

**TRACTOGRAPHIE DE LA MATIÈRE BLANCHE PAR
RÉSEAUX DE NEURONES RÉCURRENTS**

par

Philippe Poulin

Thèse présentée au Département d'informatique
en vue de l'obtention du grade de philosophiæ doctor (Ph.D.)

FACULTÉ DES SCIENCES

UNIVERSITÉ DE SHERBROOKE

Sherbrooke, Québec, Canada, 13 janvier 2022

Le 13 janvier 2022

Le jury a accepté la thèse de Philippe Poulin dans sa version finale.

Membres du jury

Professeur Maxime Descoteaux
Directeur
Département d'informatique

Professeur Pierre-marc Jodoin
Codirecteur
Département d'informatique

Professeur Kevin Whittingstall
Membre interne
Département de radiologie diagnostique
Faculté de médecine et des sciences de la santé

Professeur Christian Desrosiers
Membre externe
Département de génie logiciel et des TI
École de technologie supérieure, Montréal, Québec

Professeur Guillaume Gilet
Président-rapporteur
Département d'informatique

Sommaire

La matière blanche du cerveau fait encore l'objet de nombreuses études. Grâce à l'IRM de diffusion, on peut étudier de façon non invasive la connectivité du cerveau avec une précision sans précédent. La reconstruction de la matière blanche — la tractographie — n'est pas parfaite cependant. En effet, la tractographie tend à reconstruire tous les chemins possibles au sein de la matière blanche ; l'expertise des neuroanatomistes est donc requise pour distinguer les chemins qui sont possibles anatomiquement de ceux qui résultent d'une mauvaise reconstruction. Cette connaissance est difficile à exprimer et à codifier sous forme de règles logiques.

L'intelligence artificielle a refait surface dans les années 1990 — suite à une amélioration remarquable de la vitesse des processeurs — en tant que solution viable à plusieurs problèmes qui étaient considérés comme fondamentalement «humains» et quasi impossibles à résoudre pour une machine. Celle-ci représente un outil unique pour intégrer l'expertise des neuroanatomistes dans le processus de reconstruction de la matière blanche, sans avoir à fournir de règles explicitement. Un modèle peut ainsi apprendre la définition d'un chemin valide à partir d'exemples valides, pour ensuite reproduire ce qu'il a appris, sans répéter les erreurs classiques.

Plus particulièrement, les réseaux de neurones récurrents sont une famille de modèles créés spécifiquement pour le traitement de séquences de données. Comme une fibre de matière blanche est représentée par une séquence de points, le lien se fait naturellement. Malgré leur potentiel énorme, l'application des réseaux récurrents à la tractographie fait face à plusieurs problèmes techniques.

Cette thèse se veut très exploratoire, et détaille donc les débuts de l'utilisation des réseaux de neurones récurrents pour la tractographie par apprentissage, des problèmes qui sont apparus suite à la création d'une multitude d'algorithmes basés sur

SOMMAIRE

l'intelligence artificielle, ainsi que des solutions développées pour répondre à ces problèmes.

Les résultats de cette thèse ont démontré le potentiel des réseaux de neurones récurrents pour la reconstruction de la matière blanche, en plus de contribuer à l'avancement du domaine grâce à la création d'une base de données publique pour la tractographie par apprentissage.

Mots-clés: Neuroimagerie ; Tractographie ; Apprentissage Machine ; Réseaux de Neurones

Remerciements

Mon parcours à travers les études supérieures aura été quelque peu chaotique. Après un peu plus de six ans aux études supérieures — incluant trois enfants et une pandémie mondiale — atteindre la ligne d'arrivée du doctorat fut toute une épreuve, mais qui me remplit aujourd’hui de fierté. Évidemment, il s’agit d’un accomplissement qui ne se réussit pas seul, et je considère essentiel de mentionner tout le soutien que j’ai eu à travers les années.

D’abord, merci énormément à Lucie pour le support constant à travers toutes ces années d’études et de travail, et grâce à qui j’ai pu balancer le travail et la vie familiale avec Benoît, Florence et Robin.

Merci à mes parents Ghislain et Sylvie d’avoir toujours su m’encourager dans mes études et d’avoir eu confiance en moi, sans quoi je ne serai pas ici aujourd’hui.

Merci évidemment à Maxime Descoteaux et Pierre-Marc Jodoin qui m’ont accueilli au doctorat — même en sachant qu’un premier enfant était déjà attendu — et qui m’ont permis de concilier la vie familiale avec le doctorat. J’ai toujours été libre d’avancer à mon rythme, et je n’ai jamais senti de pression d’en faire plus. Je vous serai toujours reconnaissant d’avoir été aussi compréhensifs.

Merci à Adam Salvail et Hugo Larochelle de m’avoir rapatrié du milieu industriel en 2015 afin de commencer mes études supérieures. Merci également à Marc-Alexandre Côté pour tout le travail initial accompli sur le projet learn2track. On retrouve encore probablement des bouts de son code ici et là.

Merci aussi à François et Ann-Marie pour toutes les heures de gym et les nombreuses courses sur le Mont-Bellevue, qui m’ont motivé à retrouver la forme et de développer de bonnes habitudes malgré le temps limité entre les études et la famille.

Merci finalement à tous les collègues et amis à travers les années, du SMART —

REMERCIEMENTS

Marc-Alexandre, Stan, Mathieu — du VITAL — Antoine, Jon, Charles — du SCIL — François, Ann-Marie, Jasmeen, Guillaume, Gabrielle, Philippe, Matteo, Emmanuelle, Manon, Alex, Étienne — et du RECSUS — Luc, Joannie, Frédéric.

Abréviations

3D Trois dimensions

AD Diffusivité axiale (*Axial diffusivity*)

CPU *Central Processing Unit*

DTI Imagerie par tenseur de diffusion (*Diffusion Tensor Imaging*)

FA Anisotropie fractionnelle (*Fractional Anisotropy*)

fODF Fonction de distribution d'orientations des fibres (*Fiber Orientation Distribution Function*)

FRT *Funk-Radon Transform*

Go Gigaoctet

GPU Processeur graphique (*Graphical Processing Unit*)

GRU *Gated Recurrent Unit*

HARDI Imagerie de diffusion à haute résolution angulaire (*High Angular Resolution Diffusion Imaging*)

IRM Imagerie par résonance magnétique (*Magnetic Resonance Imaging*)

IRMd Imagerie par résonance magnétique de diffusion (*Diffusion-weighted Magnetic Resonance Imaging*)

LSTM *Long Short-Term Memory network*

MD Diffusivité moyenne (*Mean Diffusivity*)

ODF Fonction de distribution des orientations (*Orientation Distribution Function*)

RAM Mémoire vive (*Random Access Memory*)

RD Diffusivité radiale (*Radial Diffusivity*)

ABRÉVIATIONS

RNN Réseau de neurones artificiels récurrent (*Recurrent Neural Network*)

SH Harmoniques sphériques (*Spherical Harmonics*)

Table des matières

Sommaire	ii
Remerciements	iv
Abréviations	vi
Table des matières	viii
Liste des figures	xii
Liste des tableaux	xiv
Introduction	1
1 Contexte	4
1.1 Reconstruction de la matière blanche	4
1.1.1 Anatomie du cerveau	4
1.1.2 IRM de diffusion et représentations du signal	6
1.1.3 Tractographie	12
1.2 Évaluation des algorithmes de tractographie	16
1.3 Réseaux de neurones et séquences de données	19
1.3.1 Réseaux de neurones récurrents	19
1.3.2 <i>Long Short-Term Memory Network</i>	23
1.4 Éléments techniques liés à la tractographie par apprentissage avec réseaux de neurones récurrents	24
1.4.1 Signal d'entrée du modèle	25

TABLE DES MATIÈRES

1.4.2	Signal de sortie du modèle	25
1.4.3	Bruitage des tractes	26
1.4.4	Cisaillement des tractes	26
1.4.5	Ajout du voisinage en entrée	27
1.4.6	Échantillonnage semi-aléatoire des tractes	27
2	Tractographie par apprentissage	29
2.1	Introduction	32
2.2	Using Deep Learning for Tractography	33
2.2.1	Models	33
2.2.2	Tractography	35
2.3	Related Work	35
2.4	Experiments	36
2.4.1	ISMRM2015 Challenge	37
2.4.2	<i>In vivo</i> tracking	38
2.5	Conclusion	39
3	État de l'art et défis futurs en tractographie par apprentissage	41
3.1	Introduction	44
3.2	Annotated datasets and evaluation tools	47
3.2.1	The FiberCup dataset and the Tractometer tool	48
3.2.2	Tracula (2011)	50
3.2.3	HARDI Reconstruction Challenges	51
3.2.4	ISMRM Tractography Challenge (2015)	52
3.2.5	HAMLET (2018)	54
3.2.6	Datasets based on the BIL&GIN database	54
3.2.7	Datasets based on the HCP database	55
3.3	Machine learning methods for tractography	56
3.4	Results & Discussion	61
3.4.1	Results on the 2015 ISMRM Tractography Challenge	61
3.4.2	The 2015 ISMRM Tractography Challenge as an evaluation tool for ML algorithms	63
3.4.3	Other results	66

TABLE DES MATIÈRES

3.4.4	Limitations of machine learning algorithms for tractography	66
3.4.5	Proposed guidelines for a data-driven tractography evaluation framework	69
3.5	Conclusion	72
4	TractoInferno	73
4.1	Introduction	76
4.2	Datasets	78
4.2.1	Mazoyer et. al - BIL & GIN	78
4.2.2	Tsushida et. al - MRI-Share	79
4.2.3	DeLuca et. al - Bilingualism and the brain	80
4.2.4	Poldrack et. al - UCLA CNP	80
4.2.5	Tamm et. al - The Stockholm Sleepy Brain Study	80
4.2.6	Tremblay et. al - mTBI and Aging study (controls)	81
4.3	Methodology	81
4.3.1	Raw data QC	83
4.3.2	TractoFlow pipeline	84
4.3.3	TractoFlow results QC	84
4.3.4	Ensemble tractography	85
4.3.5	Bundle segmentation with RBX	86
4.3.6	Bundle segmentation QC	87
4.4	Evaluation pipeline for candidate tractograms	89
4.5	RNN-based tractography	89
4.5.1	Learn2track	90
4.5.2	DeepTract	90
4.5.3	Entrack	91
4.5.4	Gaussian distribution output	92
4.5.5	Gaussian mixture distribution output	92
4.5.6	Implementation details	93
4.6	Results & Discussion	94
4.7	Conclusion	97
4.8	Data access	98

TABLE DES MATIÈRES

4.9 Special thanks	98
4.10 Acknowledgements	98
4.A <i>TractoInferno</i> pipeline execution commands	98
4.A.1 QC DWI	98
4.A.2 TractoFlow	99
4.A.3 QC TractoFlow	99
4.A.4 SH signal fitting	99
4.A.5 Tractography (Det, Prob, PFT)	100
4.A.6 Tractography (SET)	100
4.A.7 RecoBundlesX (RBX)	100
4.A.8 QC RBX	101
Conclusion	102
A Tractographie par apprentissage spécifique aux faisceaux	110
A.1 Introduction	114
A.2 Related work	115
A.3 Methodology	116
A.3.1 Dataset	116
A.3.2 Model	116
A.3.3 Model inputs	116
A.3.4 Reference streamlines	117
A.3.5 Evaluation	118
A.4 Results	119
A.5 Conclusion	122
Bibliographie	123

Liste des figures

1.1	Éléments principaux de l'anatomie du cerveau	5
1.2	Anatomie du neurone	6
1.3	Mouvement des molécules d'eau	7
1.4	Carte d'anisotropie fractionnelle.	9
1.5	Harmoniques sphériques modifiées	11
1.6	Schéma de l'origine de l'ODF de diffusion	12
1.7	Schématisation du processus de tractographie	13
1.8	Illustration de différentes configurations de fibres	15
1.9	Représentation de la discréétisation d'un faisceau pour évaluation en masque d'incidence binaire.	16
1.10	Métriques de chevauchement et dépassement, et coefficient de Dice.	18
1.11	Applications pour un réseau de neurones récurrent	20
1.12	Schéma du LSTM	24
1.13	Traitement des données pour tractographie avec réseaux récurrents	28
2.1	Architecture of the proposed models for machine learning tractography	34
2.2	Comparison of RNN predictions against ground truth	39
2.3	Expert, FFNN and RNN bundles obtained in experiment 2.4.2	40
3.1	Data generation process of the 2015 ISMRM Tractography Challenge	52
3.2	Original and new results for the 2015 ISMRM Tractography Challenge	62
3.3	Comparison of Random-Forest and deterministic tractography	67
3.4	Comparison of various tracking methods	67
4.1	<i>TractoInferno</i> processing pipeline	82

LISTE DES FIGURES

4.2 Examples of HTML pages generated by dmriqcpcy for data QC	83
4.3 Reconstruction of the <i>Corpus Callosum</i>	95
4.4 Reconstruction of the <i>Optic Radiation</i>	96
4.5 Reconstruction of the <i>Pyramidal Tract</i>	97
A.1 BW-DeepTracker model	117
A.2 Tracked bundles on BIL&GIN test subject t0203	121

Liste des tableaux

2.1	Quantitative evaluation on the ISMRM 2015 Tractography Challenge	38
3.1	Annotated datasets	49
3.2	Main properties of data-driven methods for tractography	57
3.3	Tractometer results	63
3.4	Data differences for data-driven methods	63
4.1	Original datasets metadata	79
4.2	List of bundles in the default RBX atlas	87
4.3	Automatic rating criteria	88
4.4	<i>TractoInferno</i> test results	94
A.1	Evaluation metrics on the test set	120

Introduction

Le développement de l'imagerie par résonance magnétique dans les années 80 [75] a créé une révolution dans le domaine de la neuro-imagerie en permettant l'étude du cerveau humain de façon non invasive. Plus précisément, au début des années 90, le développement de l'imagerie par tenseur de diffusion a permis l'étude de la matière blanche et des connexions cérébrales à partir du déplacement des molécules d'eau [6]. La reconstruction de fibres de matière blanche, ou tractographie, permet notamment d'étudier la connectivité du cerveau, de guider certaines interventions chirurgicales, et même d'examiner l'évolution de pathologies neurodégénératives.

La tractographie repose généralement sur l'idée de base qu'il est possible de reconstruire une fibre de matière blanche, en faisant des sauts à répétition [65]. Pour faire la distinction entre une fibre réelle et une reconstruction, on désignera dans cette thèse la fibre reconstruite par une nouvelle terminologie, la tracte. On utilise ainsi l'information de diffusion de diverses façons afin de choisir quelle direction prendre, tout en considérant divers paramètres afin de guider le processus de reconstruction. D'autres méthodes dites globales existent également, mais reposent essentiellement sur le même concept de décisions locales.

Cette façon de faire comporte un lot de problèmes qui seront détaillés plus loin [122, 117]. La plupart de ces problèmes découlent du fait que la décision de faire un pas dans une certaine direction repose sur de l'information locale, sans jamais considérer le chemin parcouru par la tracte (on considère habituellement la direction précédente uniquement) [82]. Pour pallier à ces problèmes, on a recours, soit à un expert anatomiste qui pourra manuellement choisir les tractes valides ou pertinentes, soit à un ou plusieurs algorithmes de filtrage afin de retirer le plus de tractes invalides possible en fonction de certains critères prédéfinis.

INTRODUCTION

La solution peut alors sembler évidente : on doit intégrer plus d'information dans le processus de décision de la tractographie. Ainsi, les avancées relativement récentes en apprentissage profond (*Deep Learning*) sont extrêmement pertinentes pour ce problème, où on constate qu'il est possible d'utiliser les tractes valides, de façon supervisée, afin d'entraîner un modèle qui pourra apprendre à reproduire les décisions qui mènent à des résultats valides [103]. On appellera alors cette famille de solutions la tractographie par apprentissage. Cependant, comme le démontrera cette thèse, la solution est plus compliquée qu'elle n'y paraît.

Cette thèse se veut donc une exploration de la tractographie par apprentissage, des problèmes qui y sont reliés, et des pistes de solutions à explorer pour palier à ceux-ci. Notons d'abord que ce travail juxtapose l'intelligence artificielle et la neuro-imagerie. Bien que l'apprentissage profond soit inspiré de la neurobiologie, dans le cadre de cette thèse, ce lien est fortuit. Ainsi, toute mention de réseau de neurones sert à décrire le modèle mathématique d'un réseau de neurones artificiels, et pas pour décrire une structure anatomique. L'intelligence artificielle est utilisée ici comme un outil parmi d'autres afin d'accomplir un travail indépendant, soit la reconstruction de faisceaux de matière blanche.

Le [chapitre 1](#) présente les bases de l'IRM de diffusion, qui permet l'exploration non invasive du cerveau, puis comment celle-ci est utilisée pour la reconstruction de fibres de matière blanche. Les algorithmes traditionnels de tractographie y seront décrits, ainsi que les problèmes qui en découlent. Ensuite, on retrouve une brève explication de l'apprentissage machine, et plus spécifiquement du traitement et de l'apprentissage de séquences de données.

Le [chapitre 2](#) présente un algorithme de tractographie par apprentissage avec contexte spatial, nommé *Learn to track*. Ce nouvel algorithme propose l'utilisation des réseaux de neurones récurrents pour apprendre la direction à choisir dans le cadre de la reconstruction de fibres de matière blanche, tout en intégrant l'information du chemin déjà parcouru par une fibre afin de faire de meilleures prédictions. *Learn to track* a été le tout premier algorithme récurrent proposé pour la tractographie.

L'[annexe A](#) est une suite logique au [chapitre 2](#), puisqu'il s'agit d'une extension de *Learn to track*, avec un modèle amélioré, une base de données réelles, et une approche spécifique aux faisceaux.

INTRODUCTION

Ensuite, au fil des années, plusieurs nouveaux algorithmes de tractographie par apprentissage ont été proposés. Le [chapitre 3](#) soulève les problèmes présents à travers toutes les méthodes proposées, et qui demeurent une préoccupation lors de la comparaison des résultats obtenus. Il propose aussi plusieurs pistes de solutions afin de pouvoir comparer les résultats de futurs algorithmes.

Pour faire face aux problèmes soulevés dans le chapitre précédent, le [chapitre 4](#) présente la création d'une nouvelle base de données, *TractoInferno*, visant spécifiquement la tractographie par apprentissage et les problèmes qui y sont reliés.

Chapitre 1

Contexte

Cette thèse concerne la tractographie par apprentissage, c.-à-d. l'utilisation de l'apprentissage machine afin de développer de nouveaux algorithmes de reconstruction de la matière blanche. Ce chapitre est un survol des notions de bases nécessaires à la compréhension des concepts utilisés dans les chapitres subséquents.

1.1 Reconstruction de la matière blanche

Cette section développe les aspects essentiels liés à la tractographie. On y explique certains éléments anatomiques du cerveau et de la matière blanche qui le compose. On détaille ensuite comment acquérir et analyser des images du cerveau à l'aide du mouvement des molécules d'eau et de l'imagerie par résonance magnétique de diffusion. Finalement, on présente les principes de base de la reconstruction de la matière blanche.

1.1.1 Anatomie du cerveau

Le cerveau est un organe complexe que l'on cherche encore à comprendre. Il est composé de plusieurs milliards de neurones organisés en faisceaux, qui communiquent entre eux via les diverses connexions qu'ils forment. On identifie dans le cerveau trois milieux distincts : le liquide cérébro-spinal, la matière blanche, et la matière grise. Le

1.1. RECONSTRUCTION DE LA MATIÈRE BLANCHE

tissu qui nous intéresse est la matière blanche, qui est composée d'axones et dont le rôle est de transmettre l'information le long du neurone, du corps cellulaire jusqu'aux terminaisons axonales. La matière grise contient principalement le corps cellulaire des neurones et les terminaisons axonales, où l'information est reçue et transmise d'un neurone à un autre. Celle-ci se distingue en deux groupes : la matière grise du ruban cortical, qui recouvre la surface du cerveau et forme les *gyri* et *sulci*, ainsi que les noyaux gris centraux, qui forment des régions de connexions d'axones à l'intérieur du cerveau. Finalement, le liquide céphalo-rachidien, situé dans les ventricules et autour du cerveau, sert à protéger le cerveau. La [figure 1.1](#) illustre les structures d'intérêt dans le cerveau.

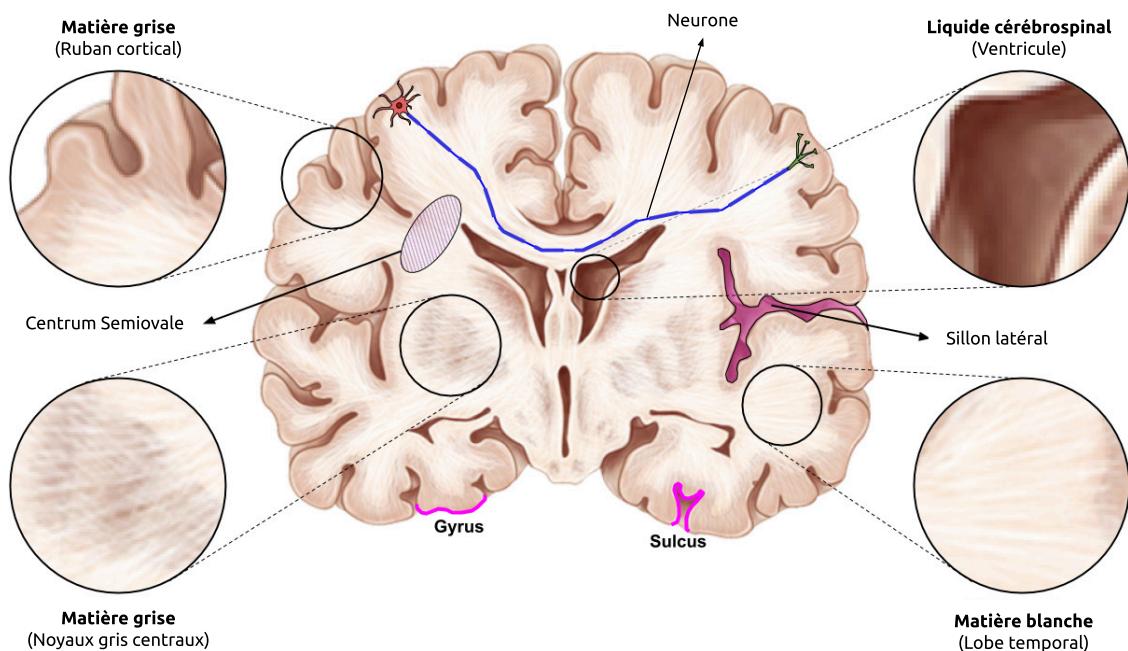


Figure 1.1 – Éléments principaux de l'anatomie du cerveau. (Figure adaptée de [118], courtoisie de François Rheault.)

