencoder, unconditional on any prior information, such as the one used in this work, might have a limited ability to provide a latent space that offers a uniform behavior for sampling purposes. Thus, in order to obtain a presumably more tractable, spatially localized distribution that might be easier to sample from, some bundles can be further split into coherent streamline groups. In this work, the corpus callosum (initially available as a single bundle in the ISMRM 2015 Tractography Challenge and HCP datasets) is split into the 6 groups defined in Rheault (2021), and each group is sampled separately.

The results in figure 5 suggest that some other bundles might have also benefited from an additional split to be able to better sample some of the seeds (e.g. the most lateral CST streamlines). This would likely allow the rejection sampling method to obtain a better fit to some of the less densely packed streamline seeds, which should offer better results when sampling them.

We did not investigate whether using another set of seed streamlines, chosen according to some non-random policy, would have allowed to sample the entire corpus callosum without the need to split it. Similarly, we did not investigate whether other sampling methods are able to more easily fit and sample from such complex distributions.

In our experiments, we consider the fornix a commissural bundle (and thus composed of a single component), following the bundles used in Maier-Hein et al. (2017). However, other works (both describing the human white matter anatomy, such as Catani et al. (2002); Catani and Thiebaut de Schotten (2008), and presenting white matter dissection methods, such as Schilling et al. (2021a)) model the fornix as a pair of left and right hemisphere component projection bundles. In our experiments, we did not take steps to ensure that the seed streamlines were equally distributed across hemispheres, and thus, when the number of seeds is low and due to the random sampling effect, components of a hemisphere might get under-represented. Such an imbalance impacts the generative sampling process, and ultimately the improvement in the white matter occupancy.

Although the bandwidth used to fit the proposed distribution influences the number of accepted samples by the rejection sampling method, this generally comes at the cost of an increased likelihood of streamlines not complying with

the anatomical plausibility criteria downstream. Given enough computational resources and time, eventually a larger compliance ratio might be obtained by using different sampling parameters and by aggregating results.

An additional advantage of our generative framework is that virtually no "prematurely terminated" streamline is generated given a set of *well-behaved* streamline seeds. In this context,"prematurely terminated" designates those streamlines that notably deviate from the length distribution expected from the bundle they belong to. In conventional streamline propagation methods, a non-negligible amount of streamlines exit the white matter mask without reaching their target endpoints in the cortex due to unfavorable local orientation decisions. The extremely low rate of length-wise discarded streamlines observed experimentally, and the overlap being conserved when applying the *ADGC* criterion on the "Fiber Cup" (see table 2), shows that GESTA is able to yield better behaved streamlines length-wise. GESTA may still provide streamlines that do not reach the cortex, as shown by the decrease in the *ADGC* criterion overlap figures on the ISMRM 2015 Tractography Challenge and HCP datasets (tables 3 and 4); however, both the gray matter mask and the seeds used might have affected such results. As required by the application, the streamline plausibility evaluation framework admits adjustments on its criteria (e.g. including other tissues and structures to the masks, etc.), parameters and thresholds.

### 5.2 Generative tractography evaluation framework

The proposed streamline compliance criteria are designed such that constraints are increasingly restrictive, *ADG* being the less demanding criterion, and *ADGC* being the most demanding. Streamlines reaching the GM largely depend on the seed streamlines given the lack of explicit modelling in our framework. Some randomly selected seeds, which have been considered as being plausibles by the underlying scoring methods, may actually not reach the cortex, and hence the sampled streamlines might show the same behavior, ultimately being discarded by the $ADGC_B$ (or $ADGC_R$) criterion. At the same time, in the proposed setting, the GM occupancy constraint is considered as a *hard* constraint, as it is a binary criterion. The WM occupancy ratio constraint, where the requirement is relaxed by introducing a volume-based tissue occupancy threshold, is considered a soft criterion. Investigating whether using a spatially-localized, variable ratio or window (e.g. softening the constraint only at the tissue interface or using partial volume effect (PVE) maps obtained from T1-weighted data) would offer a more suitable plausibility evaluation framework is considered to be beyond the aims of this work.