Le neurone est donc composé d'un corps cellulaire et d'un long filament (axone) qui permet de conduire les signaux électriques jusqu'aux terminaisons axonales, tel qu'illustré dans la [figure 1.2](#). La gaine de myéline qui recouvre l'axone permet la transmission efficace des signaux électriques.

1.1. RECONSTRUCTION DE LA MATIÈRE BLANCHE

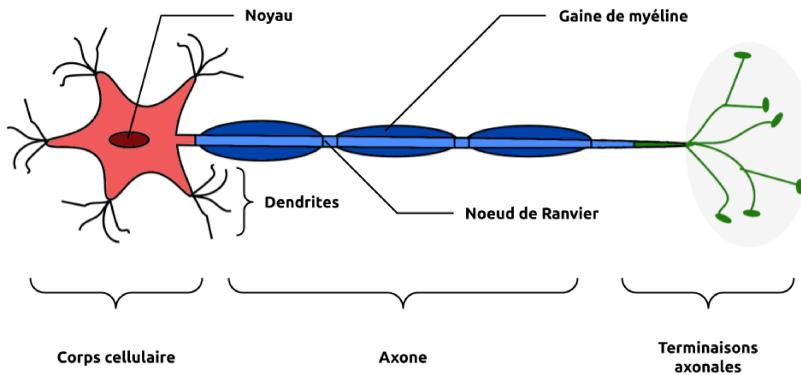


Figure 1.2 – Anatomie simplifiée du neurone.

Pour explorer le cerveau humain de manière non invasive, l'imagerie par résonance magnétique (IRM) est maintenant l'outil de choix. L'IRM est une technique créée dans les années 1980 [75] (mais dont le principe remonte aux années 1950) qui permet de reconstruire des images anatomiques en utilisant les propriétés magnétiques des tissus.

1.1.2 IRM de diffusion et représentations du signal

Le rôle de la tractographie est de reconstruire le chemin parcouru par les fibres de matière blanche du cerveau à partir du signal mesuré par l'IRM de diffusion. Comme un faisceau de fibres de matière blanche est composé de plusieurs axones rassemblés ensemble de façon compacte, le déplacement de l'eau est globalement restreint dans le sens du faisceau, tel qu'ilustré dans la figure 1.3.

À l'aide de l'IRM de diffusion, on peut obtenir une mesure liée au déplacement des molécules d'eau dans un axe choisi. Pour obtenir un modèle de la diffusion en trois dimensions, on acquiert des images de diffusion dans plusieurs directions différentes.

L'imagerie par tenseur de diffusion (DTI : *Diffusion Tensor Imaging*) — inventée en 1994 par Peter Basser [5, 6] — permet de modéliser la diffusion des molécules d'eau par un modèle gaussien en trois dimensions. Le signal atténué S_k est modélisé par le tenseur de diffusion via l'équation :

$$S_k = S_0 e^{-b\hat{\theta}_k^T \mathbf{D} \hat{\theta}_k}, \quad (1.1)$$

1.1. RECONSTRUCTION DE LA MATIÈRE BLANCHE

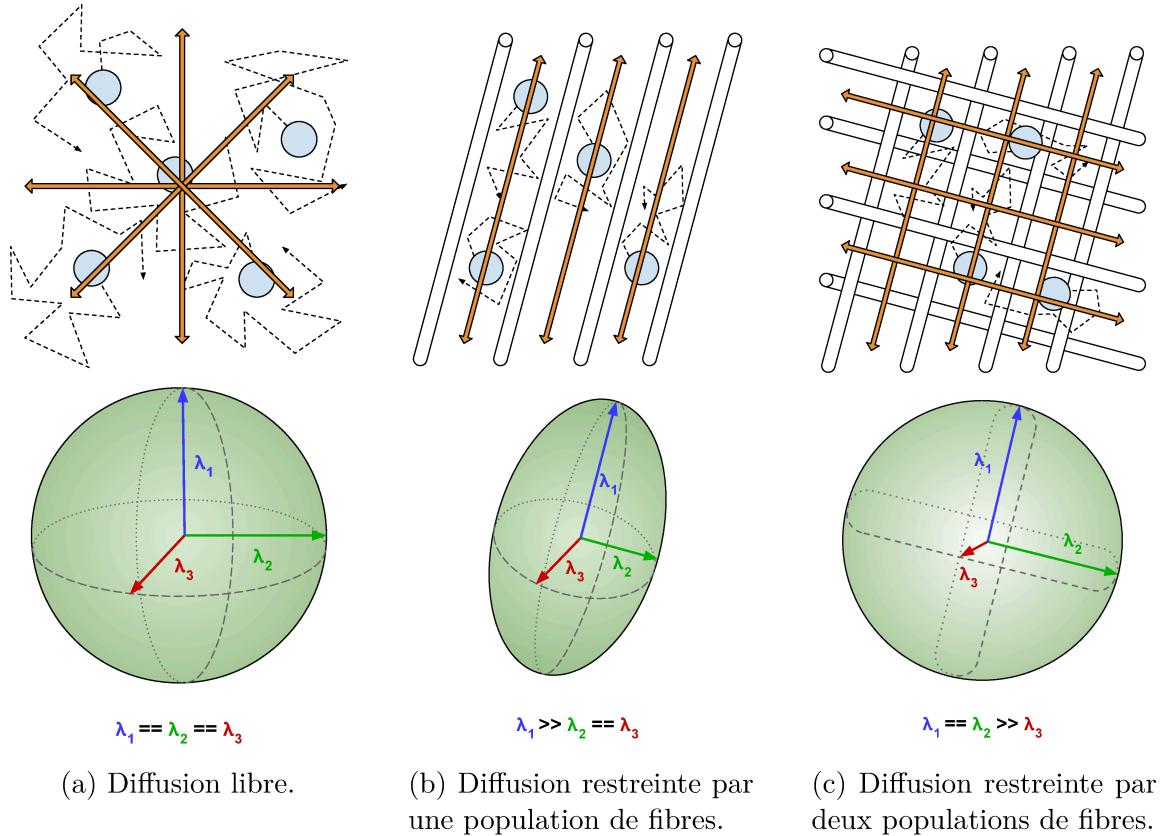


Figure 1.3 – Schéma du déplacement aléatoire des molécules d'eau dans un milieu libre ou restreint par différentes populations de fibres, avec le tenseur de diffusion associé et sa décomposition en vecteurs/valeurs propres. (a) Sans restriction, les molécules d'eau se déplacent librement ; le tenseur de diffusion est sphérique (les 3 valeurs propres sont égales). (b) Dans un milieu restreint par une seule population de fibres, les molécules se déplacent plus facilement le long de l'axe des fibres ; le tenseur de diffusion est allongé (une valeur propre est plus grande que les autres). (c) Dans un milieu restreint par deux populations de fibres perpendiculaires, le déplacement des molécules d'eau correspond aux deux axes parallèles aux fibres. L'hypothèse gaussienne n'est pas adaptée ; le tenseur de diffusion ne permet pas de distinguer l'orientation des fibres.

1.1. RECONSTRUCTION DE LA MATIÈRE BLANCHE

où S_0 représente le signal mesuré sans gradient de diffusion, \hat{g}_k le vecteur unitaire de la direction du gradient de diffusion, et b est le facteur-b (*b-value*) qui dépend de la séquence d'acquisition [74] (force, durée et délai entre l'application des gradients de champ magnétique). Un facteur-b d'une valeur de 1000 s/mm^2 est typiquement utilisé pour l'imagerie par tenseur de diffusion [18].

Le tenseur de diffusion \mathbf{D} est calculé en solvant l'[équation 1.1](#), à l'aide d'au moins six images avec des directions de gradient \hat{g}_k différentes, plus une image S_0 sans gradient de diffusion. Il correspond à la matrice de covariance symétrique de la distribution gaussienne :

$$\mathbf{D} = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{xy} & D_{yy} & D_{yz} \\ D_{xz} & D_{yz} & D_{zz} \end{bmatrix}. \quad (1.2)$$

Les éléments propres de cette matrice décrivent une surface ellipsoïde, dont l'orientation est déterminée par le premier vecteur propre. Les valeurs propres de cette matrice sont utilisées pour calculer certaines propriétés liées à la forme du tenseur, qui permettent de décrire le comportement général de la diffusion des molécules d'eau, comme la diffusion axiale (AD, [équation 1.3](#)), la diffusion radiale (RD, [équation 1.4](#)), la diffusion moyenne (MD, [équation 1.5](#)) et l'anisotropie fractionnelle (FA, [équation 1.6](#)), données par les équations suivantes :

$$AD = \lambda_1; \quad (1.3)$$

$$RD = \frac{\lambda_2 + \lambda_3}{2}; \quad (1.4)$$

$$MD = \frac{\lambda_1 + \lambda_2 + \lambda_3}{3}; \quad (1.5)$$

$$FA = \sqrt{\frac{1}{2} \sqrt{\frac{(\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_3 - \lambda_1)^2}{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}}}. \quad (1.6)$$

La carte d'anisotropie fractionnelle — la mesure la plus connue et utilisée du modèle DTI — est illustrée dans la [figure 1.4](#). Elle décrit le niveau d'anisotropie dans chaque voxel [99]. Une valeur de FA élevée indique une diffusion fortement orientée dans une seule direction, alors qu'une valeur faible indique plutôt un milieu isotrope (diffusion

1.1. RECONSTRUCTION DE LA MATIÈRE BLANCHE

égale dans toutes les directions). Par contre, le modèle gaussien du tenseur de diffusion est imparfait ; par exemple, un croisement de trois populations de fibres orthogonales produirait une valeur faible de FA, identique à un milieu isotrope.

Ce tenseur assume un déplacement gaussien des molécules d'eau, et permet une reconstruction limitée des fibres de matière blanche. En effet, en présence d'une seule population de fibres, la direction principale du modèle de tenseur est très corrélée avec l'orientation principale des fibres. Par contre, en présence d'un croisement de fibres avec différentes orientations, la direction principale du tenseur est ambiguë et empêche une reconstruction robuste. Malgré les limitations du tenseur DTI, ses propriétés sont encore utilisées aujourd'hui pour décrire et expliquer certaines observations liées à la matière blanche du cerveau.

Depuis son invention en 2002, l'imagerie de diffusion à haute résolution angulaire (HARDI : *High Angular Resolution Diffusion Imaging*) [145, 97, 144] est le standard en IRM de diffusion pour faire de la tractographie. Pour une acquisition HARDI à simple coquille (traduction libre de *single-shell*), on mesure le signal dans plusieurs directions réparties uniformément sur une sphère en espace-q, en conservant le même facteur-b. Par exemple, pour utiliser un modèle à déconvolution sphérique, on recommande d'utiliser 45 directions de gradients, et un facteur-b d'environ 3000 s/mm^2 [138, 18]). Ce signal est discret, et les orientations spécifiques peuvent varier selon le protocole employé. Ainsi, il est plus pratique de représenter le signal discret comme une fonction continue sur la sphère. La représentation la plus utilisée emploie les harmoniques

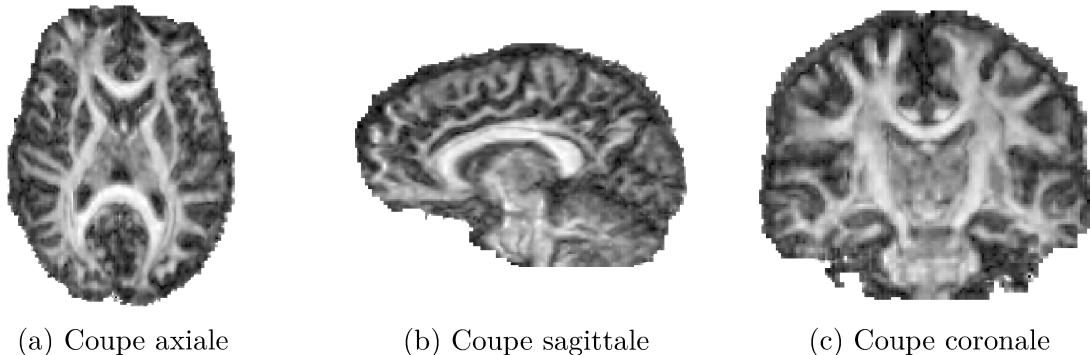


Figure 1.4 – Carte d'anisotropie fractionnelle. Image provenant d'un sujet de la base de données *TractoInferno* Poulin et al..

1.1. RECONSTRUCTION DE LA MATIÈRE BLANCHE

sphériques (SH : *Spherical Harmonics*), qui consiste à utiliser une base de fonctions sphériques et de calculer les coefficients qui approximent le mieux le signal discret mesuré par l'IRM.

Spécifiquement, on utilise couramment la base d'harmoniques sphériques modifiées définie par [34], qui forme une base réelle, symétrique et orthonormale pour représenter des fonctions sur la sphère. Cette base est illustrée dans la figure 1.5. Le signal ainsi représenté est noté par

$$S(\theta_i, \phi_i) = \sum_{j=1}^R c_j Y_j(\theta_i, \phi_i), \quad (1.7)$$

où θ_i et ϕ_i sont les directions d'acquisition du signal brut (en coordonnées polaires), $Y_j(\theta_i, \phi_i)$ sont les fonctions de la base modifiée de [34], et c_j sont les coefficients à calculer pour approximer le signal original. La valeur R représente l'indice du dernier coefficient, et est calculée en fonction de l'ordre maximal L que l'on désire utiliser :

$$R = (1/2)(L + 1)(L + 2). \quad (1.8)$$

Comme mentionné précédemment, une image acquise en appliquant un gradient de diffusion a une intensité faible si les molécules d'eau sont libres de diffuser dans la direction du gradient. Les données ont donc besoin d'être modifiées pour en extraire les directions principales de diffusion. Ainsi, la transformée de Funk-Radon (FRT) [144, 36] permet de passer du signal mesuré (inversement proportionnel au déplacement des molécules d'eau) à une fonction de distribution d'orientations (ODF : *Orientation Distribution Function*) de diffusion. Cette ODF de diffusion correspond à la distribution des orientations du déplacement des molécules d'eau, sur la sphère. L'utilisation de la base d'harmoniques sphériques pour le signal permet de définir une solution analytique pour la transformée de Funk-Radon, dont le résultat est l'ODF de diffusion, représentée également par les harmoniques sphériques.

Par contre, l'ODF de diffusion n'est pas appropriée pour la tractographie puisqu'elle est peu robuste aux croisements de fibres (les croisements à angle plus petit produisent des pics moyennés et mal définis). Un modèle comme la déconvolution sphérique [138] peut donc être utilisé pour transformer l'ODF de diffusion en ODF de

1.1. RECONSTRUCTION DE LA MATIÈRE BLANCHE

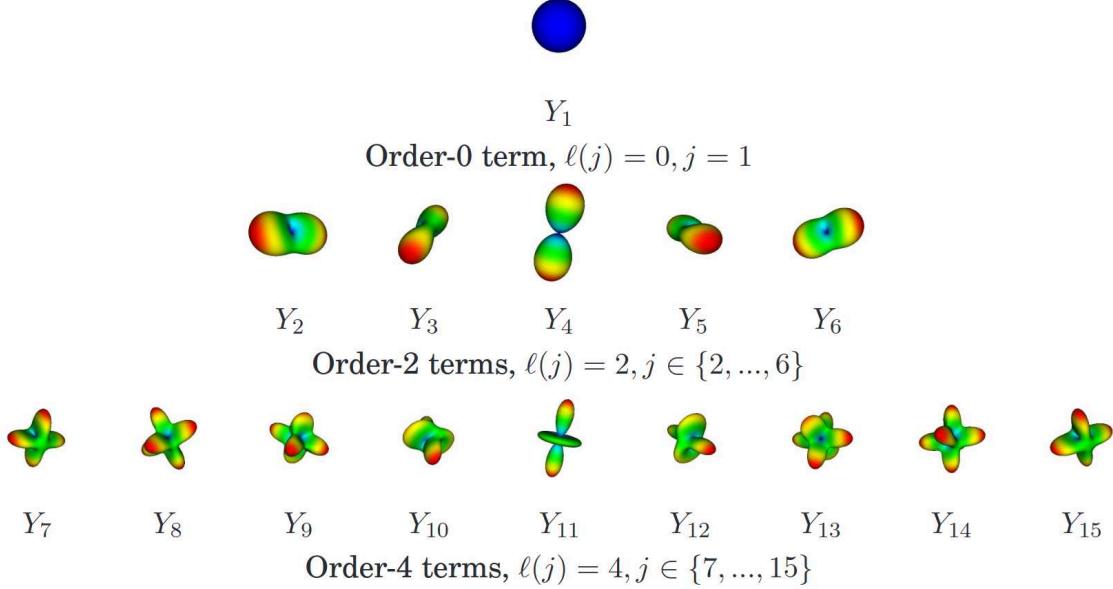


Figure 1.5 – Harmoniques sphériques modifiées de [34], jusqu'à l'ordre 4. Image tirée de [34].

fibre (fODF). Ses pics sont alignés avec la diffusion, et sont plus prononcés que l'ODF de diffusion. Cependant, la déconvolution sphérique suppose un profil de diffusion de fibres de matière blanche. Ce profil de diffusion de fibres est soit fixé manuellement, soit calculé à partir de quelques voxels spécifiques qui contiennent une seule population de fibres (comme au cœur du corps calleux, ou *corpus callosum*). Un profil de diffusion unique est généralement utilisé pour modéliser toutes les fibres à travers la matière blanche. Il s'agit d'une limitation importante, puisqu'on considère que le profil de diffusion est constant, alors qu'il varie d'un faisceau à l'autre, d'une acquisition à une autre, ou même d'un fabricant de scanners à un autre. En pratique, choisir un bon profil de diffusion est encore une question ouverte. On verra plus tard comment l'utilisation d'un modèle par apprentissage permet de se départir de l'hypothèse du profil de diffusion constant, en utilisant le signal de diffusion comme entrée du modèle.

En plus de l'acquisition HARDI à simple coquille, on peut également faire une acquisition à multiples coquilles (traduction libre de *multi-shell*). Dans ce cas, on procède à plusieurs acquisitions à simple coquille (généralement deux ou trois), mais

1.1. RECONSTRUCTION DE LA MATIÈRE BLANCHE

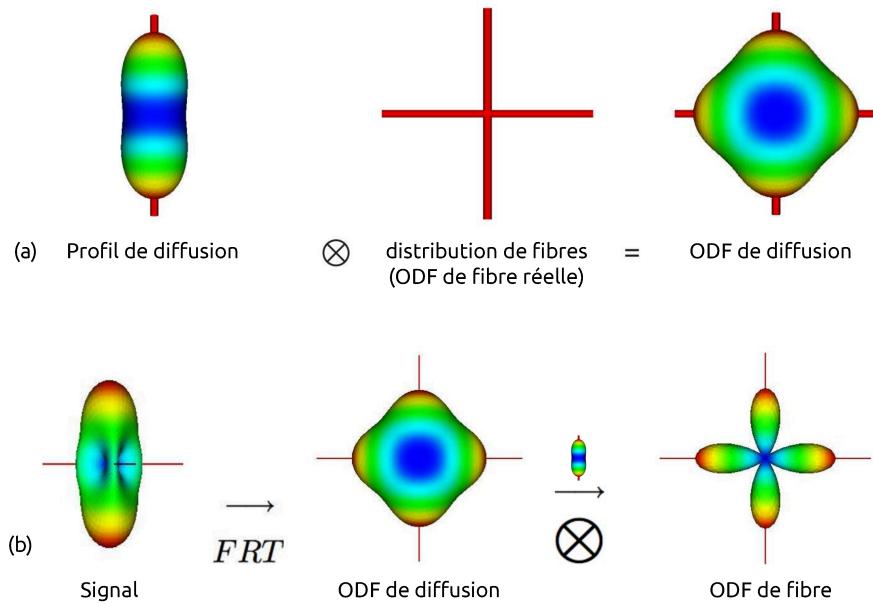


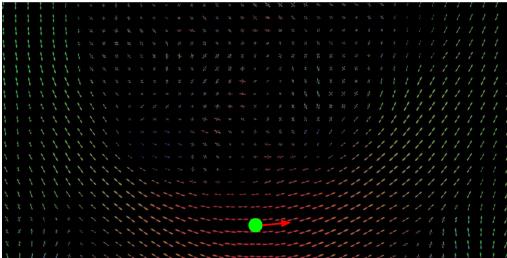
Figure 1.6 – Schéma de l'origine de l'ODF de diffusion, et modélisation du signal en ODF de fibre. (a) La convolution du profil de diffusion à une population de fibre avec la réelle distribution de fibres donne l'ODF de diffusion. (b) Le signal mesuré est transformé en ODF de diffusion via la transformée de Funk-Radon (FRT). La déconvolution de l'ODF de diffusion avec un profil de diffusion permet d'obtenir une ODF de fibre. Figure adaptée de [34].

en utilisant un facteur- b différent à chaque fois (e.g. $b = [1000, 2000, 3000] \text{ s/mm}^2$). Une amélioration du modèle à déconvolution sphérique permet d'utiliser ces données afin d'extraire des ODF spécifiques à certains tissus (matière blanche, matière grise, liquide cérébro-spinal), en supposant un profil de diffusion à multiples coquilles différent pour chaque tissu.

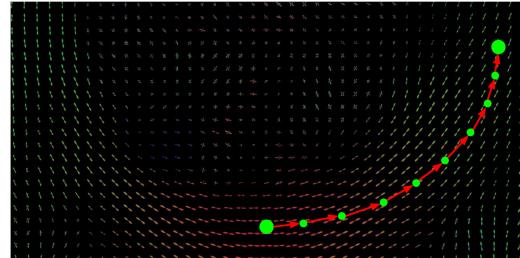
1.1.3 Tractographie

La tractographie consiste à reconstruire le plus fidèlement possible les fibres de matière blanche qui composent le cerveau. Les méthodes traditionnelles les plus couramment utilisées se basent sur l'ODF de fibre pour faire une reconstruction itérative. On choisit d'abord un point de départ (soit dans la matière blanche, soit à l'intersection de la matière blanche et de la matière grise), puis un chemin est tracé pas à

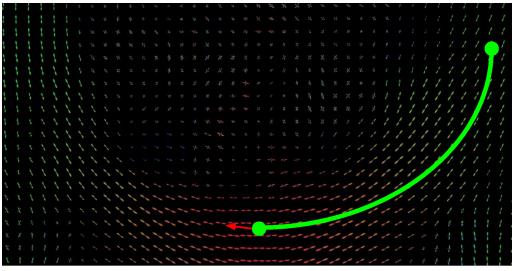
1.1. RECONSTRUCTION DE LA MATIÈRE BLANCHE



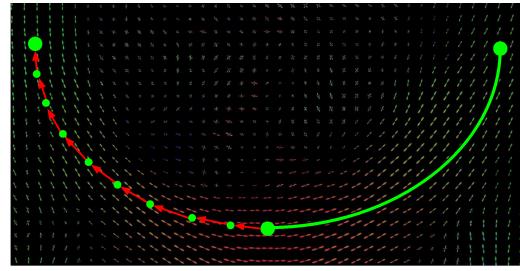
(a) Un point de départ est choisi aléatoirement dans la matière blanche (ici, au cœur du corps calleux).



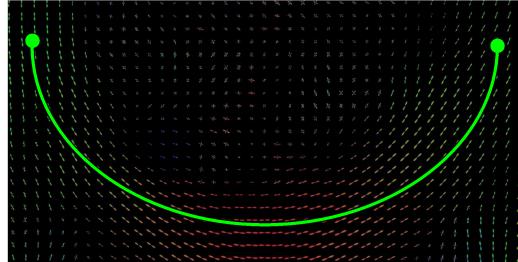
(b) De façon itérative, on choisit une direction alignée avec l'ODF de fibre à la position courante, jusqu'à atteindre un critère d'arrêt.



(c) Le processus est réinitialisé, en sens inverse pour compléter la tracte.



(d) De nouveau, l'ODF de fibre est utilisée pour choisir la direction à suivre.



(e) Les deux segments de tracte sont fusionnés pour former une reconstruction complète.

Figure 1.7 – Schématisation du processus de tractographie, avec initialisation dans la matière blanche.

pas, en prenant une décision locale qui dépend de l'ODF de fibre à chaque position. La reconstruction se termine lorsque la tracte se retrouve à l'extérieur de la matière blanche. Si le point de départ était dans la matière blanche, un deuxième segment

1.1. RECONSTRUCTION DE LA MATIÈRE BLANCHE

est généré dans le sens inverse (à partir du même point de départ), afin d'obtenir une tracte complète qui connecte idéalement deux régions de matière grise. En répétant ce processus des milliers de fois avec des points de départ différents, on reconstruit un tractogramme, où chaque tracte est représentée par une série de points dans l'espace 3D.

À partir de ce principe de base pour la tractographie, deux familles principales de méthodes existent : les méthodes déterministes et les méthodes probabilistes. Les méthodes déterministes suivent la direction principale de l'ODF de fibre la plus alignée avec la direction précédente, ce qui mène à un tractogramme plus restreint spatialement, mais qui contient moins de faux positifs et donc convient davantage à des études de connectivité. Les méthodes probabilistes échantillonnent plutôt une direction à l'intérieur d'un cône aligné avec la direction précédente, proportionnellement à l'amplitude de l'ODF de fibre. Ainsi, pour un même point de départ, on peut obtenir plusieurs tractes qui représentent l'incertitude du processus de tractographie. Le tractogramme résultant est plus varié, mais contient également plus de faux positifs [120]. Cette méthode convient donc davantage à la reconstruction de faisceaux. Chaque type de méthode a donc ses forces et ses faiblesses, et le choix de l'une ou l'autre dépend de plusieurs facteurs. D'autres algorithmes existent, dont quelques-uns sont détaillés dans le [chapitre 4](#).

Les algorithmes classiques de tractographie comportent cependant certains problèmes [117]. En effet, la tractographie basée sur des décisions locales produit des erreurs courantes puisqu'elle ne considère pas la structure globale complexe du cerveau. Il est très difficile pour un modèle local de différencier entre les différentes configurations possibles dans les zones de croisement de faisceaux. En effet, certaines configurations sont simplement impossibles à résoudre sans contexte, comme illustré dans la [figure 1.8](#). En conséquence, un tractogramme complet peut contenir jusqu'à quatre fois plus de mauvaises connexions que de connexions valides [82]. Même une amélioration de la résolution des machines ne peut résoudre ce problème. Le fait de prendre une décision locale suppose donc que le chemin déjà parcouru par une fibre n'a aucune influence sur les prochaines directions à prendre. Certes, des algorithmes globaux existent [23, 111], mais ceux-ci comportent une multitude d'hyperparamètres à optimiser, nécessitent davantage de temps de calcul, et n'obtiennent généralement

1.1. RECONSTRUCTION DE LA MATIÈRE BLANCHE

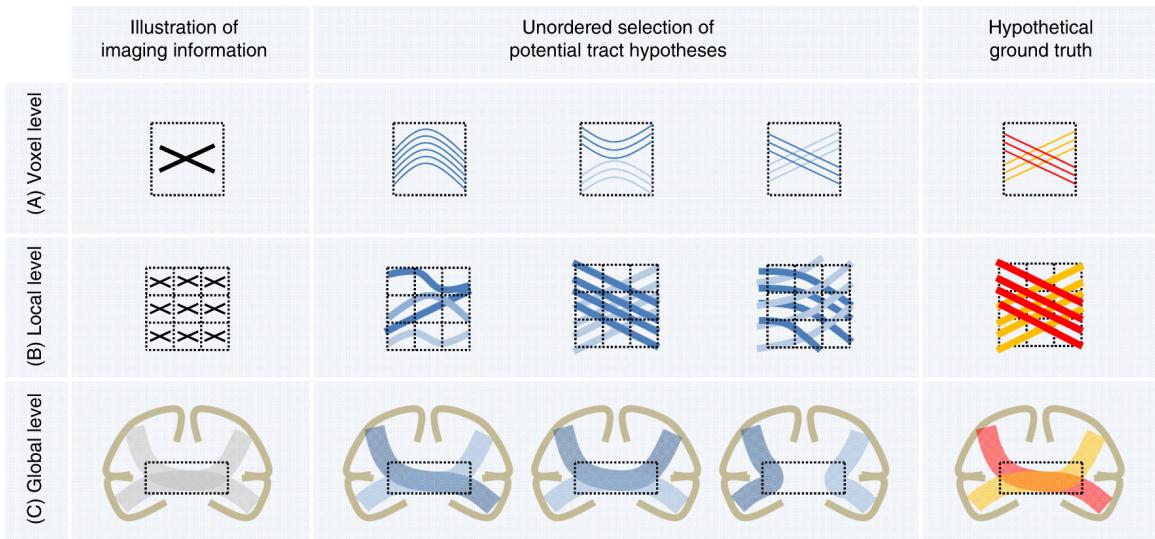


Figure 1.8 – Illustration de différentes configurations de fibres qui mènent à la même acquisition. Cet exemple met en évidence le fait que le contexte, ou le «passé» d'une tracte, est une information importante lors de la reconstruction. Les études histologiques nous permettent de savoir quelles configurations sont anatomiquement valides, mais il reste encore difficile d'intégrer cette connaissance dans le processus de reconstruction. Figure tirée de [82].

pas de meilleurs résultats que les modèles locaux [82].

Pour résoudre une partie des problèmes liés à l'incertitude des pics d'ODF de fibre dans des régions de croisements, des approches spécifiques aux faisceaux ont été proposées, soit de façon classique [115] ou à l'aide d'apprentissage machine Wasserthal et al.. Dans les deux cas, l'idée de base consiste à modifier les ODF de fibres de sorte à mieux représenter un faisceau d'intérêt, puis d'utiliser un algorithme de tractographie traditionnel pour reconstruire le faisceau. Dans le même ordre d'idées, l'approche récurrente du *Bundle-Wise Deep Tracker* (annexe A) a été développée sans nécessiter de modification aux ODF de fibres, mais plutôt en entraînant un modèle récurrent uniquement avec les tractes du faisceau d'intérêt.

1.2. ÉVALUATION DES ALGORITHMES DE TRACTOGRAPHIE

1.2 Évaluation des algorithmes de tractographie

Avant d'évaluer la tractographie par apprentissage, on doit d'abord s'intéresser à l'évaluation des méthodes de tractographie traditionnelles. En effet, l'évaluation des méthodes de tractographie est une tâche ardue, puisque la vérité terrain est généralement inconnue, à moins d'utiliser des modèles simulés. De plus, la tractographie produit en sortie des tractes, représentées par des séquences de points dans un espace en trois dimensions, qui ne peuvent pas être évaluées directement. L'évaluation d'un tractogramme dépend donc généralement de la tâche à accomplir, et plusieurs compétitions ont été organisées à travers les années pour comparer les algorithmes existants et évaluer divers aspects de la tractographie [122].

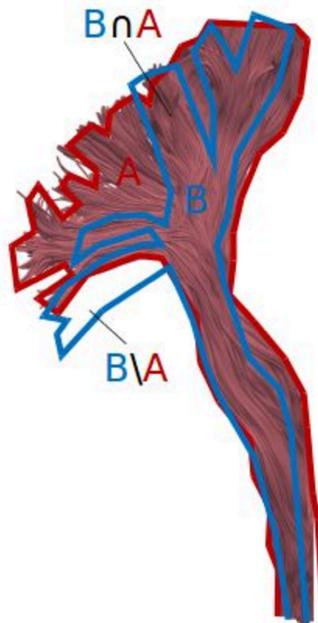


Figure 1.9 – Représentation de la discréttisation d'un faisceau pour évaluation en masque d'incidence binaire. **A (rouge)** : la discréttisation d'un faisceau de référence. **B (bleu)** : la discréttisation d'un faisceau à évaluer. L'intersection des faisceaux ($B \cap A$) ainsi que le volume externe au faisceau de référence ($B \setminus A$) sont utilisés dans le calcul des métriques d'évaluation. Figure adaptée du site *Tractometer*¹.

1. www.tractometer.org

1.2. ÉVALUATION DES ALGORITHMES DE TRACTOGRAPHIE

Pour évaluer la connectivité, on s'intéresse aux connexions corticales réalisées par les tractes générées. On peut utiliser un modèle simulé dont la vérité terrain est connue, comme les données du *ISMRM 2015 Tractography Challenge* [82], conjointement avec l'outil *Tractometer* proposé dans [27]. On évalue alors les bonnes connexions retrouvées (vrais positifs), les mauvaises connexions produites (faux positifs) et les connexions manquées (faux négatifs).

Pour évaluer la reconstruction de faisceaux, on s'intéresse plutôt au volume reconstruit par les algorithmes de tractographie. Pour ce faire, on doit préliminairement discréteriser le tractogramme en masque d'incidence binaire dans un volume 3D, comme représenté dans la [figure 1.9](#). Ainsi, à l'aide d'un volume de référence pour un faisceau donné (la vérité terrain), on peut mesurer des métriques d'évaluation comme le coefficient de Dice [39] ([équation 1.9](#)), le chevauchement [82] (OL : *overlap*) ([équation 1.10](#)) et le dépassement [82] (OR : *overreach*) ([équation 1.11](#)). La [figure 1.10](#) montre une schématisation de la conversion en masque d'incidence binaire, ainsi que du calcul des métriques d'évaluation, avec les équations suivantes :

$$\text{DICE} = \frac{2|A \cap B|}{|A| + |B|}, \quad (1.9)$$

$$\text{OL} = \frac{|A \cap B|}{|A|}, \quad (1.10)$$

$$\text{OR} = \frac{|B \setminus A|}{|A|}. \quad (1.11)$$

Alors que le coefficient de Dice est symétrique, le chevauchement et le dépassement ne le sont pas, puisqu'ils assument la présence d'une «vérité terrain» à laquelle on compare un «candidat». Le chevauchement correspond à l'intersection de A et B, divisé par le volume de A, c.-à-d. le pourcentage du volume de A couvert par B : il prend une valeur entre zéro et un. Le dépassement correspond au volume de B en excluant A, divisé par le volume de A : il prend une valeur allant de zéro à l'infini.

On observe donc qu'il n'y a pas une seule valeur permettant de classer ou d'évaluer de façon absolue les différents algorithmes de tractographie puisqu'il y a plusieurs facteurs à considérer. On verra plus loin que la tractographie par apprentissage s'intéresse davantage à la reconstruction de faisceaux, et plus spécifiquement au volume des

1.2. ÉVALUATION DES ALGORITHMES DE TRACTOGRAPHIE

faisceaux reconstruits. On travaillera donc principalement avec la discréétisation en volume 3D des tractogrammes plutôt qu'avec les tractes directement, et les algorithmes seront comparés sur la base du chevauchement, du dépassement, et du coefficient de Dice. Ces notions sont importantes, notamment pour le [chapitre 2](#) et le [chapitre 4](#), qui utilisent ces métriques comme outil de comparaison.

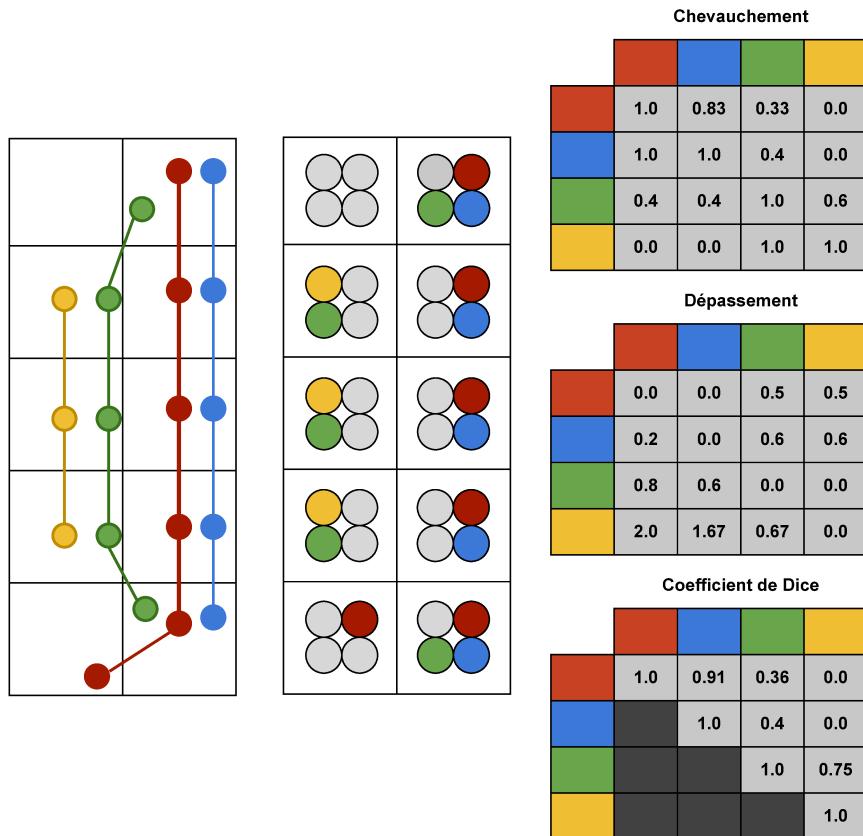


Figure 1.10 – Schématisation des métriques de chevauchement et de dépassement, ainsi que du coefficient de Dice. À gauche, 4 tractes (en jaune, vert, rouge et bleu) dans un volume de 5 x 2 voxels. Au centre, la discréétisation des tractes indiquant la présence ou l'absence d'une tracte pour chaque voxel. À droite, le calcul des métriques pour chaque paire de tractes. Comme le coefficient de Dice est symétrique, les entrées répétées sont noircies. Figure modifiée, courtoisie de François Rheault.

1.3. RÉSEAUX DE NEURONES ET SÉQUENCES DE DONNÉES

1.3 Réseaux de neurones et séquences de données

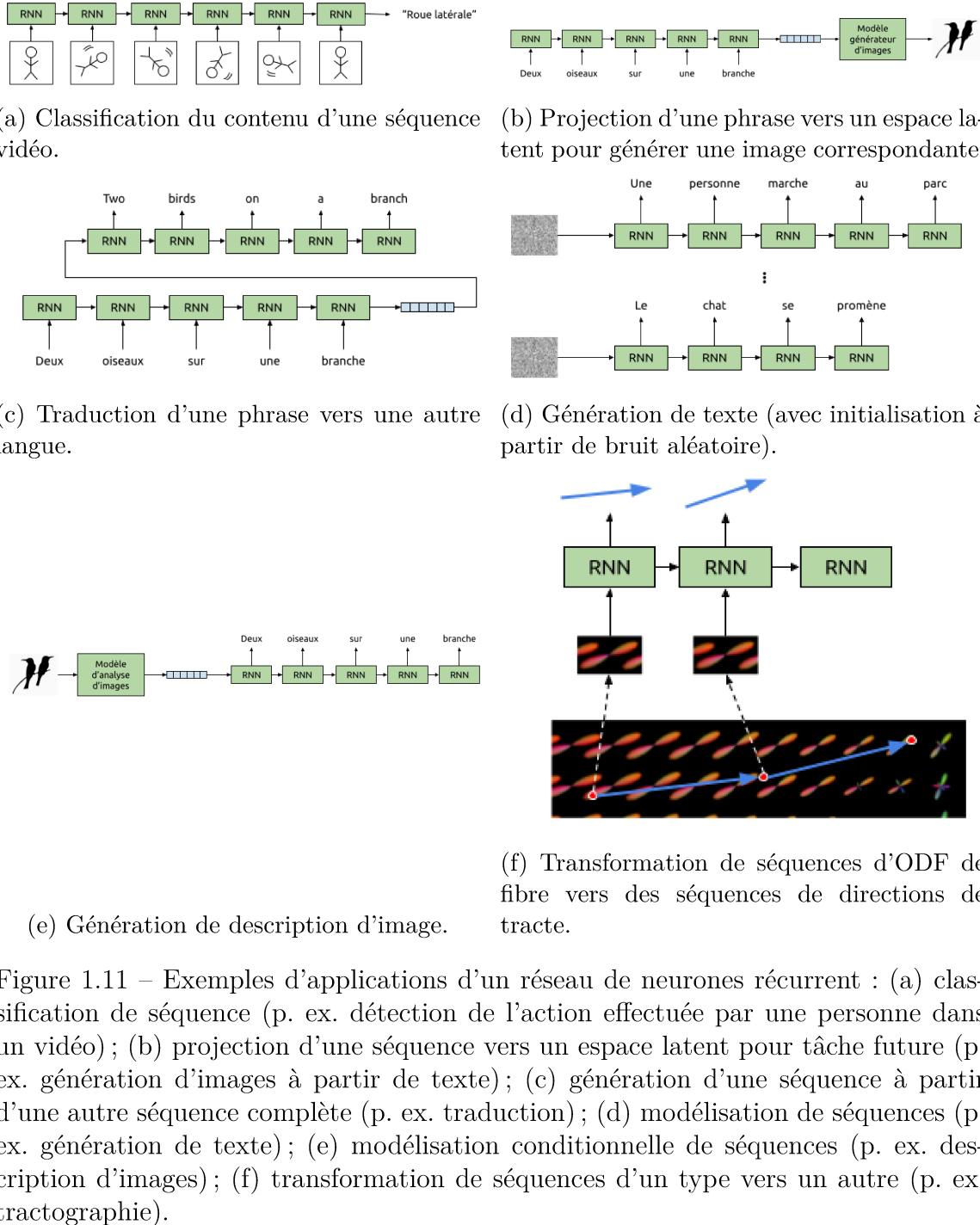
Le premier algorithme d'apprentissage machine proposé pour la tractographie était basé sur une forêt d'arbres décisionnels [93]. Depuis, plusieurs autres méthodes ont été proposées, comme le réseau récurrent à prédiction déterministe [101] ou probabiliste [11], le perceptron multicouche [160, 159], les filtres harmoniques hiérarchiques [112], et même l'apprentissage par renforcement Théberge et al.. Cette thèse privilégie l'approche par réseaux de neurones récurrents.