We decided to explicitly avoid generating streamline evaluation criteria that include partially fulfilling constraints of different nature. Specifically, we do not consider the case where streamlines could reach the GM at only one endpoint, but be within a given distance from the cortex at the other endpoint. Although this could be a valid criterion to assess the plausibility of streamlines, we chose to model separately the influence of criteria that differ in their nature. Also, this allows to keep the number and complexity of the plausibility evaluation criteria to a set of limited, well-established measures.

The allowed streamline to local orientation angle gap is relaxed for the HCP and BIL&GIN callosal homotopic datasets. Allowing a $30°$ cone when propagating local orientations into streamlines is a common choice. Given that our autoencoder is trained solely using streamline data, no local information is encoded in our latent space, and hence, the constraint is relaxed in order to overcome this limitation.

Experimental evidence (Maier-Hein et al., 2017) has shown that some of these requirements might need to be relaxed to avoid being excessively restrictive when evaluating the results of tracking methods. In the context of this work, WM and GM masks are slightly dilated, and the streamline local orientation to peak alignment criterion uses a soft condition (rather than strictly enforcing to comply with an alignment cone at each streamline segment) (see section A.1). The softened criterion allows the generated streamlines to overcome the bottleneck effect at locations along their trajectory, where conventional tracking methods have more difficulties in passing through. As shown in Canales-Rodríguez et al. (2019), current spherical deconvolution methods' ability to distinguish crossing fiber populations declines considerably below the $45$-$35°$ inter-fiber angle range. Thus, provided that appropriate seed streamlines exist, the generative framework presented in this work could be used in regions that are hard to traverse for conventional local reconstruction and streamline propagation methods.

### 5.3 Experimental choices

In our experiments, the seed streamlines are chosen randomly. We did not investigate strategies to increase the diversity of the selected reference streamlines. The proposed procedure allows to test our method without providing it any particular advantageous context. However, as revealed in figures 4 and 5, as seeds are not distributed uniformly along all possible pathways, the sampling method can draw samples preferentially from a given region of the latent space with a larger density. This is a consequence of the underlying (seed) data distribution being biased over some particular area or configuration, which results in an estimate of the distribution reflecting such bias.

When using a low percentage of the available set of the test streamlines as seeds, a given bundle might be left with an insufficient number of streamlines to sample from (see section 5.4). Note that the number of available seeds (i.e. prior to any subsampling) varies largely across bundles, and the choice is further restricted by the use of the test set. In this work, we do not aim to produce an atlas of latent-generated streamlines so as to measure the improvement in downstream tasks (such as tractometry or connectomics), but to provide evidence that building such an atlas is feasible using the presented method. Thus, we did not seek to find the optimal set of seed of streamlines that would allow for a more favorable sampling and which would maximize the generative performance.

### 5.4 Limitations

The quality of the generated streamlines is fundamentally tied to the richness of the seed streamlines. The tractography data used in this work, and the training and testing splits, had not been generated so as to make all possible (or known) intra-bundle configurations sufficiently available at generation time. As a consequence, the anatomical plausibility of the generative streamlines covering such particular structures gets affected.

Due to the lossy nature of the encoder and the decoder, the latent-generated streamlines describe smoother trajectories than those of the seed streamlines. This might impact the plausibility of the generative streamlines at the boundaries of the bundles.

The autoencoder requires that all tractograms used at train and test time be in the same space (for example, the standard-MNI space). Thus, our framework is limited by any potential accuracy loss derived from the spatial transformations involved in the registration process.

The rejection sampling procedure involves estimating the distribution of the seed streamlines in the latent space. For a bundle or segment having a single streamline, such distribution cannot be effectively determined by the method, and hence, such bundles are discarded in the generative framework. A framework reliably incorporating various fiber pathways to such bundles in order to enable the sampling process is yet to be studied.

## 6 Future work

In this work, the same generative sampling parameters (namely, the Parzen estimator bandwidth factor) are used to generate streamlines across different bundles. This allows to consistently evaluate the effect of a given set of parameters on the used success measures. However, the framework allows to provide different parameters to different bundles, which might improve the success of the method at generating streamlines for a given bundle.

Similarly, we use the same feature thresholds for all bundles to determine the plausibility of streamlines sampled from the latent space. The framework could be adapted to use bundle-specific thresholds. Using a tighter match around bundle-specific values requires an agreement on such values. Studying the restrictiveness and effects of such values, if any, is out of the scope of this work, and might be addressed in the future.