1.3.1 Réseaux de neurones récurrents

La tractographie repose sur la génération d'une séquence de directions à prendre, basées sur l'information de diffusion sous-jacente. Cependant, le chemin parcouru par une tracte peut influencer la prochaine direction à prendre afin de résoudre certains cas problématiques, comme un chevauchement de deux faisceaux de fibres (dans ce cas, même une meilleure résolution ne peut résoudre cette incertitude). Ainsi, le réseau de neurones récurrent (RNN : *Recurrent Neural Network*) est tout indiqué, puisqu'il permet d'emmagasiner de l'information afin de traiter des séquences de données. Dans sa définition originale, le RNN correspond à un réseau de neurones artificiels à une couche, mais la sortie du modèle à l'étape t de la séquence est concaténée à l'entrée de l'étape $t + 1$, puis le même modèle (donc les mêmes poids) est réutilisé récursivement pour traiter une séquence complète. La figure 1.11 illustre plusieurs applications du réseau de neurones récurrent.

Pour la tractographie, on s'intéresse au problème de transformation de séquences. Spécifiquement, on doit transformer une séquence d'informations de diffusion en séquence de directions. Comme les algorithmes de tractographie classiques, on désire prendre l'information de diffusion locale pour diriger le prochain pas de la tracte, mais tout en considérant le chemin déjà parcouru afin d'éviter des erreurs courantes. Le chapitre 2 présente la première tentative d'utiliser un réseau de neurones récurrent pour la tractographie par apprentissage. Le chapitre 4 présente quant à lui une amélioration du modèle, où on modifie également la couche de sortie et la fonction de coût afin de modéliser différentes distributions de probabilité pour prédire la prochaine direction de la tracte.

1.3. RÉSEAUX DE NEURONES ET SÉQUENCES DE DONNÉES



1.3. RÉSEAUX DE NEURONES ET SÉQUENCES DE DONNÉES

À chaque pas de la séquence, le RNN prend un vecteur d'information en entrée x_t et l'état caché du pas précédent h_{t-1} , pour calculer le nouvel état caché h_t et la sortie du modèle \hat{y}_t :

$$h_t = g(W_x x_t + W_h h_{t-1} + b), \quad (1.12)$$

$$\hat{y}_t = h_t, \quad (1.13)$$

où W_x , W_h , et b sont des poids partagés dans le temps, et g est une fonction d'activation (souvent une fonction sigmoïde ou tangente hyperbolique). Aussi, l'état caché initial h_0 est un vecteur nul. Notons que pour simplifier les équations subséquentes, la sortie \hat{y}_t et l'état caché h_t sont identiques ici. Typiquement, la sortie \hat{y}_t est calculée à partir de l'état caché via une couche de réseau de neurones à propagation avant :

$$\hat{y}_t = f(W_y h_t + b_y). \quad (1.14)$$

Une fonction de coût permet de calculer l'erreur totale d'une séquence en faisant la somme des erreurs à chaque pas :

$$\mathcal{L}(\hat{y}, y) = \sum_{t=1}^T \mathcal{L}(\hat{y}_t, y_t), \quad (1.15)$$

où \hat{y} est la séquence des éléments \hat{y}_t prédits par le modèle, et y est la séquence des éléments à prédire y_t , toutes deux d'une taille T .

Pour entraîner un réseau récurrent avec une méthode de descente de gradient, comme pour un réseau de neurones à propagation avant, on doit calculer la dérivée de la fonction de coût par rapport aux poids. Cependant, pour une prédition donnée, on doit propager l'erreur à travers les applications récursives du modèle. La technique appelée rétropropagation à travers le temps (BPTT : *Backpropagation Through Time*) consiste à dérouler le modèle en dérivées partielles, puis reculer à travers les étapes récursives jusqu'au pas initial. Par simplicité, on dénote comme W la concaténation des matrices de poids $[W_x, W_h]$ (on ignore le paramètre de biais b pour la suite), et comme E l'erreur totale $\mathcal{L}(\hat{y}, y)$. Le gradient de l'erreur totale est donc

$$\frac{\partial E}{\partial W} = \sum_{t=1}^T \frac{\partial E_t}{\partial W}. \quad (1.16)$$

1.3. RÉSEAUX DE NEURONES ET SÉQUENCES DE DONNÉES

Les poids sont mis à jour en utilisant une descente de gradient avec un taux d'apprentissage α :

$$W \leftarrow W - \alpha \frac{\partial E}{\partial W}. \quad (1.17)$$

Le gradient de l'erreur d'une prédiction par rapport à W à l'étape t (sans considérer les étapes récursives) est

$$\frac{\partial E_k}{\partial W} = \frac{\partial E_k}{\partial h_k} \cdot \frac{\partial h_k}{\partial W}. \quad (1.18)$$

Or, l'état caché h_k dépend aussi des états cachés précédents, qui eux dépendent aussi de W . En tenant compte de toutes les étapes récursives du modèle, on obtient le gradient

$$\frac{\partial E_k}{\partial W} = \frac{\partial E_k}{\partial h_k} \cdot \left[\sum_{i=1}^k \left(\frac{\partial h_k}{\partial h_i} \cdot \frac{\partial h_i}{\partial W} \right) \right]. \quad (1.19)$$

$$\frac{\partial h_k}{\partial h_i} = \prod_{t=i+1}^k \frac{\partial h_t}{\partial h_{t-1}} \quad (1.20)$$

Sachant que $W = [W_x, W_h]$, en utilisant la définition de h_t donnée par l'équation 1.12, la dérivée de h_t s'écrit

$$\frac{\partial h_t}{\partial h_{t-1}} = g'(W_x x_t + W_h h_{t-1}) \cdot \frac{\partial}{\partial h_{t-1}} [W_x x_t + W_h h_{t-1}] \quad (1.21)$$

$$= g'(W_x x_t + W_h h_{t-1}) W_h. \quad (1.22)$$

En combinant l'équation 1.22 avec l'équation 1.19 et l'équation 1.20, on obtient

$$\frac{\partial E_k}{\partial W} = \frac{\partial E_k}{\partial h_k} \cdot \left[\sum_{i=1}^k \left(\prod_{t=i+1}^k g'(W_x x_t + W_h h_{t-1}) W_h \cdot \frac{\partial h_i}{\partial W} \right) \right]. \quad (1.23)$$

Si la dérivée de la fonction d'activation g est plus petite que un (c'est le cas pour la sigmoïde et la tangente hyperbolique), le produit tend vers zéro, et les poids ne seront pas mis à jour. Ainsi, le réseau de neurones récurrent, dans sa définition originale, souffre du problème de «disparition de gradient» [61], et peut difficilement se «souvenir» d'information au-delà d'un certain nombre d'étapes.

1.3. RÉSEAUX DE NEURONES ET SÉQUENCES DE DONNÉES

1.3.2 Long Short-Term Memory Network

Le modèle récurrent à longue mémoire courte (LSTM : *Long Short-Term Memory Network*) — illustré dans la figure 1.12 — a été proposé comme solution au problème de disparition de gradient [60, 61]. Ce modèle récurrent permet en quelque sorte de contrôler l’information stockée dans le modèle et de l’utiliser pour faire de futures prédictions sans souffrir du problème de disparition de gradient. Alors que le RNN original possède une seule couche de réseau de neurones, le LSTM en possède trois supplémentaires qui agissent en tant que «portes» pour contrôler le flot d’information dans le modèle. Chacune de ces portes est composée d’un vecteur prédit par une couche de neurones artificiels suivis d’une fonction sigmoïde pour contenir uniquement des valeurs entre zéro et un, qui bloquent ou laissent passer de l’information via une multiplication élément par élément (produit de Hadamard). De plus, le LSTM possède également une cellule (*cell state*) dans laquelle il peut stocker et retirer de l’information.

La figure 1.12 illustre l’architecture du LSTM, défini par les équations suivantes :

$$f_t = \sigma(W_f \cdot [x_t, h_{t-1}] + b_f), \quad (\text{porte d'oubli}) \quad (1.24)$$

$$i_t = \sigma(W_i \cdot [x_t, h_{t-1}] + b_i), \quad (\text{porte d'entrée}) \quad (1.25)$$

$$o_t = \sigma(W_o \cdot [x_t, h_{t-1}] + b_o), \quad (\text{porte de sortie}) \quad (1.26)$$

$$\tilde{c}_t = \tanh(W_c \cdot [x_t, h_{t-1}] + b_c), \quad (\text{cellule candidate}) \quad (1.27)$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t, \quad (\text{cellule}) \quad (1.28)$$

$$h_t = o_t * \tanh(c_t), \quad (\text{état caché}) \quad (1.29)$$

$$\hat{y}_t = h_t, \quad (\text{sortie}) \quad (1.30)$$

où $*$ désigne une multiplication élément par élément. La porte d’oubli contrôle l’information conservée de la cellule passée vers la nouvelle cellule ; la porte d’entrée contrôle l’information transmise de la cellule candidate à la nouvelle cellule ; la porte de sortie contrôle l’information transmise de la cellule vers le nouvel état caché. Le modèle récurrent à information contrôlée (GRU : *Gated Recurrent Unit*) est une autre architecture récurrente proposée pour pallier aux problèmes du RNN original, similaire au

2. <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>

1.4. ÉLÉMENTS TECHNIQUES LIÉS À LA TRACTOGRAPHIE PAR APPRENTISSAGE AVEC RÉSEAUX DE NEURONES RÉCURRENTS

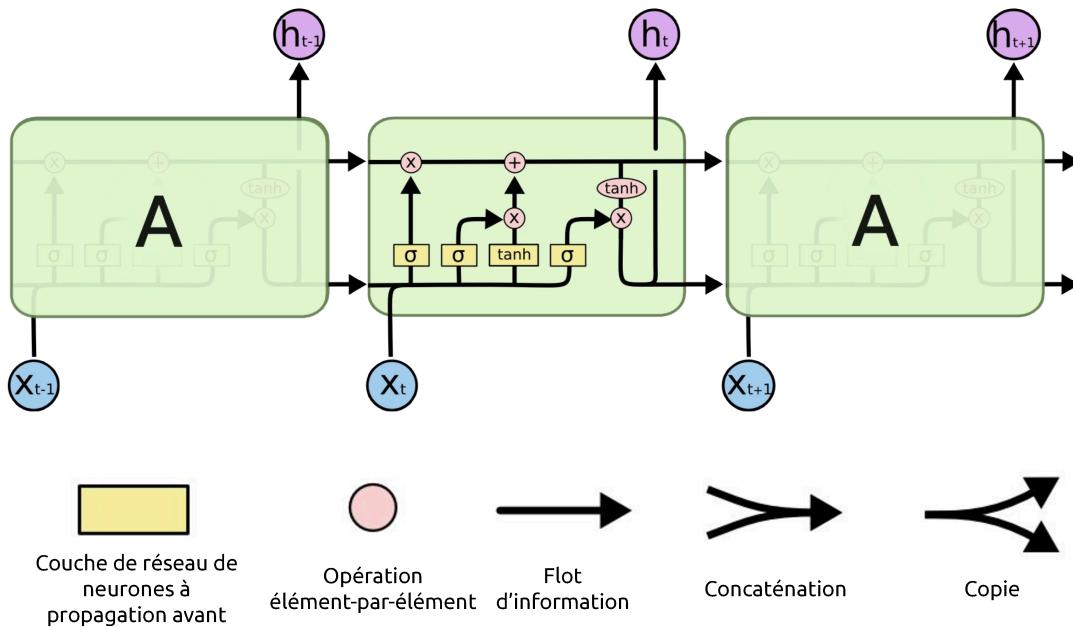


Figure 1.12 – Schéma du LSTM. Chaque ligne indique le transfert de vecteur d'une opération à une autre, c.-à-d. le flot d'information. Chaque couche de réseau de neurones à propagation avant possède une fonction d'activation, notée par σ pour une sigmoïde ou \tanh pour une fonction tangente hyperbolique. La cellule c_t (non identifiée dans le schéma) correspond au flot d'information continu dans la région supérieure du module A. Figure adaptée du blogue de [Christopher Olah](#)².

LSTM, et sera mentionné dans le [chapitre 2](#).

1.4 Éléments techniques liés à la tractographie par apprentissage avec réseaux de neurones récurrents

L'entraînement d'un modèle récurrent pour reconstruire des tractes de matière blanche a ses particularités bien précises uniques à ce domaine. Dans le cadre de ce projet, diverses méthodes de transformation et d'augmentation des données ont été développées afin de permettre l'utilisation de ce type de modèle pour la tractographie, spécifiquement dans le cadre d'une large base de données comme proposée dans le

1.4. ÉLÉMENTS TECHNIQUES LIÉS À LA TRACTOGRAPHIE PAR APPRENTISSAGE AVEC RÉSEAUX DE NEURONES RÉCURRENTS

[chapitre 4](#). Comme certains de ces détails ne se retrouvent dans aucun article publié, ils se retrouvent ici pour éclairer le lecteur sur le défi technologique que représente la tractographie par apprentissage avec les réseaux récurrents.

1.4.1 Signal d'entrée du modèle

Pour la tractographie par apprentissage, tel que décrit dans le [chapitre 2](#), l'entrée du réseau est une séquence de signaux de diffusion, correspondant aux points d'une tracte de référence. Ce signal d'entrée peut prendre plusieurs formes : le signal brut discret, les coefficients SH représentant le signal, les coefficients SH représentant l'ODF de fibre, ou même les pics de l'ODF de fibre. À titre de référence, la méthode décrite dans le [chapitre 2](#) utilise le signal de diffusion, mais projeté sur 100 directions réparties uniformément sur la sphère, alors que le [chapitre 4](#) utilise plutôt les coefficients SH représentant le signal.

1.4.2 Signal de sortie du modèle

La sortie du modèle dépend du type d'information qui doit être prédite. Dans le cas simple où on prédit une direction (un vecteur 3D), le modèle calcule trois scalaires en sortie, sans appliquer de fonction d'activation. Ces trois scalaires sont ensuite interprétés comme un vecteur en trois dimensions.

Pour prédire une distribution de probabilité, comme proposé dans le [chapitre 4](#), quelques modifications sont apportées. D'abord, le nombre de neurones en sortie est ajusté pour obtenir le bon nombre de paramètres. Ensuite, certains paramètres ont besoin de transformations spécifiques. Par exemple, pour obtenir un paramètre de variance (strictement positif) pour une gaussienne, on applique une exponentielle à la valeur de sortie. Ou encore, pour obtenir les paramètres de mélange d'une mixture de Gaussiennes, on applique une fonction exponentielle normalisée (*softmax*) pour que la somme des valeurs de sortie soit un (1) et qu'elles représentent une distribution de probabilité.

1.4. ÉLÉMENTS TECHNIQUES LIÉS À LA TRACTOGRAPHIE PAR APPRENTISSAGE AVEC RÉSEAUX DE NEURONES RÉCURRENTS

1.4.3 Bruitage des tractes

Les tractes de référence utilisées pour entraîner les modèles récurrents ne sont pas parfaites. En pratique, elles sont globalement valides, c'est-à-dire qu'elles connectent les bonnes régions entre elles, mais peuvent être localement bruitées. En effet, les algorithmes ne produisent généralement pas des tractes lisses, en particulier les algorithmes probabilistes qui ont une tendance intrinsèque à changer régulièrement de direction. Ainsi, pour éviter d'apprendre le bruit et les biais spécifiques des algorithmes de référence, on peut ajouter explicitement un léger bruit gaussien à chaque coordonnée (avec une attention particulière pour ne pas inverser deux pas subséquents).

1.4.4 Cisaillement des tractes

L'initialisation du processus de tractographie peut se faire n'importe où dans la matière blanche, alors que les tractes de référence sont complètes, c'est-à-dire qu'elles débutent et terminent près de la matière grise plutôt que dans la matière blanche.

Pour un modèle local, ce type d'initialisation ne pose aucun problème, mais devient problématique pour un modèle récurrent qui n'a jamais été exposé à ce type de données en entraînement. Comme l'initialisation dans la matière blanche a plusieurs avantages, tels qu'une meilleure couverture spatiale et une réduction du biais de longueur [113], il serait judicieux de proposer une méthode qui répond à ce problème.

Une solution simple est d'augmenter les données de chaque lot d'entraînement par un cisaillement des tractes de référence à un endroit aléatoire de chacune, de sorte que la position initiale de celles-ci ne soit plus uniquement près de la matière grise. Par contre, on ne doit pas appliquer cette méthode sur toutes les tractes, on doit plutôt choisir une certaine probabilité de cisaillement, qui doit représenter la distribution des données de test. Ainsi, la probabilité est fixée à 50%, puisqu'au moment d'évaluer un modèle, celui-ci est initialisé dans la matière blanche une fois sur deux afin de reconstruire une demi-tracte, puis celle-ci est lue en sens inverse afin de la compléter et de reconstruire une tracte complète.

1.4. ÉLÉMENTS TECHNIQUES LIÉS À LA TRACTOGRAPHIE PAR APPRENTISSAGE AVEC RÉSEAUX DE NEURONES RÉCURRENTS

1.4.5 Ajout du voisinage en entrée

Pour améliorer les prédictions du modèle, l'information la plus pertinente serait de voir le signal d'entrée plus loin que la position courante avant de prendre une décision. Pour ce faire, on ajoute le signal dans un voisinage autour de la coordonnée courante, à une distance fixe dans les six axes principaux (la distance correspond à la distance du pas effectué entre chaque prédition). En conséquence, la dimensionnalité du vecteur d'entrée est multipliée par sept (la position courante plus les six voisins).

1.4.6 Échantillonnage semi-aléatoire des tractes

Étant donné les contraintes de mémoire graphique (GPU) et mémoire vive (RAM), il est impossible de charger tous les volumes de diffusion en même temps. En effet, une base de données comme *TractoInferno* a un taille totale d'environ 350 Go, et les futures bases de données tendent à être de plus en plus lourdes en mémoire avec la quantité grandissante du nombre d'acquisitions et l'amélioration de la résolution des images. Ainsi, un échantillonnage aléatoire des tractes à travers tous les volumes est une mauvaise stratégie. Pour être plus judicieux avec l'utilisation de la mémoire, un échantillonnage semi-aléatoire est proposé, où un nombre maximal de volumes de diffusion par lot d'entraînement est spécifié, puis les tractes sont échantillonées aléatoirement parmi les volumes de ce sous-ensemble. De plus, pour éviter de charger de nouveau volumes en mémoire à chaque lot, un nombre de cycles par sous-ensemble de volumes est aussi spécifié. Finalement, pour un traitement efficace des données, le travail d'interpolation du signal est fait en parallèle sur CPU, puis les données traitées sont envoyées sur le GPU un lot à la fois pour entraîner le modèle. La [figure 1.13](#) présente le schéma du processus complet de traitement parallèle des données.

1.4. ÉLÉMENTS TECHNIQUES LIÉS À LA TRACTOGRAPHIE PAR APPRENTISSAGE AVEC RÉSEAUX DE NEURONES RÉCURRENTS

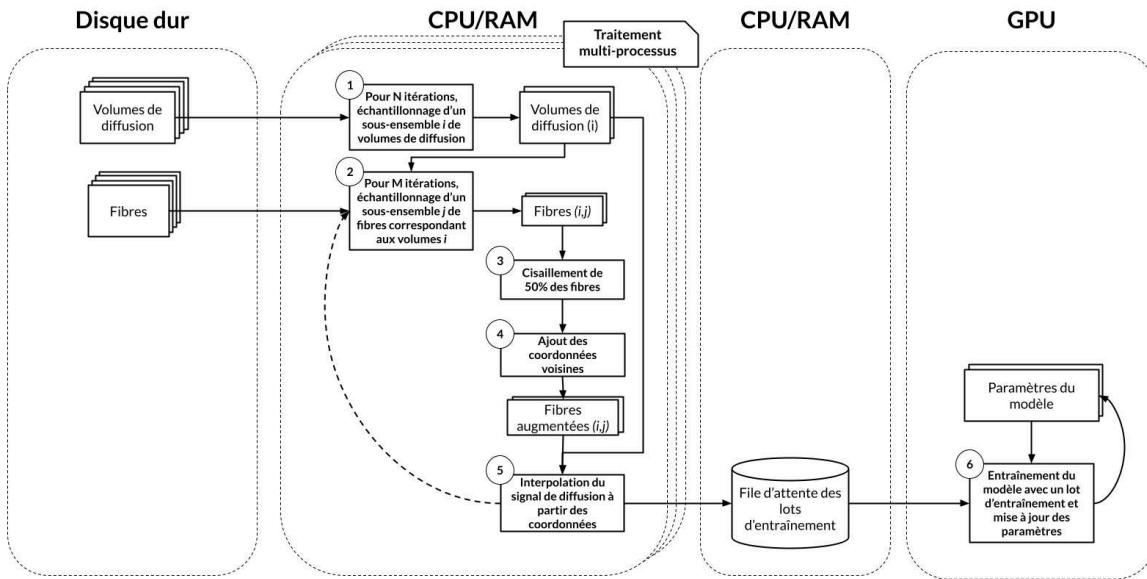


Figure 1.13 – Traitement et augmentation des données pour l’entraînement d’un modèle de tractographie.

1 : Pour N itérations, un ensemble de volumes de diffusion est échantillonné, qui sera utilisé pendant M itérations afin de réduire les transferts de données en mémoire vive (RAM).

2 : Pour M itérations, un ensemble de tractes est échantillonné à partir des volumes échantillonnés en 1.

3, 4 : Les tractes sont cisaillées et augmentées des coordonnées voisines.