An enhanced latent space seeding method might favorably shape the underlying streamline distribution for the presented sampling method. For example, using a strategy to favor dissimilar reference picks could possibly allow the sampling method to generate candidates that would span a richer set of intra-bundle configurations. This would eventually lead to larger improvements in the white matter spatial coverage. The impact of other seeding methods (e.g. adaptive, bundle-aware, etc., or based on bundle summarizing methods) in order to draw streamline samples more uniformly across all latent space regions remains to be studied. Whether other sampling methods are more robust to the data distribution is yet to be investigated.

We chose not to use existing streamline propagation methods for streamlines that did not comply with some of the criteria. Streamlines that fail to reach the GM tissue at one end and which are currently discarded by the $ADGC_B$ and $ADGC_R$ criteria could be recovered for the purpose by incorporating a surface flow method (St-Onge et al., 2018).

The proposed generative framework being unconditional, it lacks explicit spatial constraints to harness the streamline candidate generation. Self-supervised approaches for autoencoders, as the one proposed by Chen et al. (2021), could be investigated as a means to improve the learned latent space. Similarly, introducing constraints in the latent space might be useful to provide an autoencoder capable of generating outputs with some desired attributes. Providing some further level of guarantee on the generative streamlines' behavior, such as learning appropriate termination patterns, would require introducing some additional parameterization to train the autoencoder. Likewise, a closer streamline feature-to-latent dimension space correlation might improve the latent space-based generative capabilities offered by such network. Leveraging the autoencoder to improve the learned latent representations, and thus increasing the sampled streamlines' success rate in terms of their plausibility according to some criterion by introducing constraints into the training process is left for a separate piece of work.

Provided that a large and complete set of generative streamlines is built, our approach could be used to substitute streamlines that do not comply with a given plausibility criterion (e.g. reaching the GM) for a (latent space) nearest neighbor in the set that is guaranteed to comply with the plausibility criteria. Additional experiments would be required to investigate the benefits of this procedure in downstream tasks, such as connectomics.

The presented streamline evaluation framework can be used in the context of any deep learning-powered generative approach. Extending its capabilities to include other plausibility criteria (e.g. including tissue maps, or bundle segmentation masks, or microstructure-informed features, etc.) is left as a future work.

## 7 Conclusions

In this work, we introduced the first deep generative tractography framework. We demonstrate that an autoencoder can be used for a variety of tasks in tractography (i) without any changes to the architecture, and (ii) being trained only once. The same autoencoder can be used to (i) filter a tractogram, removing implausible streamlines, and (ii) to generate new streamlines to enhance the spatial coverage of an existing tractogram and its white matter bundles. The GESTA (*Generative Sampling in Tractography using Autoencoders*) framework is shown to successfully generate anatomically plausible streamlines across a variety of both synthetic and *in-vivo* human brain datasets by exploiting the latent space of an already trained autoencoder. The proposed generative tractography method may help tractometry and connectomics derivatives by providing a more anatomically reliable white matter mapping.

## Acknowledgments

This work has been partially supported by the Centre d'Imagerie Médicale de l'Université de Sherbrooke (CIMUS); the Axe d'Imagerie Médicale (AIM) of the Centre de Recherche du CHUS (CRCHUS); and the Réseau de Bio-Imagerie du Québec (RBIQ)/Quebec Bio-imaging Network (QBIN) (FRSQ - Réseaux de recherche thématiques File: 35450). This research was enabled in part by support provided by Calcul Québec (www.calculquebec.ca/en/) and Compute Canada (www.computecanada.ca). We also thank the research chair in Neuroinformatics of the Université de Sherbrooke. Finally, thanks to Carl Lemaire, Étienne St-Onge, and Antoine Théberge for their insightful comments and discussions. Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

## References

B. Anctil-Robitaille, C. Desrosiers, and H. Lombaert. Manifold-aware cycleGAN for high-resolution structural-to-DTI synthesis. In N. Gyori, J. Hutter, V. Nath, M. Palombo, M. Pizzolato, and F. Zhang, editors, *Computational Diffusion MRI (CDMRI) Medical Image Computing and Computer Assisted Intervention (MICCAI) Workshop*, Mathematics and Visualization, pages 213–224, Lima, Peru, 10 2020. Medical Image Computing and Computer Assisted Intervention (MICCAI) Society, Springer.