5 : À partir des coordonnées des tractes augmentées, le signal d’entrée est interpolé pour former un lot de données d’entraînement, puis envoyé dans une file d’attente. Dès qu’une place se libère dans la file d’attente, retourner à l’étape 2 (ou à l’étape 1 si les M itérations sont terminées).

6 : Tant qu’il y a des données, un lot de données est envoyé de la file d’attente vers le processeur graphique (GPU) pour une itération d’entraînement du modèle.

Chapitre 2

Réseaux de neurones récurrents pour la tractographie par apprentissage

Résumé

Cet article propose le premier modèle de tractographie par apprentissage qui utilise un réseau de neurones récurrent, qui peut intégrer l'information du chemin complet d'une tracte afin de faire une prédiction. Un volume de diffusion synthétique (simulé) est utilisé pour entraîner le modèle et évaluer ses prédictions, à l'aide de l'outil standardisé *Tractometer*. Un tractogramme non-filtré est utilisé comme jeu de données d'entraînement.

Contributions de la publication

- Présentation du premier algorithme de tractographie par apprentissage qui utilise un réseau de neurones récurrent pour intégrer le contexte spatial de la tracte.
- Comparaison de plusieurs versions de l'algorithme proposé.

- Amélioration relative de la couverture spatiale de 108% comparativement à la soumission moyenne, et de 35% comparativement à la méthode utilisée pour générer les données d’entraînement.
- Seul algorithme obtenant une couverture spatiale moyenne de plus de 50% tout en conservant le débordement spatial sous 50%, comparé à toutes les soumissions au *ISMRM 2015 Tractography Challenge*.
- Évaluation quantitative de l’algorithme sur un sujet synthétique, et production de résultats qualitatifs sur un sujet réel de la base de données *Human Connectome Project*.

Contributions des auteurs

- Écriture du manuscrit et analyse des résultats (Philippe Poulin)
- Programmation de l’algorithme (Marc-Alexandre Côté et Philippe Poulin)
- Développement de l’outil d’évaluation Tractometer (Marc-Alexandre Côté et Jean-Christophe Houde)
- Création de la figure [figure 2.3](#) (Marc-Alexandre Côté)
- Dissection virtuelle des faisceaux (Laurent Petit)
- Génération des tractes d’entraînement et développement du sujet synthétique du *ISMRM 2015 Tractography Challenge* (Peter F. Neher et Klaus H. Maier-Hein)
- Supervision du projet (Hugo Larochelle et Maxime Descoteaux)

Commentaires

Cet article a été publié à MICCAI (*Medical Image Computing & Computer Assisted Intervention*) en 2017. Cette conférence est très sélective, et a généralement un taux d’acceptation sous les 30%. Il s’agit de la meilleure conférence dans le domaine de l’analyse d’images médicale.

Learn to Track: Deep Learning for Tractography

Philippe Poulin¹, Marc-Alexandre Côté², Jean-Christophe Houde²,
Laurent Petit⁴, Peter F. Neher³, Klaus H. Maier-Hein³, Hugo
Larochelle¹, Maxime Descoteaux²

¹ Computer Science Department, Université de Sherbrooke, Sherbrooke, Canada

² Sherbrooke Connectivity Imaging Laboratory (SCIL), Computer Science
Department, Université de Sherbrooke, Sherbrooke, Canada

³ Medical Image Computing (MIC), German Cancer Research Center (DKFZ),
Heidelberg, Germany

⁴ Groupe d’Imagerie Neurofonctionnelle, IMN, CNRS, CEA, Université de Bordeaux,
France

Abstract

We show that deep learning techniques can be applied successfully to fiber tractography. Specifically, we use feed-forward and recurrent neural networks to learn the generation process of streamlines directly from diffusion-weighted imaging (DWI) data. Furthermore, we empirically study the behavior of the proposed models on a realistic white matter phantom with known ground truth. We show that their performance is competitive to that of commonly used techniques, even when the models are used on DWI data unseen at training time. We also show that our models are able to recover high spatial coverage of the ground truth white matter pathways while better controlling the number of false connections. In fact, our experiments suggest that exploiting past information within a streamline’s trajectory during tracking helps predict the following direction.

2.1. INTRODUCTION

2.1 Introduction

Tractography is currently at the heart of human brain connectomics studies [148]. However, recent biases and limitations of existing tractography pipelines have been highlighted [27], such as the reconstruction of many non-existent connections (false positive streamlines), poor spatial extent of existing connections and the difficulty of injecting anatomical priors beyond manual dissection and tissue classes from T1-weighted segmentations.

Currently, tracking algorithms depend on local models with assumptions on the nature of the underlying DWI signal. In 2015, [92] proposed a machine learning approach to fiber tractography based on a random-forest classifier. They successfully demonstrated how a purely data-driven approach can be used to reconstruct streamlines from the raw diffusion signal. Their method works well on 2D synthetic data and shows promising qualitative results on *in vivo* data. However, it has yet to be shown how well machine learning (and particularly deep learning) approaches can perform quantitatively on more realistic data and how well they can generalize to unseen data. In this paper, our main contributions are the first deep learning models for this problem and their evaluation, namely (1) a local reconstruction model based on a multilayer perceptron, (2) a sequential reconstruction model based on a recurrent neural network, (3) a careful quantitative evaluation of performances on the phantom of the ISMRM 2015 Tractography Challenge, and (4) a qualitative examination of the streamlines generated in unseen data during training. Our method outperforms or is competitive with the current state-of-the-art deterministic and probabilistic tractography algorithms robust to crossing fibers. In particular, out of 96 other tractography methods, this is the only approach able to recover more than 50% of spatial coverage of ground truth bundles while producing overreaching false connections below 50%. Our recurrent neural network is a promising deep learning solution for tractography based on raw DWI. It includes a notion of history of followed directions, which makes it robust to crossing fibers, robust to a wide range of geometries and allows the flexibility to include priors and learn how to reduce false-positive connections.

2.2. USING DEEP LEARNING FOR TRACTOGRAPHY

2.2 Using Deep Learning for Tractography

Given a diffusion dataset and sequences of spatial coordinates, the goal is to train a model to predict tracking directions to follow. In the context of tractography, a deterministic model can be used in an iterative process for streamline creation.

We chose to focus on deep learning models because of their well-known ability to discover and extract meaningful structures directly from raw data [76]. Our models are based on two types of deep learning models : a *Feed-Forward Neural Network* (FFNN), and a *Recurrent Neural Network* (RNN) [56]. While the FFNN is a local model and serves as a good baseline, it has the same weaknesses as existing methods, i.e. it is not able to learn streamline structures. To address this weakness, we used an RNN, because this family of models can process whole sequences as input. In our case, treating streamlines as sequences of coordinates in 3D space, our hypothesis is that a recurrent model should be able to learn the fiber or bundle structure through the diffusion signal in order to make better predictions and solve classic problems like fiber crossing.

Model inputs: As in [92], to be independent of the gradient scheme, the raw diffusion signal is first resampled to have D gradient encodings evenly distributed on the sphere (we used $D = 100$). We also normalized each diffusion-weighted images by the b=0 image. A streamline is represented as a sequence \mathbf{S} of M equally-spaced spatial coordinates $P_i = (x_i, y_i, z_i)$. The diffusion signal is evaluated at each of these points, using trilinear interpolation in the voxel space. This results in a sequence of M vectors with D dimensions representing the diffusion information along the streamline. In all our models, we also tried giving the previous direction as a supplementary input, as in [92]. Note that the spatial coordinates are not given as input to the model. This choice allows the model to be invariant to brain size or translation, reducing the preprocessing needed before feeding data to the model and improving generalization.

2.2.1 Models

FFNN The FFNN sees all streamline coordinates as individual, independent local data points. The output of the model is a 3-dimensional normalized vector. The

2.2. USING DEEP LEARNING FOR TRACTOGRAPHY

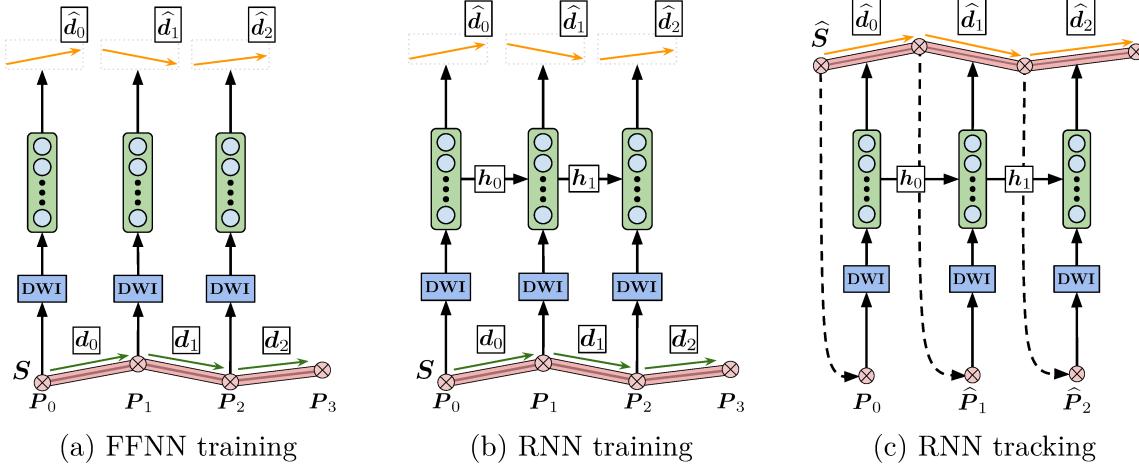


Figure 2.1 – Architecture of the proposed models. (a) Given a streamline \mathbf{S} , diffusion information is evaluated at each point P_i using trilinear interpolation ($\text{DWI}(P_i)$). The resulting vector is provided to the FFNN to predict a direction $\hat{\mathbf{d}}_i$ (orange), which is compared against its associated target direction \mathbf{d}_i (green). (b) Unlike the FFNN, the RNN has recurrence connexions through each step, allowing to send information to itself through the sequence. (c) Given a starting point P_0 , a new streamline $\tilde{\mathbf{S}}$ is generated by iteratively predicting a new direction $\hat{\mathbf{d}}_i$, and feeding the estimated new position \hat{P}_{i+1} back to the model. Note how the predicted direction $\hat{\mathbf{d}}_i$ gets influenced by prior information along the streamlines through $\mathbf{h}_{j < i}$.

model is represented in Figure 2.1a. To remove the directional ambiguity when no previous direction is given, we choose to consider the output vector as an *undirected axis* instead of a direction. To this end, the loss function is defined as the negative squared cosine similarity.

RNN The general idea behind the RNN is to model an internal state that is updated with each new observation in the input sequence and can be used to make predictions. Through its updatable internal state, the model can “remember” relevant features about the past. In this case, we used a Gated Recurrent Unit (GRU) [22] type of RNN.

Figure 2.1b shows that for each point P_i in the streamline, the diffusion information $\mathbf{DWI}(P_i)$ is used to update the internal state \mathbf{h}_i of the model. From there, at each step along the streamline, the model makes a prediction of the direction to

2.3. RELATED WORK

follow $\hat{\mathbf{d}}_i$. The loss function is defined as the mean squared error (MSE) between the model’s prediction $\hat{\mathbf{d}}_i$ and the target \mathbf{d}_i (i.e. the next normalized segment of the streamline).

2.2.2 Tractography

Tractography is performed by using a fully trained model. Streamlines generation follows an iterative process as in classical streamline-based tractography techniques [89] as illustrated in Figure 2.1c. From a seed point $P_0 = (x_0, y_0, z_0)$, a new streamline is created with the initial seed $\hat{\mathcal{S}} = \{P_0\}$. Next, the model is given the DWI data at the previous streamline coordinate P_i to obtain a predicted direction $\hat{\mathbf{d}}_i$. A next point $P_{i+1} = P_i + \alpha \hat{\mathbf{d}}_i$ is then computed, where α is a chosen step size, as in standard streamline-based tractography algorithms. Points are generated to iteratively produce a streamline until a desired criterion is met (e.g. too high curvature, exiting WM mask). The whole process is repeated as many times as required to produce a full tractogram.

2.3 Related Work

The work of Neher *et al.* [92] hypothesizes that tractography can be improved by considering local neighborhood features, adding a directional prior to promote straight fibers, and using a fiber deflection protocol to help the model recover from mistakes. More precisely, their model makes predictions based on a voting mechanism, using local direction proposals from multiple sample positions in the vicinity of the current location. Each direction proposal is obtained by a classification over 100 possible directions (weighted using the previous direction), along with a streamline termination probability. If fiber termination is the more likely option, a deflection is attempted by rotating the sample point 180° around the previous direction and classifying a second time.

In our current approach, the problem is framed as a regression task over normalized directions instead of a vote over discretized directions. This means that to produce a prediction, fewer computations are needed at the output, compared to computing and

2.4. EXPERIMENTS

voting over many proposals. Regression also allows the model to output more precise directions and thus be more suitable at exploiting smaller variations in direction. In addition, if straight fibers are supported by the data, a directional prior should not be necessary and a deep learning model should be able to learn the right structure, which is why our model does not include such a prior.

While Neher *et al.* consider the neighborhood of the current position, they do not consider the full evolution of the streamline up to each point. Our hypothesis is that there are high-order dependencies between the next direction in the streamline and all previous directions. Consequently, our recurrent approaches have a natural mechanism for integrating past information along the streamline to predict a next direction. These two approaches are not exclusive however, and would probably benefit from each other.

Finally, in a deep learning context, learning a stopping criterion along with the direction to follow is more complex. It would require careful engineering and balancing of the loss function in order for one not to overcome the other during training, especially using a recurrent approach. This is beyond the scope of this paper and left for future work.

2.4 Experiments

We quantify the performance of our methods on the 2015 ISMRM challenge dataset [81] and evaluate using the Tractometer connectivity metrics [27]. In doing so, we can compare ourselves to the 96 original challenge submissions [26]. We then qualitatively evaluate our method when tracking on *in vivo* data.

Tracking parameters: Tracking was done using $1/2$ voxel step size (1.0 mm for ISMRM challenge, 0.625 mm for HCP). Seeding was done using 1 seed per voxel in the WM mask, and tracking was done using a dilated WM mask. All streamlines leaving the dilated mask were automatically terminated and streamlines shorter than 20 mm or longer than 200 mm were discarded. Streamlines with a half-cone curvature higher than 20° were also discarded.

2.4. EXPERIMENTS

2.4.1 ISMRM2015 Challenge

For the first experiment, we choose to reproduce the training environment of Nehler *et al.* [92], by using training data generated specifically for the subject of interest. They determined the optimal method for generating their training data to be a constrained spherical deconvolution (CSD) deterministic streamline tractography (DET) of MRtrix [141]. Staying coherent with their approach, a tractogram was generated using CSD-DET on a denoised and distortion-corrected version of the ISMRM2015 challenge dataset. The resulting 92K streamlines were then split into training and validation sets (using splits of 90% and 10%).

We trained models with one to four layers, varying the layer size between 500 and 1000, and used the Adam optimizer with early stopping. Full code is available [online](#)¹. For each type of model, the one with the best validation error was chosen for tracking & tractometer evaluation.

We report in Table 2.1 the valid connections ratio and the associated number of valid bundles (true positives), the invalid connections ratio and the associated number of invalid bundles (false negatives), the volumetric bundle overlap and overreach in percentages. Drawing conclusions from only one of these last two metrics can be misleading (e.g. a model can have both the best overlap and the worst overreach). These metrics are related to precision and recall measures, and are combined into the F_1 -measure. Note that the “_PD” model suffix indicates when the previous direction was given as input to the model. We report as baselines the ISMRM mean results and submission 6_1[26], which is a CSD-DET based method comparable to what was used to generate our training data.

We see that the local model (FFNN) is already competitive with the mean ISMRM challenge scores. Its ability to estimate the main diffusion axis and tuning its predictions according to the streamlines seen during training allows it to improve all mean scores except number of invalid connections (+3.9 VC, +7.6 IC, -10 NC, +1.6 VB, -214 IB, +12.2 OL, -1.6 OR, +11.6 F1). Surprisingly, adding the previous direction as input worsened the model’s performance. We think that the optimization process allowed the model to achieve a good performance by generally simply copying the

1. https://github.com/ppoulin91/learn2track/tree/miccai2017_submission

2.4. EXPERIMENTS

Table 2.1 – Quantitative evaluation on the ISMRM 2015 Tractography Challenge.

Model	Connections (%)			Bundles		Avg. Bundle (%)		
	Valid	Invalid	No	Valid	Invalid	Overlap	Overreach	F ₁
AVG RESULTS	53.6	19.7	25.2	21.4	281	31.0	23.0	44.2
TEAM 6_1	69.9	26.2	3.9	23	74	47.7	32.3	56.0
FFNN	57.5	27.3	15.2	23	67	43.2	21.4	55.8
FFNN_PD	14.8	64.2	21.1	22	100	36.7	40.7	45.3
RNN	66.1	25.3	8.6	21	36	7.7	12.0	14.2
RNN_PD	41.6	45.6	12.8	23	130	64.4	35.4	64.5

previous direction given as input. Indeed, we looked at the generated streamlines, and while they do a good job of covering the brain (the FFNN_PD model recovered 22 bundles out of 25), they are mostly straight and miss important connections.

Going from the local model to the recurrent model (RNN) provided different insights. Without the previous direction as input, the model generated more relative valid connections, but overall very few streamlines (as seen in the overlap metric, 7.7%). With the previous direction however, the model achieved very good coverage of the challenge bundles (64.4% overlap), while dropping a bit below 50% VC. It achieved the best F1 score over all our models. **In comparison, no submission in the ISMRM2015 challenge achieved an overlap higher than 50% while keeping overreach under 50%** [26, 81]. Figure 2.2 shows how the left CST is reconstructed with high coverage and low overreach. We believe that the recurrent model, being able to accumulate “memories” about the past of the streamline, is able to extract information of the previous direction without committing the same mistakes as the local model. This ability to “memorize” is what makes this model stand apart from classic methods.

2.4.2 *In vivo* tracking

Using the models trained in the first experiment (section 2.4.1), we tracked on an unseen brain (HCP subject #100307). As a gold standard we used a virtual ROI-based dissection made by an expert neuroanatomist [90, 19]. Streamlines used for the dissection were generated using Particle Filtering Tractography [53] using default

2.5. CONCLUSION

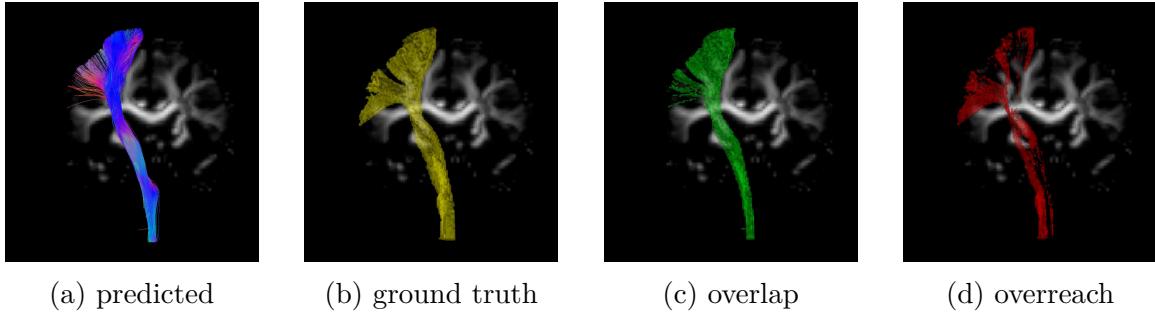


Figure 2.2 – (a) Left CST generated by the RNN model using the ISMRM2015 challenge data. (b) Ground truth mask as defined by the ISMRM2015 challenge. (c) Overlapping and (d) overreaching voxels of the generated bundle with respect to the ground truth mask.

parameters and based on a spherical harmonics 8 multi-shell constrained spherical deconvolution reconstruction [50, 64]. The resulting bundles are shown in Figure 2.3. Visual evaluation shows results that are in line with the first experiment. The local model does a good job of recovering the bundles, but has poor coverage. The recurrent model is much more similar to the expert segmentation in most of the recovered bundles. We suspect that the RNN would gain even more by training on much larger datasets with multiple subjects.

2.5 Conclusion

We propose the first deep learning alternatives to traditional local modeling approaches to tractography based on raw DWI. Our FFNN model provides the first performance baseline for local deep models. We also present a novel approach where the past of the streamline is considered by a recurrent model in order to make better predictions. Compared to the other ISMRM2015 submissions, this proved to be the only technique able to recover more than 50% of spatial coverage while producing overreaching false connections below 50%. We show that deep learning models can generalize to new DWI unseen at training time. These novel results show that deep learning is a promising approach to tractography.