B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee. A reproducible evaluation of ants similarity metric performance in brain image registration. *NeuroImage*, 54(3):2033–2044, 02 2011.

M. Battocchio, S. Schiavi, M. Descoteaux, and A. Daducci. Improving tractography accuracy using dynamic filtering. In N. Gyori, J. Hutter, V. Nath, M. Palombo, M. Pizzolato, and F. Zhang, editors, *Computational Diffusion MRI (CDMRI) Medical Image Computing and Computer Assisted Intervention (MICCAI) Workshop*, Mathematics and Visualization, pages 45–54, Lima, Peru, 10 2020. Medical Image Computing and Computer Assisted Intervention (MICCAI) Society, Springer.

M. Battocchio, S. Schiavi, M. Descoteaux, and A. Daducci. Bundle-o-graphy. In *29th Annual Meeting & Exhibition of the International Society for Magnetic Resonance in Medicine (ISMRM)*. International Society for Magnetic Resonance in Medicine, 05 2021. Abstract #0867.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transasctions on Pattern Analysis Machine Intelligence*, 35(8):1798–1828, 08 2013.

E. J. Canales-Rodríguez, J. H. Legarreta, M. Pizzolato, G. Rensonnet, G. Girard, J. R.-P. no, M. Baraković, D. Romascano, Y. Alemán-Gómez, J. Radua, E. Pomarol-Clotet, R. Salvador, J.-P. Thiran, and A. Daducci. Sparse wars: A survey and comparative study of spherical deconvolution algorithms for diffusion mri. *NeuroImage*, 184:140–160, 01 2019.

M. Catani and M. Thiebaut de Schotten. A diffusion tensor imaging tractography atlas for virtual in vivo dissections. *Cortex*, 44(8):1105–1132, 09 2008. Special Issue on "Brain Hodology - Revisiting disconnection approaches to disorders of cognitive function".

M. Catani, R. J. Howard, S. Pajevic, and D. K. Jones. Virtual in vivo interactive dissection of white matter fasciculi in the human brain. *NeuroImage*, 17(1):77–94, 09 2002.

M. Chamberland, K. Whittingstall, D. Fortin, D. Mathieu, and M. Descoteaux. Real-time multi-peak tractography for instantaneous connectivity display. *Frontiers in Neuroinformatics*, 8:59, 05 2014.

M. Chamberland, B. Scherrer, S. P. Prabhu, J. Madsen, D. Fortin, K. Whittingstall, M. Descoteaux, and S. K. Warfield. Active delineation of Meyer's loop using oriented priors through MAGNEtic tractography (MAGNET). *Human Brain Mapping*, 38(1):509–527, 01 2017.

Y. Chen, C. Zhang, Y. Song, N. Makris, Y. Rathi, W. Cai, F. Zhang, and L. J. O'Donnell. Deep fiber clustering: Anatomically informed unsupervised deep learning for fast and effective white matter parcellation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Strasbourg, France, 09-10 2021. Springer International Publishing.

Q. Chenot, N. Tzourio-Mazoyer, F. Rheault, M. Descoteaux, F. Crivello, L. Zago, E. Mellet, G. Jobard, M. Joliot, B. Mazoyer, and L. Petit. A population-based atlas of the human pyramidal tract in 410 healthy participants. *Brain Structure and Function*, 224(2):599–612, 03 2019.

M.-A. Côté, G. Girard, A. Boré, E. Garyfallidis, J.-C. Houde, and M. Descoteaux. Tractometer: Towards validation of tractography pipelines. *Medical Image Analysis*, 17(7):844–857, 10 2013. Special Issue on the 2012 Conference on Medical Image Computing and Computer Assisted Intervention.

A. Daducci, A. D. Palù, A. Lemkaddem, and J.-P. Thiran. COMMIT: Convex optimization modeling for microstructure informed tractography. *IEEE Transactions on Medical Imaging*, 34(1):246–257, 01 2015.