While we believe that deep learning will be able to discover new pathways by

2.5. CONCLUSION

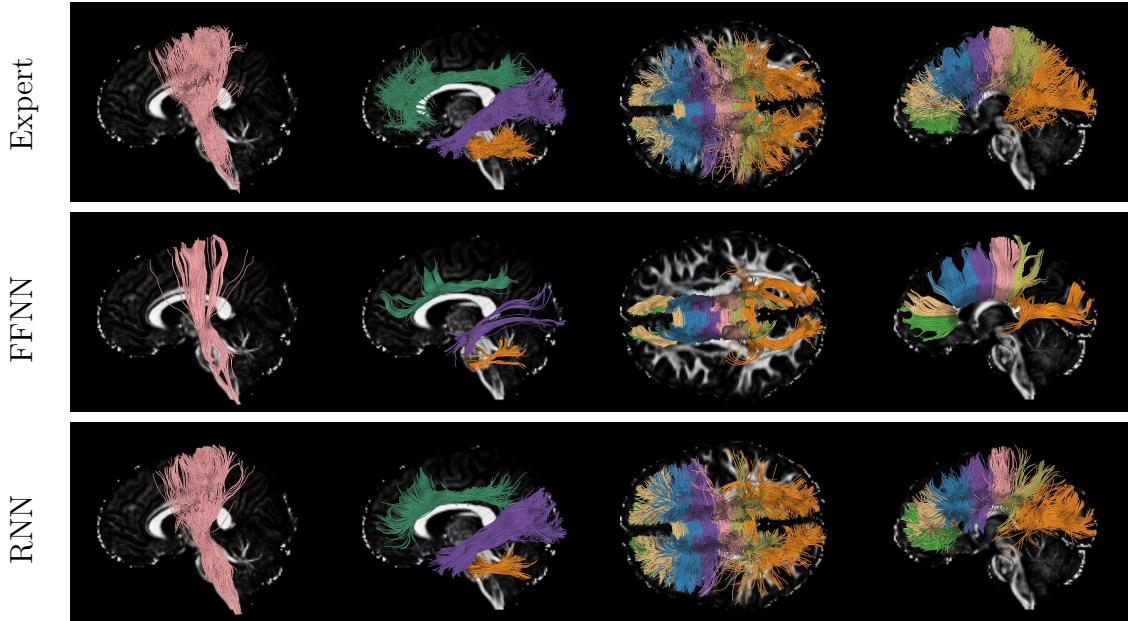


Figure 2.3 – Expert, FFNN and RNN bundles obtained in experiment 2.4.2. Colored bundles (in order) : [CST], [Cingulum, ILF, MCP], [Rostrum, Genu, Rostral Body, Anterior Midbody, Posterior Midbody, Isthmus, Splenium].

learning the global streamline structure, we still do not have enough accurate data to explore this area of research. In future works, as data become available, we plan on training on incomplete datasets (i.e. removing one or more bundles) in order to see the reconstruction and discovery capabilities of our models. Furthermore, we will explore how modifying the output of the RNN (e.g. predicting the parameters of a distribution) can improve the power of the model.

Chapitre 3

État de l'art et défis futurs en tractographie par apprentissage

Résumé

La tractographie par apprentissage fait face à plusieurs problèmes, dont celui que le modèle final dépend fortement des données utilisées pour l'entraîner. De plus, les données utilisées pour évaluer un modèle donné vont également influencer les résultats obtenus. À travers les années, plusieurs nouveaux algorithmes de tractographie par apprentissage ont été proposés, mais sans jamais partager exactement les mêmes données, ce qui rend la comparaison entre les modèles impossible. Cet article résume les différents modèles proposés, les problèmes à résoudre afin de pouvoir comparer de futurs modèles, ainsi que des pistes de solution et de bonnes pratiques à appliquer pour le futur.

Contributions de la publication

- Revue des bases de données d'intérêt pour la tractographie par apprentissage.

- Revue des modèles proposés pour la tractographie par apprentissage.
- Énumération des problèmes liés à la comparaison des résultats existants.
- Proposition de solutions futures pour adéquatement entraîner, évaluer et comparer des modèles.

Contributions des auteurs

- Revue de littérature et écriture du manuscrit (Philippe Poulin)
- Recherche, révision et édition du manuscrit (Daniel Jörgens)
- Supervision du projet et révision du manuscrit (Pierre-Marc Jodoin et Maxime Descoteaux)

Commentaires

Cet article est le résultat de quelques années à constater la difficulté à comparer les résultats des nouveaux algorithmes proposés. Il a été publié dans le journal *Magnetic Resonance Imaging*, en décembre 2019, dans le cadre d'une édition spéciale sur l'intelligence artificielle en IRM.

Tractography and machine learning: Current state and open challenges

Philippe Poulin¹, Daniel Jörgens², Pierre-Marc Jodoin^{*1}, Maxime Descoteaux*

¹ Computer Science Department, Université de Sherbrooke, Sherbrooke, Canada

² Department of Biomedical Engineering and Health Systems, KTH Royal Institute of Technology, Stockholm, Sweden

* Co-last author

Abstract

Supervised machine learning (ML) algorithms have recently been proposed as an alternative to traditional tractography methods in order to address some of their weaknesses. They can be path-based and local-model-free, and easily incorporate anatomical priors to make contextual and non-local decisions that should help the tracking process. ML-based techniques have thus shown promising reconstructions of larger spatial extent of existing white matter bundles, promising reconstructions of less false positives, and promising robustness to known position and shape biases of current tractography techniques. But as of today, none of these ML-based methods have shown conclusive performances or have been adopted as a *de facto* solution to tractography. One reason for this might be the lack of well-defined and extensive frameworks to train, evaluate, and compare these methods.

In this paper, we describe several datasets and evaluation tools that contain useful features for ML algorithms, along with the various methods proposed in the recent years. We then discuss the strategies that are used to evaluate and compare those methods, as well as their shortcomings. Finally, we describe the particular needs of ML tractography methods and discuss tangible solutions for future works.

3.1. INTRODUCTION

3.1 Introduction

In the field of diffusion magnetic resonance imaging (dMRI), tractography refers to the process of inferring streamline structures that are locally aligned with the underlying white matter (WM) dMRI measurements [65]. A simple approach to obtain such streamlines is an iterative process in which, starting from a seed point, an estimate of the local tissue orientation is determined and followed for a certain step length before repeating the orientation estimation at the new position. The tracking procedure may be deterministic [166, 7] (at each point, the algorithm follows the strongest orientation) or probabilistic [9, 140, 141] (at each point, the algorithm samples a direction closely aligned with the strongest orientation). Tracking may also be global as some methods recover streamlines all at once [111, 84, 63]. In between the local and global methods is the category of shortest-path methods, including front evolution, simulated diffusion, geodesic, and graph-based approaches [65]. Ultimately, the collection of all trajectories created in that way is called a tractogram.

In traditional methods, the estimate of the local tissue fiber orientation is usually inferred from an explicit and local model which fits the (local) diffusion data. These local models include diffusion tensor models [7, 99], multi-tensor models [17, 83], and other methods that aim at reconstructing the fiber orientation distribution function (fODF) like constrained spherical deconvolution (CSD) [138, 37], to name a few. However, the choice of the best model is by itself difficult [122, 82], as it depends on various factors such as data acquisition protocol or targeted WM regions, and therefore has a direct influence on the quality of an obtained tractogram [27]. Moreover, traditional methods based on local orientation alone are prone to make common mistakes, such as missing the full spatial extent of bundles and producing a great amount of false positive connections [82].

Another important factor for the performance of a tractography method are the actual rules that regard the progression of a single step as well as simple global properties of an individual streamline. Traditional methods may define several engineered, or “manually-defined”, high-level rules with the aim of improving the anatomical plausibility of the recovered tractogram. Instances of these are constraints on streamline length (i.e. filtering streamlines that are too long or too short), streamline shape

3.1. INTRODUCTION

(e.g. filtering streamlines with sharp turns), or progression rules that make streamlines “bounce off” the WM border when they are about to leave the WM mask with a certain angle [53, 93]. In the same way as modeling noise and artifacts, and defining the right local model, also the design of these high-level rules has a direct impact on the performance of a tractography method [27, 82].

To address these inherent difficulties, recent proposals suggest that machine learning (ML) algorithms, supervised or unsupervised, may be used to implicitly learn a local, global or contextual fiber orientation model as well as the tracking procedure. Approaches ranging from the application of self-organizing maps (SOM) [41, 40], random forests (RF) [93, 92], Multilayer Perceptrons (MLP) [67, 160], Gated Recurrent Units (GRU) [101, 11, 102], as well as Convolutional Neural Networks (CNN) [157] and Autoencoders [80], have been employed at the core of tractography to drive streamline progression. Apart from the differences in their underlying architecture, these ML methods differ substantially in aspects of the exact problem formulation, e.g. definition of the input data to the model, modeling the predictions as a regression [101, 80] or classification problem [92, 67], or even the general tractography approach, i.e. whole-brain [101, 11] or bundle-specific [157, 80, 112, 102]. The fact alone that these approaches differ in several aspects, makes it difficult to draw conclusions on the value of each of the individual modeling choices.

Furthermore, while the above mentioned approaches constitute the main ideas for applying ML directly to the process of tractography, machine learning and especially deep learning (DL) methods have been applied in related fields. Stacked U-Nets were proposed to segment the volume of individual white matter bundles from images of fODF peaks [155]. It was also suggested to predict fiber orientations from raw diffusion data based on convolutional neural networks (CNN) [71]. Several ideas for streamline clustering or streamline segmentation have been proposed, including a CNN based on landmark distances [94], a long short-term memory (LSTM)-based siamese network for rotation invariant streamline segmentation [66], and a CNN approach for streamline clustering based on the sequence of their coordinates [58, 59]. Even though the mentioned works are closely related to tractography and contribute to the common goal of improved analysis of the white matter anatomy of the human brain, we restrict our focus exclusively on the direct application of ML (and especially

3.1. INTRODUCTION

DL) for tractography, with the explicit goal of producing streamlines and addressing the weaknesses of traditional methods. For that reason, we refer the interested reader to the respective references for more details.

An important factor for effectively advancing this field of research is a common and appropriate methodology for training and evaluating the performance of different approaches, which is currently lacking. Over the years, multiple challenges have been proposed to assess the performance of conventional tractography methods, and a clear and exhaustive review is provided by Schilling et al. [122]. However, we argue that the design of these challenges is typically inappropriate for ML methods. In fact, the *2015 ISMRM Tractography Challenge* [82] (along with the *Tractometer* evaluation tool [27]) has been adopted as the tool of choice for benchmarking new ML tractography pipelines [101, 93, 160, 11]. Unfortunately, several inherent flaws arising specifically in the context of ML make it difficult to perform a fair comparison between the results obtained from different ML pipelines. In particular, diffusion data preprocessing is left to participants (**dissimilar inputs**), tracking seeds and a tracking mask are not always given (**varying test environment**), the test diffusion volume is sometimes used for training (**data contamination**), training streamlines are not provided (**disparate training data**), and testing on a single synthetic subject means that any computed estimator of a model’s performance is unreliable (**small sample size**). Against the background of a prospectively increasing number of ML-based approaches tackling the problem of tractography, a carefully designed evaluation framework that appropriately addresses the specific requirements of ML methods has the potential to support and facilitate research in this field in the upcoming years.

In this paper, we follow a threefold strategy. First, we introduce the currently available datasets and evaluation tools along with useful features and weaknesses regarding machine learning. Then, we provide a comprehensive review of existing ML-based tractography approaches and derive a set of key concepts distinguishing them from each other. Subsequently, we identify and discuss the strategies for evaluation of tractography pipelines and identify issues and limitations arising when applied to ML-based tractography methods. We finally describe important features for an appropriate evaluation framework the community ought to adopt in the near future

3.2. ANNOTATED DATASETS AND EVALUATION TOOLS

to better promote data-driven streamline tractography and point out the potential advantages for research in data-driven streamline tractography.

3.2 Annotated datasets and evaluation tools

Over the years, many diffusion MRI datasets were produced and annotated, either as part of a challenge or research papers. In this section, we overview several datasets that have been used to train and/or validate supervised learning algorithms for tractography. Specifically, we selected datasets that offer both diffusion data and streamlines. Selected datasets also needed to have either clearly defined evaluation metrics, or to be large enough (more than 50 subjects) to be considered as standalone training sets. We include datasets that are either publicly available or simply mentioned in a research paper without a public release.

We excluded datasets or challenges focused on non-human anatomy (e.g. rat or macaque), where the ground truth is harder to define and results might be harder to generalize to human anatomy (for data-driven algorithms), like the *2018 VOTEM Challenge* [135] (my.vanderbilt.edu/votem/). Moreover, we left out datasets focused only on pathological cases like the *2015 DTI Challenge* [107], because we consider it too early for data-driven tractography algorithms, at least until more conclusive results on healthy subjects. We also excluded tractography atlases when tracking was done on a single diffusion volume, usually averaged over multiple subjects (e.g. HCP842 [167]), because results tend to be overly smooth and unsuited for ML methods. However, we include a recent case when tracking was done for each subject: the 100-subjects WM atlas of Zhang et al. [169].

While all the selected datasets are useful in one way or another for data-driven methods, they differ in multiple ways, which are detailed in the following subsections and summarized in Table 3.1. The listed properties are the following:

- **Name:** The dataset name and reference
- **Year:** The year of publication of the dataset or paper using the dataset
- **Public:** Is the dataset (diffusion data and streamlines) publicly available?
- **Real:** Is the diffusion data a real acquisition or is it simulated?
- **Human:** Does the diffusion data represent the human brain anatomy?

3.2. ANNOTATED DATASETS AND EVALUATION TOOLS

- **Subjects:** The number of subjects or acquisitions
- **Bundles:** The number of bundles or tracks (if streamlines are available)
- **GT:** Is a ground truth known? For real acquisitions, streamlines validated by a human expert (e.g. neuroanatomist) are considered as GT despite the fact that these annotations are subject to inter-rater and intra-rater variations.
- **Metrics:** Well-defined evaluation metrics are available with this dataset.
- **Split:** Is the dataset split into a training and testing set that future works can rely on?

Note that the notion of “ground truth” refers to an indisputable biologically-validated label assigned to an observed variable. In medical imaging, such ground truth may be obtained with a biopsy [136], throughout careful complementary analysis [13] or by having several experts agreeing on a given diagnostic [12]. Unfortunately, such restrictive definition of a ground truth is unreachable most of the time, especially for white matter tracks obtained from tractography, where no expert can truly assess the existence (or non-existence) of a given streamline in a human brain from MRI images only. In fact, only synthetically-generated streamlines or man-made phantoms can be considered as real “ground truth”. Despite that, for the purpose of this paper, we also use the term “ground truth” for any data that has been manually validated by a human expert, typically a neuro-anatomist. In the medical imaging field, this annotated data would be called a *gold standard*, while in the artificial intelligence community, it might be called *weakly annotated data*. Although such annotations do not meet the fundamental definition of a ground truth, it is nonetheless widely accepted by the medical imaging AI community [87].

3.2.1 The FiberCup dataset and the Tractometer tool

Original FiberCup Tractography Contest (2009) Fillard et al. proposed the *FiberCup Tractography Contest* [48, 105] in conjunction with the 2009 MICCAI conference. The goal was to quantitatively compare tractography methods and algorithms using a clear and reproducible methodology. They built a realistic diffusion MR 7-bundle phantom with varying configurations (crossing, kissing, splitting, bending). The organizers acquired diffusion images with b-values of 2000, 4000, and

3.2. ANNOTATED DATASETS AND EVALUATION TOOLS

Table 3.1 – Annotated datasets.

Name	Year	Public	Real	Human	Subjects	Bundles	GT	Metrics	Split
Fibercup [48]	2009	✓	✓		1	7	✓	✓	
Simulated Fibercup [164]	2012	✓			1	7	✓	✓	
Tracula [168]	2011		✓	✓	67	18	✓		
HARDI 2012 [29]	2012	✓			2	7	✓	✓	✓
HARDI 2013 [28]	2013	✓			2	20	✓	✓	✓
ISMRM 2015 [82, 27]	2015	✓			1	25	✓	✓	
HAMLET [112]	2018		✓	✓	83	12			✓
PyT (BIL&GIN) [20]	2018		✓	✓	410	2	✓		
BST (BIL&GIN) [115]	2018		✓	✓	39	5	✓	✓	
TractSeg (HCP) [155]	2018	✓	✓	✓	105	72	✓		
Zhang et al. (HCP) [169]	2018		✓	✓	100	58 + 198	✓		

6000 s/mm², and used isotropic resolutions of 3mm and 6mm, resulting in 6 different diffusion datasets. Contestants were provided all datasets (but not the ground truth) and were free to apply any preprocessing they wanted on the diffusion images. Evaluation was done by choosing 16 specific voxels, or seed points, in which a unique fiber bundle is expected. Participants were expected to submit a single fiber bundle for each of those seed voxels. Quantitative evaluation was done by comparing the 16 pairs of candidate and ground truth fibers using a symmetric Root Mean Square Error (sRMSE).

While the *FiberCup Tractography Contest* makes a good test case for simple configurations, it does not represent a true human anatomy and does not impose a choice of b-value and preprocessing, which can induce significant differences in data-driven methods. Also, it does not provide any training streamlines, and is thus useful only as a validation tool for ML-based methods. Furthermore, the fact that it contains only one subject makes it hard to evaluate the true generalization capability of an ML method trained and tested on that dataset. However, it is the only dataset that provides seed points in order to have a uniform test environment, which is of utmost importance when comparing ML-based algorithms. In the end, it is unclear if for ML-based methods there would be any correlation between a good performance on the FiberCup contest and good performance on human anatomy.

Tractometer evaluation tool (2013) In 2013, Côté et al. developed the *Tractometer* evaluation tool, to be used alongside the original FiberCup data, with the

3.2. ANNOTATED DATASETS AND EVALUATION TOOLS

aim of providing quantitative measures that better reflect brain connectivity studies. Using a Region of Interest (ROI)-based filtering method, a complete tractogram can be evaluated on global connectivity metrics, such as the number of valid and invalid bundles. Furthermore, they propose two seeding masks: a complete mask (mimicking a brain WM mask), and a ROI mask (mimicking GM-WM interfaces). The tractometer was designed to address the fact that “metrics are too local and vulnerable to the seeds given, and, as a result, do not capture the global *connectivity* behavior of the fiber tracking algorithm”[27].

Simulated FiberCup (2014) In 2014, Neher et al. proposed a simulated version of the FiberCup, allowing new tracking algorithms to be tested using multiple acquisition parameters [91]. The simulated data can be used alongside the *Tractometer* tool designed for the original phantom.

Wilkins et al. also developed a synthetic version of the FiberCup dataset, but did not publicly release the data [164]. Unfortunately, with regards to ML methods, the simulated FiberCup dataset suffers from the same shortcomings as the original FiberCup dataset as it contains only one non-human subject whose data is not split *a priori* into a training and testing set.

3.2.2 Tracula (2011)

Yendiki et al. [168] published the Tracula method for automated probabilistic reconstruction of 18 major WM pathways. It uses prior information on the anatomy of bundles from a set of training subjects. The training set was built from 34 schizophrenia patients and 33 healthy controls, using a 1.5T Siemens scanner as part of a multi-site MIND Clinical Imaging Consortium [163]. The diffusion images include 60 gradient directions acquired with a b-value of 700 s/mm^2 , along with 10 b=0 images, with an isotropic resolution of 2mm. Whole-brain deterministic tracking was performed, followed by expert manual labeling using ROIs for 18 major WM bundles. The dataset also includes a measure of the inter-rater and intra-rater variability for the left and right uncinate.

To our knowledge, this is the earliest apparition of a large-scale human dataset with expert annotation of streamlines. It is also the only dataset that includes a

3.2. ANNOTATED DATASETS AND EVALUATION TOOLS

measure of inter-rater and intra-rater variability, which is a desirable feature for ML methods (also discussed later in Section 3.4.5). Unfortunately, the complete set of diffusion images and streamlines has been incorporated into the method and is not public.

3.2.3 HARDI Reconstruction Challenges

HARDI Reconstruction Challenge (2012) Daducci et al. organized the *2012 HARDI Reconstruction Challenge* [28] at the ISBI 2012 conference. The goal of the challenge was to quantitatively assess the quality of intra-voxel reconstructions by measuring the predicted number of fiber populations and the angular accuracy of the predicted orientations. A training set was released prior to the challenge, and a test set was used to score the algorithms. As such, the 2012 HARDI dataset contains diffusion images but no streamlines.

Participants could request a custom acquisition (only once) by sending a list of sampling coordinates in q-space, and the organizers would then produce a simulated signal for the given parameters. A $16 \times 16 \times 5$ volume was then produced, containing seven different bundles attempting to recreate realistic 3-D configurations. The metrics proposed by the authors are ill-posed for ML-based methods because of the limited context available and the focus on local performances. Like the *FiberCup*, it would only be useful as a validation tool given the lack of training streamlines, a limited number of bundles (only seven) and a limited number of non-human subjects (only two).

HARDI Reconstruction challenge (2013) The *2013 HARDI Reconstruction Challenge* [29] was organized one year later at the ISBI 2013 conference. For ML-based methods, three improvements are relevant compared to the 2012 challenge: a more realistic simulation of the diffusion signal, a new evaluation system based on connectivity analyses and a larger set of 20 bundles. Indeed, data-driven methods try to learn an implicit representation without imposing a model on the signal, which means that the signal used for training and testing should be as close as possible to that in clinical practice. Furthermore, the main benefit of data-driven methods is

3.2. ANNOTATED DATASETS AND EVALUATION TOOLS

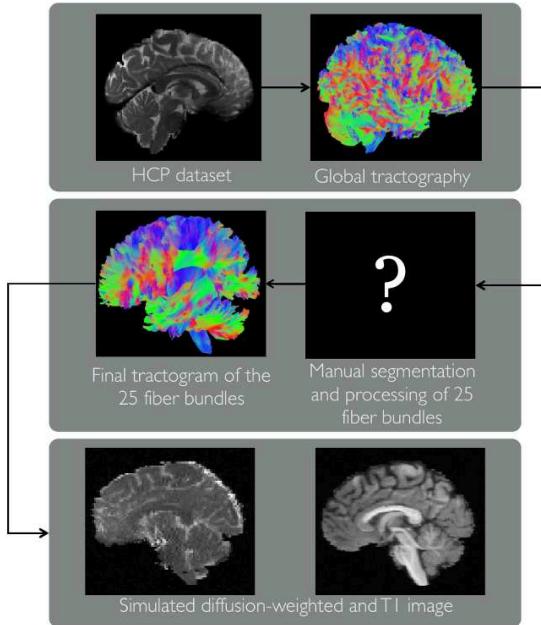


Figure 3.1 – 2015 ISMRM Tractography Challenge data generation process (Taken from www.tractometer.org)

the ability to use context in order to make good predictions in a multitude of configurations, which means they have the potential to particularly improve connectivity analyses. Therefore, it would be a better validation tool for ML-based methods than the 2012 HARDI Reconstruction Challenge. Nonetheless, the dataset suffers from an inherent limitation as it contains only two non-human subjects.

3.2.4 ISMRM Tractography Challenge (2015)

This dataset has been designed for a tractography challenge organized in conjunction with the 2015 ISMRM conference [82]. During the challenge, participants were asked to reconstruct streamlines from a synthetic human-like diffusion-weighted MR dataset which was simulated with the aim of replicating a realistic, clinical-like acquisition, including noise and artifacts. The available data consists of a diffusion dataset with $32 b=1000 \text{ s/mm}^2$ images and one $b=0$ image, with 2mm isotropic resolution, as well as a T1-like image with 1mm isotropic resolution. Since all data was generated from an expert segmentation of 25 bundles, in theory, a perfect tracking algorithm

3.2. ANNOTATED DATASETS AND EVALUATION TOOLS

should only produce exactly these specific bundles. Unfortunately, as for the HARDI and FiberCup datasets, the *2015 ISMRM Tractography Challenge* contains data from a limited number of subjects (only one) and lacks a clear separation between training and testing data. Nonetheless, in combination with the *Tractometer* tool [27], this dataset has often been used to assess ML-based tractography methods. Figure 3.1 shows the data generation process for the challenge.

Once a tractogram has been generated using the challenge diffusion data, the tractometer tool uses a “bundle recognition algorithm” [52] to cluster the streamlines into bundles. The generated bundles are then compared to the ground truth, producing groups of “valid bundles” and “invalid bundles”, depending on which regions of the brain the streamlines connect. Streamlines that do not correspond to a ground truth bundle are classified as “No connections” streamlines. The metrics computed by the modified *Tractometer* for the *Tractography Challenge* are as follows:

- **Valid bundles** (VB): The number of correctly reconstructed ground truth bundles.
- **Invalid Bundles** (IB): The number of reconstructed bundles that do not match any ground truth bundles.
- **Valid Connections** (VC): The ratio of streamlines in valid bundles over the total number of produced streamlines.
- **Invalid Connections** (IC): The ratio of streamlines in invalid bundles over the total number of produced streamlines.
- **No Connections** (NC): The ratio of streamlines that are either too short or do not connect two regions of the cortex over the total number of produced streamlines.
- **Bundle Overlap** (OL): The ratio of ground truth voxels traversed by at least one streamline over the total number of ground truth voxels.
- **Bundle Overreach** (OR): The ratio of voxels traversed by at least one streamline that do not belong to a ground truth voxel over the total number of ground truth voxels.
- **F1-score** (F1): The harmonic mean of recall (OL) and precision (1-OR).

The definition of streamline-oriented metrics (VB, IB, VC, IC, NC) and volume-oriented metrics (OL, OR, F1) means that there is no single number that can fully

3.2. ANNOTATED DATASETS AND EVALUATION TOOLS

assess the performance of an algorithm. For example, deterministic methods often score higher on streamline-oriented metrics compared to probabilistic methods. As such, a thorough review of all scores must be performed in order to properly compare algorithms, and in many cases, the choice of an algorithm over another may depend on a specific use-case (e.g. bundle reconstruction vs. connectivity analysis).

3.2.5 HAMLET (2018)

To validate their method, Reisert et al. [112] used a dataset of 83 human subjects from two independent cohorts. The first cohort comprises 55 healthy volunteers, all scanned by a Siemens 3T TIM PRISMA MRI scanner. The second cohort has 28 volunteers scanned with a Siemens TIM TRIO. The first cohort was used for training while the second one was used for testing. Subjects in the second cohort were scanned twice for test-retest experiments, some unique characteristic to that dataset. The reference streamlines were obtained by first tracking the whole brain with global tractography, and then by segmenting the streamlines for 12 bundles with a selection algorithm in MNI space. Unfortunately, the recovered streamlines have not been manually validated by an expert.

3.2.6 Datasets based on the BIL&GIN database

Bundle-Specific Tractography (2018) Rheault et al. proposed a bundle-specific tracking method based on anatomical priors that improves tracking in the centrum semiovale crossing regions [115]. Using multiple tractography algorithms, they tracked and segmented five bundles (Arcuate Fasciculus - AF left/right, Corpus Callosum - CC, Pyramidal Tracts - PyT left/right) in 39 subjects from the BIL&GIN database [86]. To compare algorithms, they used an automatic bundle segmentation method based on clear anatomical definitions. In addition, they defined several performance metrics, such as *bundle volume*, *ratio of valid streamlines*, and *efficiency*. However, the tractograms and automatic bundle segmentation procedure were neither made public nor validated by an expert. Such a dataset, along with the evaluation procedure, could be extremely useful to assess if data-driven methods can reliably learn the structure of a specific bundle and reconstruct it in unseen subjects.

3.2. ANNOTATED DATASETS AND EVALUATION TOOLS

A population-based atlas of the human pyramidal tract (2018) Chenot et al. created a streamline dataset of the left and right PyT based on a population of 410 subjects [20], also from the BIL&GIN database [86]. To do so, they combined manual ROIs along the bundles’ pathway and the bundle-specific tractography algorithm of Rheault et al. [115]. The quality of the segmentations and the high number of subjects would make this a noteworthy training dataset for data-driven methods. Unfortunately for ML methods, only two bundles were examined. Furthermore, while the probability maps of the atlas have been rendered public, the tractograms are still unavailable.

3.2.7 Datasets based on the HCP database

TractSeg (2018) Wasserthal et al. proposed a data-driven method for fast WM tract segmentation without tractography [155]. In doing so, they built an impressive dataset of 72 manually-validated bundles for 105 subjects from the Human Connectome Project (HCP) diffusion database [149, 54]. Tractograms were obtained via a four-step semi-automatic approach:

1. Tractography (Multi-Shell Multi-Tissue CSD [140])
2. Initial tract extraction (TractQuerier [154])
3. Tract refinement (Manual ROIs [130] + QuickBundles [49])
4. Manual quality control and cleanup

To the best of our knowledge, this is the largest public database to include both diffusion data and reference streamlines. No further preprocessing of the diffusion data is needed because of the standard procedure of [54]. The authors defined volume-oriented metrics such as Dice score [132], but did not offer any streamline-oriented metrics as their method predicts a volume segmentation. The high number of subjects and bundles makes this a remarkable training set.

In a subsequent paper, the same authors re-used a subset of 20 bundles of the TractSeg dataset to train and validate their TOM ML algorithm [157]. However, as for original 72-bundle dataset, the TOM dataset does not come with a predefined set of training and testing data and no formal evaluation protocol that users could rely on has been proposed.

3.3. MACHINE LEARNING METHODS FOR TRACTOGRAPHY

Zhang et al. (2018) Zhang et al. [169] built a WM fiber atlas using 100 HCP subjects. They first generated streamlines for all subjects using a two-tensor unscented Kalman filter method [110], and sampled 10,000 streamlines from each subject after a tractography registration step. Then, using a hierarchical clustering method, the authors generated an initial WM fiber atlas of 800 clusters. Finally, an expert neuroanatomist reviewed the annotations in order to accept or reject each cluster, and provided the correct annotations when the initial annotation was rejected. The final, proposed atlas is comprised of 58 bundles (each composed of multiple clusters), along with “198 short and medium range superficial fiber clusters organized into 16 categories according to the brain lobes they connect” [169].

While the atlas is public, the sampled streamlines from the 100 subjects are all merged into the single template. In order for ML methods to benefit from this dataset, the streamlines would need to be separated back into the space of the particular original subjects. For this reason, we do not consider this dataset to be “public”, in the context of machine learning.

3.3 Machine learning methods for tractography

For this review, we regard all supervised machine learning methods published in peer-reviewed journals, conferences or on arXiv (arxiv.org) and biorXiv (biorxiv.org). We added the requirement that methods needed to be specifically designed for tractography, i.e. with the purpose of predicting a contextual streamline direction (and not reconstructing a local, non-conditional fODF or clustering streamlines). This criterion includes whole-brain as well as bundle-specific tractography methods. A summary of the main properties for all reviewed methods is provided in Table 3.2.

Random Forest classifier To the best of our knowledge, Neher et al. were the first to propose a machine learning algorithm for (deterministic) tractography [92]. They employ a RF classifier to learn a mapping from raw diffusion measurements to a directional proposal for streamline continuation. After collecting several of such proposals in a local neighborhood of the current streamline position (radius: 25% of the smallest side length of a voxel), these are aggregated in a voting scheme to finally

3.3. MACHINE LEARNING METHODS FOR TRACTOGRAPHY

Table 3.2 – Main properties of data-driven methods for tractography.

Method	Model	Temporal context	Spatial context	dMRI input	Prediction	Implicit stop
Neher et al. [93]	RF	1 last direction	50 samples	Resampled DWI	Classification	✓
Poulin et al. [101]	GRU	Full	1x1x1 voxel	SH	Regression	
Poulin et al. [102]	GRU	Full	1x1x1 voxel	SH	Regression	
Benou et al. [11]	GRU	Full	1x1x1 voxel	Resampled DWI	Classification	✓
Jörgens et al. [67]	MLP	2 last directions	1x1x1 voxel	Raw DWI	Regression	
Wegmayr et al. [160]	MLP	4 last directions	3x3x3 voxels	SH	Regression	
Wasserthal et al. [157]	CNN	N/A	Entire WM	fODF peaks	Regression	
Reisert et al. [112]	CNN-like	N/A	Entire WM	SH	Regression	✓

RF: Random Forest; MLP: Multilayer perceptron; GRU: Gated recurrent unit; CNN: Convolutional neural network; SH: Spherical harmonics coefficients; fODF: fiber Orientation Distribution Function; Implicit stop: indicates if a method learns its tracking stopping criterion or if it relies on a usual explicit criterion.

arrive at a single direction in which to grow the streamline.

To define reference streamlines for their experiments, the authors employ several tractography pipelines and train their classifier on each of the resulting tractograms. They determine the best trained model by evaluating the performance of each on a replication of the FiberCup phantom (based on the *Tractometer* metrics of [27]). Finally, comparing the performance of the latter to all other reference pipelines, they report a superior performance of their tracking model over all other approaches. While tractograms were scored on a simulated phantom (i.e. no real anatomy), extended experiments presented in a subsequent paper [93] confirm the superiority of their approach on the *2015 ISMRM Tractography Challenge* dataset (simulated data of a human anatomy).

Gated Recurrent Unit (GRU) Tracking Hypothesizing “that there are high-order dependencies between” the local orientation at a point of a streamline and the orientations at all other points on the same streamline, Poulin et al. proposed a recurrent neural network (RNN) based on a GRU [101] to learn the tracking process. Their method implements an implicit model mapping diffusion measurements to local streamline orientations which not only depends on measurements in a local context, but on all data previously seen along the extent of a particular streamline. As opposed to [93, 92], the RNN model is implemented as a regression approach. In their experiments, the authors show that a recurrent model (when trained on ref-

3.3. MACHINE LEARNING METHODS FOR TRACTOGRAPHY

erence streamlines obtained using deterministic CSD-based tractography [141]) was able to outperform most of the original submissions in the *2015 ISMRM Tractography Challenge* with respect to the *Tractometer* scores (discussed in section 3.2.4).

DeepTracker In a subsequent paper, Poulin et al. [102] again suggested using a GRU, but in a bundle-specific fashion. While the model architecture is very similar, it was trained on a dataset of 37 real subjects, each with a curated set of streamlines for bundles. After training a single model for each of the selected bundles, the authors showed promising results compared to existing methods, perhaps indicative that the difficult task of learning to track streamlines necessitates more data than previously thought.

DeepTract More recently, Benou et Riklin Raviv [11] proposed a GRU-based recurrent neural network similar to that of Poulin et al. [101]. In their method, they directly use the resampled diffusion signal as input to the model (like Neher et al. [93]), in order to estimate a discrete, streamline-specific fODF representation which they refer to as “conditional fODF” (CfODF). Instead of predicting a 3D orientation vector using a regression approach, the authors implement their model as a classifier enabling them to interpret the probabilities obtained for discrete sampled directions (i.e. the classes) as the mentioned CfODF. This fODF-based formulation further allows for an inherently defined criterion for streamline termination based on the entropy of the CfODF. The proposed model can be employed for both deterministic and probabilistic tractography.

Like Poulin et al. [101], the authors trained and tested their method on the *2015 ISMRM Tractography Challenge* dataset. They report results after training their method on the dataset ground truth as well as on streamlines obtained with the MITK diffusion tool [165].

Multi-Layer Perceptron Point-Wise Prediction Jorgens et al. [67] propose a multi-layer perceptron (MLP) to predict the next step of a streamline. Like [93, 92, 101], their method takes as input the diffusion signal and thus avoids explicit dMRI model-fitting. The authors implemented different configurations of their

3.3. MACHINE LEARNING METHODS FOR TRACTOGRAPHY

proposed MLP such as three different input scenarios (point-wise input vs region-wise input with and without considering previous orientations), different approaches to aggregate the output (maximum likelihood, mathematical expectation of the categorical prediction and regression) as well as the voting scheme proposed by Neher et al. [92]. Results reveal that the best configurations are those having the previous two directions included in the input of the network thus showing that temporal context is a key component for data-driven tractography. Also, the regression and classification approaches led to similar results and the use of region-wise information did not provide any substantial improvement over the use of point-wise information.

Like Poulin et al. [101] and Benou et Riklin Raviv [11], the authors trained and tested their method on the *2015 ISMRM Tractography Challenge* dataset (but did not use the *Tractometer* tool). Unfortunately, they did not estimate the tracking capabilities of their method as they only measured point-wise angular errors when predicting the next step of a streamline.

Multi-Layer Perceptron Regression Tracking A similar approach suggested by [160] employs a MLP to predict the next direction of a streamline through regression. At each point, the input of the model is given by all diffusion measurements in a cubic neighborhood, along with a certain number of previous steps for the current streamline. In that way, the authors provide the ML model directly with diffusion information in a local neighborhood (spatial context) as well as a notion of “history” of the current streamline (temporal context). Defining their reference streamlines as tractograms obtained with a standard tractography method from *in vivo* datasets, they train their model on three subjects from the HCP database. Experimental validations on the *2015 ISMRM Tractography Challenge* dataset reveal that their model outperforms some ML methods [93, 101] in most *Tractometer* metrics. However, as demonstrated by low overlap scores, the authors acknowledge that their model produces “rather confined bundles with little spread”, especially in contrast to [93, 101]. While the strength of this model is to explicitly provide information from a local neighborhood, like for Jorgens et al. [67], the notion of context along the streamline is limited and needs to be defined before training. Since the ideal temporal context (in terms of streamline length, or steps) is still unknown, this could potentially

3.3. MACHINE LEARNING METHODS FOR TRACTOGRAPHY

prohibit the model from taking advantage of all information relevant to streamline continuation.

Tract orientation mapping using an encoder-decoder CNN Wasserthal et al. [157] proposed a data-driven, bundle-specific tracking method. As opposed to the other ML methods reported in this paper, the authors do not try to directly reconstruct streamlines *per se*. Instead, their proposed *Tract Orientation Mapping* (TOM) method predicts bundle-specific fODF peaks that are then used by a deterministic tracking method. First, CSD is used to extract three principal directions in all WM voxels. Then, a U-Net CNN [119] is trained to map these fODF peaks to bundle-specific peaks, i.e. peaks that are only relevant for the streamlines of a given bundle. Their CNN takes as input 9 channels (the three fODF peaks) and outputs 60 channels, i.e. a 3D bundle-specific fODF vector for each of the 20 bundles they are looking to recover. While the recovered bundle-specific peaks can be used in different ways, the authors show that using them directly as input to a deterministic MITK diffusion tractography gives some of the best results. The approach was trained and tested on 105 HCP subjects, each with reference streamlines produced by a semi-automatic dissection of 20 large WM bundles (which they recently rendered public [157]).

HAMLET In a similar line of thought, in their HAMLET project (*Hierarchical Harmonic Filters for Learning Tracts from Diffusion MRI*) Reisert et al. [112] map raw spherical harmonics of order 2 to a spherical tensor field. In that sense, like Wasserthal et al. [157], their ML method does not output streamlines but instead voxel-wise bundle-specific tensors that can subsequently be used as input to a classical tractography method. The magnitude of the produced 3×3 tensor indicates the presence of a specific bundle whereas the tensor orientation predicts the local streamline direction. Their method implements a multi-resolution CNN with rotation covariant convolution operations. They trained and tested their method on two in-house datasets comprising a total of 83 human subjects. The 12 bundles and their associated reference streamlines have been obtained with global tractography and automatic bundle selection method. Unfortunately, the reference data was not manually validated by a human expert, and they did not perform any comparisons against

3.4. RESULTS & DISCUSSION

other tractography methods.

3.4 Results & Discussion

3.4.1 Results on the 2015 ISMRM Tractography Challenge

The *2015 ISMRM Tractography challenge* is the only dataset that has been used to assess performance of several data-driven tractography methods and is thus, as of today, the only available common ground on which to compare methods. It was used by four different papers namely, the Random-Forest of Neher et al. [93], the GRU of Poulin et al. [101] and Benou et Riklin Raviv [11], and the MLP of Wegmayr et al. [160]. Experimental results reported by the authors have been transcribed in Table 3.3, and compared with original submissions in Figure 3.2. Note that the metrics marked as *not available* (N/A) are those the authors did not report in their original paper.

As can be seen, results vary a lot and there is no clear trend showing which method performs best, especially given the nature of the evaluation metrics. As mentioned in section 3.2.4, methods can be evaluated using both *streamline-oriented* metrics and *volume-oriented* metrics, which are not always correlated. For example, a method may have a large number of valid connections but a low overlap (like the MLP of Wegmayr et al.) which means that although the model was able to recover most valid bundles, the generated streamlines do not properly cover the spatial extent of those bundles. Also, a method can be more conservative and score best in terms of invalid connections and overreach like the GRU of Benou et Riklin Raviv, but at the same time have a low ratio of valid connections and a poor bundle overlap. On the other hand, the Random-Forest of Neher et al. does not score best in any category, but is competitive according to all metrics (its large F1-score underlines that it is a more balanced method compared to MLP and DeepTract). On top of that, all methods were trained and evaluated differently, so any comparison based on the reported results should be done with extreme care.

3.4. RESULTS & DISCUSSION

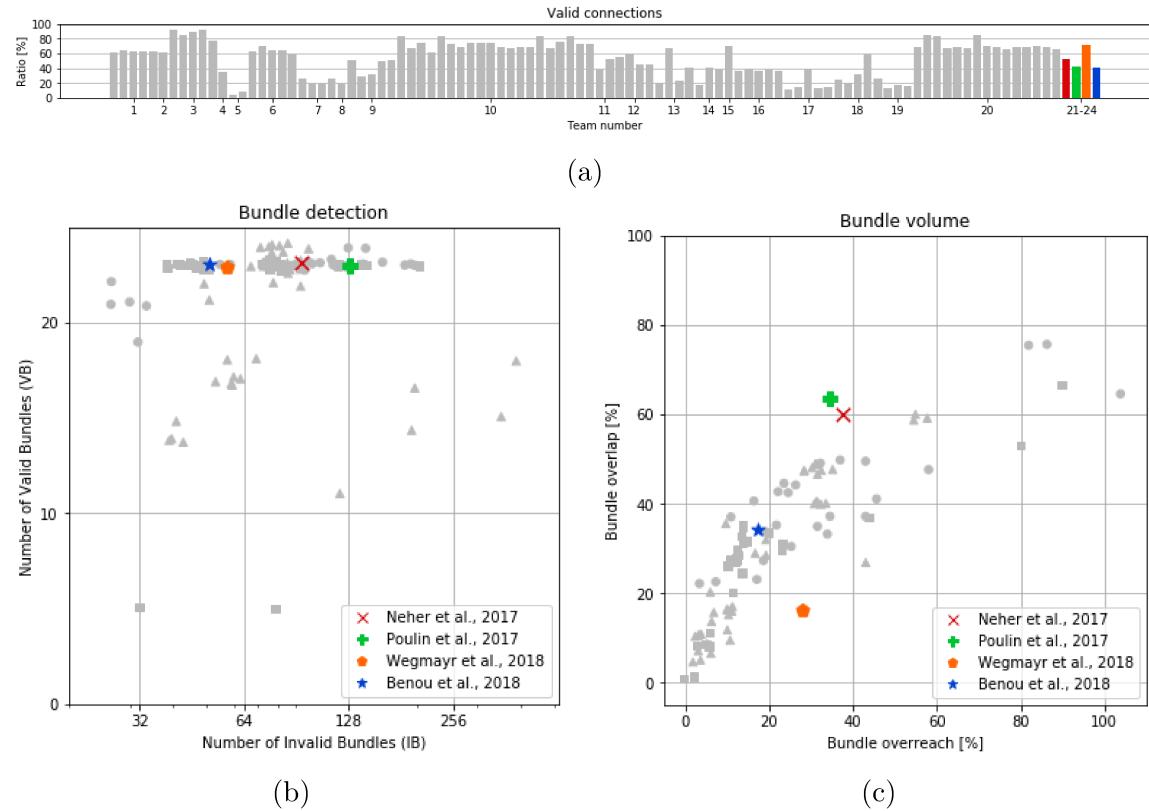


Figure 3.2 – 2015 ISMRM Tractography Challenge original submissions (1-20) and new results (21-24)

3.4. RESULTS & DISCUSSION

Table 3.3 – Tractometer results. The Bundles and Connections (%) metrics are *streamline-oriented metrics* whereas the Avg. bundle (%) metrics are *volume-oriented metrics*

Model	Bundles		Connections (%)			Avg. bundle (%)		
	Valid	Invalid	Valid	Invalid	No connection	Overlap	Overreach	F1-score
Random-Forest [93]	23	94	52	N/A	N/A	59	37	61
GRU [101]	23	130	42	46	13	64	35	65
MLP [160]	23	57	72	N/A	N/A	16	28	26
GRU (DeepTract) [11]	23	51	41	33	23	34	17	44

Table 3.4 – Differences in data

Method	Preprocessing	WM mask	Training subjects	Reference streamlines
Random-Forest [93]	<i>dwidenoise + dwipreproc</i>	Not needed	5 HCP subjects	CSD (Deterministic)
GRU [101]	None	Ground Truth	Challenge subject	CSD (Deterministic)
MLP [160]	<i>dwipreproc</i>	N/A	3 HCP subjects	iFOD (Probabilistic)
DeepTract [11]	N/A	Not needed	Challenge subject	Q-Ball (Probabilistic)

3.4.2 The 2015 ISMRM Tractography Challenge as an evaluation tool for ML algorithms

As mentioned before, the *2015 ISMRM Tractography Challenge* has been adopted as the *de facto* evaluation tool to compare ML tractography methods. However, the strengths and weaknesses of that tool should be thoroughly reviewed to understand and trust any technique reporting results with it. In this section, we present what we consider to be important issues with the way in which this tool has been used to assess the performance of data-driven methods. In particular, we detail the discrepancies between the four ML-based methods, differences that may explain some of the results in Table 3.3 and potentially undermine any conclusion that one could draw from it. Let us mention that some of these issues with the 2015 ISMRM dataset are typical for the field of tractography as a whole.