M. Descoteaux, R. Deriche, T. R. Knösche, and A. Anwander. Deterministic and probabilistic tractography based on complex fibre orientation distributions. *IEEE Transactions on Medical Imaging*, 28(2):269–286, 02 2009.

P. Fillard, M. Descoteaux, A. Goh, S. Gouttard, B. Jeurissen, J. Malcolm, A. Ramirez-Manzanares, M. Reisert, K. Sakaie, F. Tensaouti, T. Yo, J.-F. Mangin, and C. Poupon. Quantitative evaluation of 10 tractography algorithms on a realistic diffusion MR phantom. *NeuroImage*, 56(1):220–234, 05 2011.

V. S. Fonov, A. C. Evans, K. Botteron, C. R. Almli, R. C. McKinstry, and D. L. Collins. Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, 54(1):313–327, 2011.

G. Girard, K. Whittingstall, R. Deriche, and M. Descoteaux. Towards quantitative connectivity analysis: reducing tractography biases. *NeuroImage*, 98:266–278, 09 2014.

M. F. Glasser, S. M. Smith, D. S. Marcus, J. L. R. Andersson, E. J. Auerbach, T. E. J. Behrens, T. S. Coalson, M. P. Harms, M. Jenkinson, S. Moeller, E. C. Robinson, S. N. Sotiropoulos, J. Xu, E. Yacoub, K. Ugurbil, and D. C. V. Essen. The human connectome project's neuroimaging approach. *Nature Neuroscience*, 19(9):1175–1187, 09 2016.

B. Jeurissen, M. Descoteaux, S. Mori, and A. Leemans. Diffusion MRI fiber tractography of the brain. *NMR in Biomedicine*, 32(4):e3785, 04 2019. e3785 NBM-17-0045.R2.

J. H. Legarreta, L. Petit, F. Rheault, G. Theaud, C. Lemaire, M. Descoteaux, and P.-M. Jodoin. Filtering in tractography using autoencoders (FINTA). *Medical Image Analysis*, 72:102126, 08 2021.

S. Luo, J. Zhou, Z. Yang, H. Wei, and Y. Fu. Diffusion MRI super-resolution reconstruction via sub-pixel convolution generative adversarial network. *Magnetic Resonance Imaging*, 88:101–107, 05 2022.

K. H. Maier-Hein, P. Neher, J.-C. Houde, E. Caruyer, A. Daducci, T. Dyrby, B. Stieltjes, and M. Descoteaux. Tractography challenge ISMRM 2015 data, 05 2015. URL `https://doi.org/10.5281/zenodo.572345`.

K. H. Maier-Hein, P. F. Neher, J.-C. Houde, M.-A. Côté, E. Garyfallidis, J. Zhong, M. Chamberland, F.-C. Yeh, Y. C. Lin, Q. Ji, W. E. Reddick, J. O. Glass, D. Q. Chen, Y. Feng, C. Gao, Y. Wu, J. Ma, H. Renjie, Q. Li, C.-F. Westin, S. Deslauriers-Gauthier, J. O. O. González, M. Paquette, S. St-Jean, G. Girard, F. M. Rheault, J. Sidhu, C. M. W. Tax, F. Guo, H. Y. Mesri, S. Dávid, M. Froeling, A. M. Heemskerk, A. Leemans, A. Boré, B. Pinsard, C. Bedetti, M. Desrosiers, S. Brambati, J. Doyon, A. Sarica, R. Vasta, A. Cerasa, A. Quattrone, J. Yeatman, A. R. Khan, W. Hodges, S. Alexander, D. Romascano, M. Barakovic, A. Auría, Ó. Esteban, A. Lemkaddem, J.-P. Thiran, H. E. Cetingul, B. L. Odry, B. Mailhe, M. Nadar, F. Pizzagalli, G. Prasad, J. E. Villalón-Reina, J. Galvis, P. M. Thompson, F. D. S. Requejo, P. L. Laguna, L. M. Lacerda, R. Barrett, F. Dell'Acqua, M. Catani, L. Petit, E. Caruyer, A. Daducci, T. B. Dyrby, T. Holland-Letz, C. C. Hilgetag, B. Stieltjes, and M. Descoteaux. The challenge of mapping the human connectome based on diffusion tractography. *Nature Communications*, 8(1349):1–13, 11 2017.

B. Mazoyer, E. Mellet, G. Perchey, L. Zago, F. Crivello, G. Jobard, N. Delcroix, M. Vigneau, G. Leroux, L. Petit, M. Joliot, and N. Tzourio-Mazoyer. BIL&GIN: A neuroimaging, cognitive, behavioral, and genetic database for the study of human brain lateralization. *NeuroImage*, 124:1225–1231, 01 2016. Sharing the wealth: Brain Imaging Repositories in 2015.

P. F. Neher, F. B. Laun, B. Stieltjes, and K. H. Maier-Hein. Fiberfox: Facilitating the creation of realistic white matter software phantoms. *Magnetic Resonance in Medicine*, 72(5):1460–1470, 2014.

L. J. O'Donnell, A. Daducci, D. Wassermann, and C. Lenglet. Advances in computational and statistical diffusion MRI. *NMR in Biomedicine*, 32(4):e3805, 04 2019. e3805 nbm.3805.

N. Painchaud, Y. Skandarani, T. Judge, O. Bernard, A. Lalande, and P.-M. Jodoin. Cardiac segmentation with strong anatomical guarantees. *IEEE Transactions on Medical Imaging*, 39(11):3703–3713, 11 2020.

P. Poulin, F. Rheault, E. St-Onge, P.-M. Jodoin, and M. Descoteaux. Bundle-wise deep tracker: Learning to track bundle-specific streamline paths. In *27th Joint Annual Meeting of the International Society for Magnetic Resonance in Medicine (ISMRM) and the European Society for Magnetic Resonance in Medicine and Biology (ESMRMB)*, Paris, France, 06 2018. International Society for Magnetic Resonance in Medicine. Abstract #0041.

P. Poulin, D. Jörgens, P.-M. Jodoin, and M. Descoteaux. Tractography and machine learning: Current state and open challenges. *Magnetic Resonance Imaging*, 64:37–48, 12 2019. Artificial Intelligence in MRI.

D. A. Raffelt, R. E. Smith, G. R. Ridgway, J.-D. Tournier, D. N. Vaughan, S. Rose, R. Henderson, and A. Connelly. Connectivity-based fixel enhancement: Whole-brain statistical analysis of diffusion mri measures in the presence of crossing fibres. *NeuroImage*, 117:40–55, 08 2015.

M. Reisert, I. Mader, C. Anastasopoulos, M. Weigel, S. Schnell, and V. Kiselev. Global fiber reconstruction becomes practical. *NeuroImage*, 54(2):955–962, 01 2011.

F. Rheault. Population average atlas for RecobundlesX, 03 2021. URL `https://doi.org/10.5281/zenodo.4630660`.

F. Rheault, M. Roy, S. Cunnane, and M. Descoteaux. Bundle-specific fornix reconstruction for dual-tracer PET-tractometry. bioRxiv, 09 2018. URL `https://www.biorxiv.org/content/10.1101/423459v1`.

F. Rheault, E. St-Onge, J. Sidhu, K. H. Maier-Hein, N. Tzourio-Mazoyer, L. Petit, and M. Descoteaux. Bundle-specific tractography with incorporated anatomical and orientational priors. *NeuroImage*, 186:382–398, 02 2019.

F. Rheault, P. Poulin, A. V. Caron, E. St-Onge, and M. Descoteaux. Common misconceptions, hidden biases and modern challenges of dMRI tractography. *Journal of Neural Engineering*, 17(1):011001, 02 2020.

K. G. Schilling, V. Nath, C. Hansen, P. Parvathaneni, J. Blaber, Y. Gao, P. Neher, D. B. Aydogan, Y. Shi, M. Ocampo-Pineda, S. Schiavi, A. Daducci, G. Girard, M. Barakovic, J. R.-P. no, D. Romascano, G. Rensonnet, M. Pizzolato, A. Bates, E. Fischi, J.-P. Thiran, E. J. Canales-Rodríguez, C. Huang, H. Zhu, L. Zhong, R. Cabeen, A. W. Toga, F. Rheault, G. Theaud, J.-C. Houde, J. Sidhu, M. Chamberland, C.-F. Westin, T. B. Dyrby, R. Verma, Y. Rathi, M. O. Irfanoglu, C. Thomas, C. Pierpaoli, M. Descoteaux, A. W. Anderson, and B. A. Landman. Limits to anatomical accuracy of diffusion tractography using modern approaches. *NeuroImage*, 185:1–11, 01 2019.

K. G. Schilling, C. M. Tax, F. Rheault, C. Hansen, Q. Yang, F.-C. Yeh, L. Cai, A. W. Anderson, and B. A. Landman. Fiber tractography bundle segmentation depends on scanner effects, vendor effects, acquisition resolution, diffusion sampling scheme, diffusion sensitization, and bundle segmentation workflow. *NeuroImage*, 242:118451, 11 2021a.

K. G. Schilling, C. M. W. Tax, F. Rheault, B. A. Landman, A. W. Anderson, M. Descoteaux, and L. Petit. Prevalence of white matter pathways coming into a single white matter voxel orientation: The bottleneck issue in tractography. *Human Brain Mapping*, pages 1–18, 12 2021b.

R. E. Smith, J.-D. Tournier, F. Calamante, and A. Connelly. SIFT2: Enabling dense quantitative assessment of brain white matter connectivity using streamlines tractography. *NeuroImage*, 119:338–351, 10 2015.

E. St-Onge, A. Daducci, G. Girard, and M. Descoteaux. Surface-enhanced tractography (set). *NeuroImage*, 169: 524–539, 04 2018.

E. St-Onge, N. Al-Sharif, G. Girard, G. Theaud, and M. Descoteaux. Cortical surfaces integration with tractography for structural connectivity analysis. *Brain Connectivity*, 11(7):505–517, 09 2021.

A. Théberge, C. Desrosiers, M. Descoteaux, and P.-M. Jodoin. Track-to-learn: A general framework for tractography with deep reinforcement learning. *Medical Image Analysis*, 72:102093, 08 2021.

L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.

F. Zhang, A. Daducci, Y. He, S. Schiavi, C. Seguin, R. E. Smith, C.-H. Yeh, T. Zhao, and L. J. O'Donnell. Quantitative mapping of the brain's structural connectivity using diffusion MRI tractography: A review. *NeuroImage*, 249: 118870, 04 2022.

Y. Zhang, J. Zhang, K. Oishi, A. V. Faria, H. Jiang, X. Li, K. Akhter, P. Rosa-Neto, G. B. Pike, A. C. Evans, A. W. Toga, R. Woods, J. C. Mazziotta, M. I. Miller, P. C. M. van Zijl, and S. Mori. Atlas-guided tract reconstruction for automated and comprehensive examination of the white matter anatomy. *NeuroImage*, 52(4):1289–1301, 10 2010.

# A Appendix

## A.1 Streamline evaluation framework parameterization

In order to determine the anatomical plausibility of the streamlines generated from the latent space, a number of choices are made concerning each of the evaluated aspects. An analysis on the thresholding values presented in table 1 is provided in the following:

- Streamline geometry: the streamline length is set to be within the $20\,\mathrm{mm}$ to $220\,\mathrm{mm}$ range. The streamline winding is set to be below $330°$ for the "Fibercup" and ISMRM 2015 Tractography Challenge datasets; $340°$ for the HCP data; and $360°$ for the BIL&GIN callosal homotopic data. The BIL&GIN callosal homotopic data threshold is set to $360°$ due to its particular nature: following the anatomical filtering method in Legarreta et al. ((2021)) and the homotopic segment definition, it had been observed that a non-negligible proportion of streamlines extracted by the underlying method described closed loops.
- Local orientation alignment geometry: $75\%$ of the streamline local orientation values are required to be within a $30°$ cone with respect to the closest local fODF peak for the "Fibercup" and ISMRM 2015 Tractography Challenge datasets, which had been tracked using the same local tractography method, and $40°$ for the HCP and BIL&GIN callosal homotopic data. Such a degree of tolerance allows to account for differences in the tracking methods and parameters, which possibly result in tractograms of varying intrinsic attributes. At the same time, this might allow to traverse regions that are hard-to-track for conventional streamline propagation methods. $90°$ values are masked in the check to accommodate for the peak support absent segments (such as those overreaching the WM mask at the streamline endpoints). The fODF peak maps used in this work contain the $5$ largest peaks at each voxel.
- White matter tissue occupancy: streamlines are required to lie within the WM mask boundaries. We chose to skip the last $N = 10$ consecutive vertices at each streamline point when checking the WM occupancy. The generative framework provides streamline samples whose endpoints may exceed the boundaries of the WM. If the occupancy requirement is not relaxed, such streamlines would be discarded. Our streamlines have a fixed number of vertices, and thus the step size is variable across streamlines: shortest streamlines (in physical units) have a smaller step size when compared to longest streamlines. After investigation, skipping $N = 10$ consecutive vertices at each endpoint was chosen as a conservative upper bound to guarantee that the WM occupancy criterion would not penalize excessively the shortest streamlines. The WM tissue occupation is set as a binary criterion for the "Fiber Cup" dataset, but is softened for the rest of the datasets to require at least $95\%$ of the streamline vertices to be contained in the WM tissue. Such a relaxation comes as a necessary compromise for the partial volume effects and streamline trajectories slightly veering off locally.
- Gray matter tissue connectivity: streamline endpoints are required to be located in the GM mask.

The GESTA framework does not have an accurate control over the streamlines' trajectory, or their starting and termination vertices. Thus, the WM and GM tissue masks are dilated using $2$ iterations and a structuring element with a connectivity equal to $1$. The generative streamline candidates are trimmed to the corresponding brain mask. The brain masks are also eroded using $2$ iterations and a structuring element with a connectivity equal to $1$, except for the "Fiber Cup" dataset mask, which is used "as is" given the particular binary nature of the available structural data.

Due to the aforementioned lack of a precise control over the streamline termination, latent-generated raw streamlines are found to overreach the GM mask frequently. The generated raw streamlines are thus trimmed to lie within the computed brain masks for each dataset or subject. Using an eroded version of the brain masks allows to ensure that streamlines would not enter the CSF. Thus, we favor streamlines whose endpoints lay in the dilated GM mask for the $ADGC_B$ (or $ADGC_R$) criterion. In this work, we use binary GM tissue maps. We do not apply additional post-processing to the generative streamlines to only report on the generative ability of the autoencoder-based framework.

### A.2 Acronyms

The bundles considered for the ISMRM 2015 Tractography Challenge and the HCP datasets include: corpus callosum (CC); left/right cingulum; left/right cortico-spinal tract (CST); fornix; left/right fronto-pontine tract (FPT); left/right inferior cerebellar peduncle (ICP); left/right inferior longitudinal fasciculus (ILF); middle cerebellar peduncle (MCP); left/right optic radiation (OR); left/right parieto-occipital pontine tract (POPT); left/right superior cerebellar peduncle (SCP); left/right superior longitudinal fasciculus (SLF); and left/right uncinate fasciculus (UF).

The corpus callosum splits used for the ISMRM 2015 Tractography Challenge dataset and the HCP subject are: frontal lobe (most anterior part) (CC_Fr_1); frontal lobe (most posterior part) (CC_Fr_2); occipital lobe (CC_Oc); parietal lobe (CC_Pa); pre-/post-central gyri (CC_Pr_Po); and temporal lobe (CC_Te).

The BIL&GIN callosal homotopic data had been defined using the following 26 gyral-based regions of interest: angular gyrus (AG); cingulum (Cing); cuneus (Cu); fusiform gyrus (FuG); hippocampus (Hippo); inferior frontal gyrus (IFG); inferior occipital gyrus (IOG); inferior temporal gyrus (ITG); insula (Ins); lateral fronto-orbital gyrus (LFOG); lingual gyrus (LG); middle frontal gyrus (MFG); middle fronto-orbital gyrus (MFOG); middle occipital gyrus (MOG); middle temporal gyrus (MTG); parahippocampal gyrus (PHG); post-central gyrus (PoCG); pre-central gyrus (PrCG); pre-cuneus (PrCu); rectus gyrus (RG); superior frontal gyrus (SFG); supramarginal gyrus (SMG); superior occipital gyrus (SOG); superior parietal gyrus (SPG); superior temporal gyrus (STG); and temporal pole (TPole).