Table 3.4 presents a summary of the differences in how the tool is used. Note that the *not available* (N/A) mark is used for any information the authors did not mention in their original paper.

Dissimilar inputs The four ML methods use a different preprocessing pipeline. Among the proposed algorithms, two applied MRtrix’s *dwidenoise* or *dwipreproc* (www.mrtrix.org), another one denoised using [85] and corrected for eddy currents

3.4. RESULTS & DISCUSSION

and head motion, and another one did not apply any preprocessing at all. Moreover, some used the diffusion signal directly as input, while others resampled it to a specific number of gradient directions. In some cases, spherical harmonics were fitted to the signal and the SH coefficients were fed as input to the model. Finally, the non-recurrent models are also given a variable number of previous streamline directions as input.

The output of each of these pipelines contain various degrees of information. For example, fODF peaks are in theory already aligned with the major WM pathways, and information may be lost depending on the specific model used to recover the peaks from the diffusion signal. On the other hand, using the raw diffusion signal might contain more information but is more difficult to understand and process, and thus a data-driven model might require more capacity to use such an input. Without a thorough investigation of the information contained in each output, any variations in the *Tractometer* results could be attributed to the variations in preprocessing. Since we currently do not have any indication of what is useful for data-driven algorithms, it is impossible to compare ML methods if they do not use the same input data.

Varying test environment Since no white matter mask is provided, it must be computed by each participant in case it is needed for tracking. Out of the four ML methods that were evaluated on the challenge, two needed WM masks; one used the ground truth mask, and the other did not mention how the mask was computed. Furthermore, since no tracking seeds are supplied with the data either, their arrangement entirely depends on the WM mask (and the number of seeds per voxel, which is also not given).

Given the nature of streamline tractography, small variations of the tracking mask or the tracking seeds could have a substantial impact on the resulting streamlines and by that also on the obtained evaluation metrics. Also, even though computing a stopping criterion within the algorithm is a worthy improvement, it is a different task than tracking, and should be evaluated separately. Consequently, all methods should be provided the same tracking mask and seeds to reduce as much as possible the number of free variables during evaluation.

3.4. RESULTS & DISCUSSION

Data contamination The use of ML methods requires special care when dealing with available data. Since machine learning models are obtained by deriving implicit rules **directly from given data** (i.e. *training data*), testing the true generalization capabilities of these rules must be done using a **different and unseen set of data** (i.e. *test data*).

Two methods suffer from data contamination, or *leakage* [68]: the GRU in [101] and the MLP in [160]. Here, data contamination refers to the usage of the same diffusion data for training and testing. This means that the true generalization capabilities of the tested method on new, thus unseen subjects are still unknown, since the model has already seen the specific diffusion patterns that are needed in order to “explore” at test time, and therefore has been given an “unfair” advantage.

Disparate training data All methods used different reference streamlines and subjects for training. As mentioned earlier, some employed the test diffusion data directly, while others relied on a varying number of subjects from the HCP database. Two methods used deterministic CSD tracking [141] to generate reference streamlines, one used QBI tracking [1] (probabilistic) and the last one used iFOD tracking [140] (also probabilistic). In order to provide a uniform basis for comparison, the same comprehensive streamline training set should be available to every algorithm.

Simulation as a substitute for human acquisition While the diffusion signal of the 2015 ISMRM dataset is typical of that of a human brain, it is nonetheless obtained through simulation. As such, results on that dataset should not be seen as a measure of future performance on real human subjects, at least not without further empirical evaluation. Furthermore, at the given resolution and using this particular configuration of 25 bundles, false positive streamlines that would otherwise be plausible given the underlying anatomy of a real scan might be impossible to avoid. Indeed, some authors tried training their models using the ground truth bundles, and still produced over 50 invalid bundles in both cases [93, 11].

Small sample size The *2015 ISMRM Tractography Challenge* dataset has only one subject, which makes it hard to assess the future performance of a data-driven

3.4. RESULTS & DISCUSSION

algorithm [109]. In order to compute unbiased estimates of future performance, a richer testset with more subjects is needed. Also, given more subjects, bootstrapping methods [43] (i.e. sampling with replacement) could help to build more accurate estimators.

3.4.3 Other results

Some authors report local performance measures, such as the mean angular error [67]. However, local metrics do not take into account compounding errors, which can have a major effect on global structure. Consequently, global evaluation metrics should be preferred.

Tractography papers often report a visual evaluation on unseen, *in vivo* subjects, as a qualitative evaluation. For example, Figures 3.3 and 3.4 compare some of the proposed data-driven approaches with standard tractography methods on white matter bundles with known anatomy. However, in absence of a ground truth or the expertise of a neuroanatomist, it is hard to draw definitive conclusions on the quality of such results. In addition, Reisert et al. [112] presented correlation plots to assess reproducibility, but only offered qualitative comparisons with the reference streamlines without any quantitative results. To gain trust in these data-driven methods, a more rigorous approach is needed.

Finally, most ML methods offer a reduction in computation time compared to traditional methods. This is a non-negligible benefit, should these methods be adopted in practice.

3.4.4 Limitations of machine learning algorithms for tractography

Bundle exploration Supervised machine learning algorithms, by definition, learn to reproduce structured information from a given training dataset. In the case of tractography, the structure of bundles is an integral part of the training dataset, and bundle reconstruction is indeed at the core of all the proposed data-driven tractography methods. As such, trained methods should be expected to be able to reproduce those bundles. However, they should not be expected to be able to “extrapolate”

3.4. RESULTS & DISCUSSION

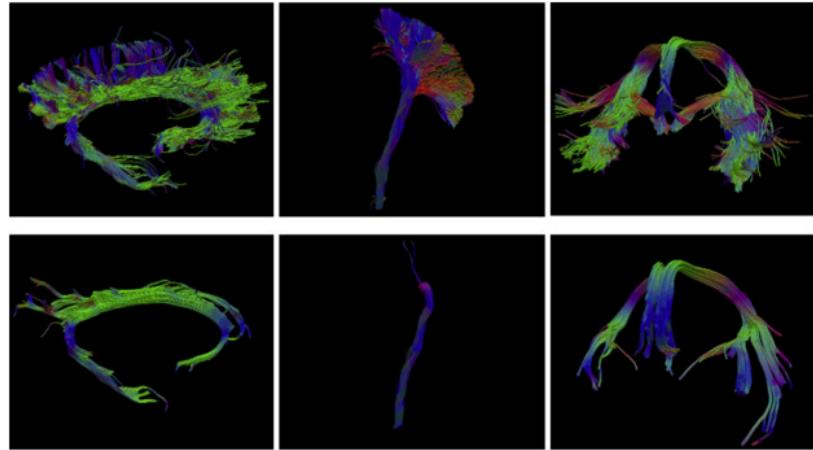


Figure 3.3 – Comparison between the RF of Neher et al. (top row), and classical deterministic CSD streamline tractography (bottom row). Results obtained on HCP subject 992774. (*Taken from [93] with authorization from the authors*)

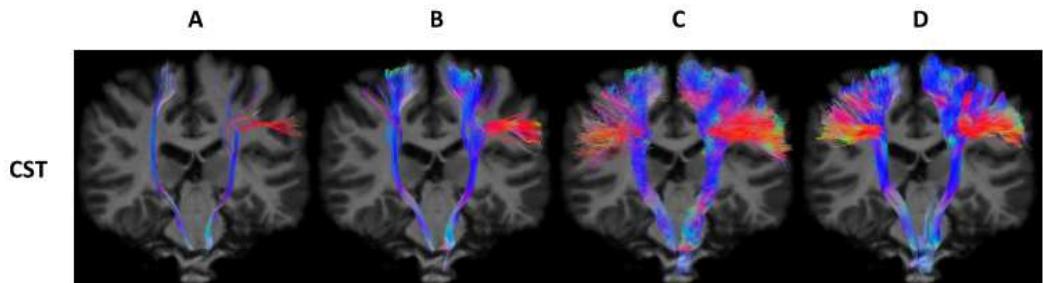


Figure 3.4 – Comparison of various tracking methods: A: Deterministic, B: Deterministic Bundle-Specific (DET-BST) [115], C: Probabilistic particle filter BST (PROB-PF-BST) [53], D: DeepTracker [102]. Results obtained on a BIL&GIN subject. (*Taken from [102] with authorization from the authors*)

3.4. RESULTS & DISCUSSION

the brain and “discover” new bundles unseen in training. In fact, it is still unknown if data-driven methods can extract enough information from the diffusion signal to solve all possible fiber configurations, even when provided with temporal or spatial context.

Until it is shown otherwise, machine learning algorithms for tractography should not be expected to reconstruct bundles unseen during training, and thus should not be trusted to extend their reconstruction capabilities to new connections or new brain regions like traditional methods.

Signal noise Working with diffusion MRI can be difficult because it usually provides a low SNR compared to other modalities, and can be subject to different noise distributions, such as non-central Chi distributions (which includes the Rician distribution), or the Gaussian distribution [25, 150, 8]. Having access to lots of training data from different sources can mitigate this problem, but this is still a challenge for data-driven methods, as it requires large training sets covering the full spectrum of potential noise profiles. Standard diffusion MRI pre-processing software is especially useful in this case to remove as much noise as possible and make the data more manageable for data-driven models. However, if data seen at test time is subject to a different noise distribution than what was seen while training, the model might fail where a classic tractography method would have worked.

Predetermined input dimensionality Varying parameters in diffusion MRI protocols, such as the number of gradient directions, their orientation, and the b-value (or multiple b-values) used to probe the tissue structures, can affect data-driven models that directly rely on the raw data as input. Given that ML models generally take a vector of information with fixed dimensionality as input, the number of gradient directions would be predetermined and could not be changed after training of such models. Changing the orientation of the gradients or changing the b-value after training would also change the input signal distribution, and is therefore generally not supported by machine learning models. That is to say, the scanning protocol should stay the same between training and testing.

Introducing an intermediate step of processing the raw diffusion signal, e.g. re-

3.4. RESULTS & DISCUSSION

sampling the diffusion signal [93, 11] or fitting spherical harmonics [101, 102, 160, 112], is on way to overcome the fixed number of gradient directions and their orientations, but it is still unknown how the model will be affected by changing the underlying measured signal. Furthermore, supporting different b-values (or even a varying number of shells) remains an open problem for models using a fixed-size input.

3.4.5 Proposed guidelines for a data-driven tractography evaluation framework

Considering the ML tractography evaluation issues previously underlined, we discuss in this section the fundamental elements of a better framework we believe the community should adopt in the upcoming years. We start with the essential characteristics such a framework should have, followed with useful features.

Essential characteristics First and foremost, an ideal data-driven tractography evaluation framework should come with a public and free-to-use dataset that anyone could easily rely on. The dataset should include images of real human acquisitions along with a careful expert selection of ground truth streamlines. It is important to avoid any bias towards a specific tractography algorithm. In order to achieve this, the streamlines could be first generated by a large number of different (and ideally orthogonal) deterministic, probabilistic and global algorithms and then segmented by expert annotators according to strict anatomical definitions for a given number of bundles. While such manual annotation would be tedious, time consuming and even error prone, we consider this an indispensable step towards building a realistic and useful dataset for ML-based development. The need for such a gold standard that quantifies human variability is well-known in other fields, such as automatic image segmentation, cell counting or in machine learning [70, 44, 15, 98]. Despite the fact that simulated brain images come with a pixel-accurate set of ground truth streamlines that can be generated in a matter of seconds, by definition synthetic diffusion signals are over-simplistic pictures of real data and, as such, cannot provide any guarantee of subsequent performance for data-driven methods on real data.

Although there is no consensus regarding the most desirable features a ML tractog-

3.4. RESULTS & DISCUSSION

raphy algorithm should have and how it should be evaluated, by its very nature, any ML evaluation framework should aim at measuring how an algorithm can faithfully reproduce a task it was trained for. As such, a reasonable dataset should include a sufficiently large number of well-separated training and testing images. Thus, statistics resulting from such a dataset would not suffer from contamination and the reported metrics would be reliable and unbiased estimates of the true generalization power of a ML algorithm. In addition, to ensure that the observed differences between multiple algorithms are resulting from the intrinsic properties of the model and not caused by some feature of the evaluation framework, the number of free variables should be reduced to a minimum. Consequently, the tracking masks and seeds should be provided together with clearly preprocessed diffusion data, so that the proposed methods can be evaluated in equal conditions. There should be multiple “classes” of input data, depending on whether an algorithm supports DWI samples, SH coefficients or fODF peaks. Furthermore, the initial diffusion signal should have the same statistical properties for the training and the testing set. Finally, the acquired images should ideally be acquired at different MRI scanners with different acquisition protocols in order to avoid overfitting issues.

Evaluation metrics should also be bound to the purpose of tractography algorithms. Considering that tractography is mostly used for bundle reconstruction, tractometry studies and connectivity analyses, an ideal evaluation framework should include two sets of metrics : 1) metrics measuring how a ML method can faithfully reproduce a set of predefined bundles it was trained to recover (tractometry), and 2) metrics measuring how it can connect matching regions of the brain, i.e. produce valid connections (connectivity). Furthermore, since many applications use tractography algorithms to produce a large number of streamlines (with many false positives), which are then filtered out by a post-processing algorithm such as RecoBundles [52], the framework should report results before and after post-processing. This would underline the true recall power of a data-driven algorithm, which is a fundamental characteristic of tract-based and connectivity-based applications [82].

Lastly, the size of an ideal dataset is of primary importance. While a small-sized dataset could be prone to overfitting, it would be costly to create a very large dataset and also difficult to ensure a coherent manual annotation. One rule of thumb that

3.4. RESULTS & DISCUSSION

can be used to identify the “correct” size of a dataset is through the inspection of the learning curve of several algorithms [10]. These curves show the model performance as a function of the training sample size. Typically, the performance of several models saturates for a sufficient dataset size. Although imperfect, this procedure is a good heuristic for estimating the size of the dataset.

Other useful features Despite any thorough manual annotation protocol, manually annotated bundles can be subject to non-negligible inter-rater and intra-rater variability. As such, a useful characteristic of a ML tractography dataset would be a measure of those variations. This would be obtained by having several experts annotating the dataset, and at least one expert annotating it twice or more times. Such measures would provide a minimal bound beyond which a data-driven algorithm could be considered “as good as an expert”. Another very useful tool would be an openly accessible online evaluation system. Given such a system, people could upload their test results in order to compare them with the test ground truth. In that way, an automatic ranking procedure similar to that of Kaggle could be used to sort various ML algorithms based on their achieved scores. While no ranking method is perfect, it would nonetheless provide a common evaluation framework that people could rely on.

An ideal dataset would also cover the whole field of diffusion MRI acquisition protocols, from HCP-like research acquisitions to clinical acquisitions. It would include single b-value as well as multiple b-values data, along with more sophisticated acquisition protocols such as b-tensor encoding. It would also need low resolution images together with high-resolution images. Since data harmonization is also a problem for data-driven algorithms, acquisition from several sites are needed for test-retest studies. Annotated pathological cases would complete the dataset by allowing careful preliminary studies on how ML-based methods can be relied on in unhealthy patients.

Finally, since tractography is used more and more in pre-clinical applications, a subset of manually annotated rodent or macaque brains would be of great interest to train and test future ML algorithms (like the *2018 VOTEM Challenge* [135], for example).

This is, of course, the ultimate wish list. But, in the era of open data and open

3.5. CONCLUSION

science, it needs to be done by the community, for the community. We can already see this work in progress with more and more accessible and reproducible data being published every year.

3.5 Conclusion

In this paper, we provided an exhaustive review of the current state of the art of machine learning methods in the field of tractography. We described the existing datasets that comprise both diffusion data and reference streamlines, which could generally be useful for new tracking methods based on ML. In particular, we thoroughly examined the widely used evaluation tool for data-driven tracking methods, the *2015 ISMRM Tractography Challenge*, and detailed flaws and shortcoming when used to assess data-driven algorithms. Based on our findings, we suggested good practices that we believe would foster the development of a new evaluation framework for ML-based tractography methods with the potential to effectively advance this field of research.

There is no doubt that machine learning tractography will have an important role to play in the future to solve some of the open problems of tractography. At the moment, however, all existing methods show theoretical potential and in limited test cases. Methods have yet to make solid demonstrations of their performance and efficiency in practice. There is still no ML-based tractography tool that is a scalable and usable on any given diffusion MRI dataset. This is true for healthy datasets but even more so for pathological brains. Hence, it is fair to say that ML-based tractography is still at its infancy and is not ready for “prime-time”, but is nonetheless a very fertile field of research to make meaningful contributions to the field of connectivity mapping.

Chapitre 4

TractoInferno : Une base de données pour la tractographie par apprentissage

Résumé

Les méthodes de tractographie par apprentissage publiées jusqu'à maintenant sont difficilement comparables pour plusieurs raisons, la principale étant que les données d'entraînement et de test varient d'une publication à l'autre. Pour corriger ce problème, la nouvelle base de données *TractoInferno* est un outil indispensable. Il s'agit d'une base de données variées, libre d'accès pour la communauté scientifique. Elle contient toutes les données prétraitées nécessaires pour entraîner et évaluer un algorithme de tractographie par apprentissage.

Contributions de la publication

- Construction de la plus grande base de données multisite pour la tractographie par apprentissage, contenant 284 sujets répartis

entre 6 sites d’acquisition.

- Publication libre d'accès de *TractoInferno* pour la communauté scientifique.
- Création d'un outil d'évaluation standardisé pour évaluer et comparer de futurs algorithmes.
- Production de résultats de référence pour des méthodes classiques et des méthodes basées sur les réseaux de neurones récurrents.

Contributions des auteurs

- Rassemblement des données originales (Arnaud Boré, Guillaume Theaud, Louis de Beaumont, Samuel Guay, et Philippe Poulin)
- Traitement des données (Guillaume Theaud et Philippe Poulin)
- Révision des données traitées (Guillaume Theaud, Maxime Descoteaux et Philippe Poulin)
- Support pour l'exécution de la méthode SET (Étienne St-Onge)
- Écriture du manuscrit (Philippe Poulin)
- Révision du manuscrit (Pierre-Marc Jodoin et Emmanuelle Renaud)

Commentaires

TractoInferno le résultat d'un travail acharné à rassembler, traiter et réviser des données afin de construire une base de données sans précédent. Il s'agit d'un effort considérable, autant au niveau humain que technique, qui a pour but d'avoir un impact majeur dans le domaine de la tractographie. Cet article a été soumis à Nature Scientific Data en décembre 2021.

TractoInferno: A large-scale, open-source, multi-site database for machine learning dMRI tractography

Philippe Poulin^{1,2}, Guillaume Theaud¹, François Rheault¹, Étienne St-Onge¹, Arnaud Boré¹, Emmanuelle Renauld¹, Louis de Beaumont^{3,4}, Samuel Guay^{3,4}, Pierre-Marc Jodoin², Maxime Descoteaux¹

¹ Sherbrooke Connectivity Imaging Laboratory (SCIL), Computer Science Department, Université de Sherbrooke, Sherbrooke, Canada

² Videos and Images Theory and Analytics Laboratory (VITAL), Université de Sherbrooke, Sherbrooke, Canada

³ Montreal Sacred-Heart Hospital Research Centre, Montréal, QC, Canada

⁴ Department of Surgery, Université de Montréal, Montréal, QC, Canada

Abstract

TractoInferno is the world's largest open-source multi-site tractography database, including both research- and clinical-like human acquisitions, aimed specifically at machine learning tractography approaches and related ML algorithms. It provides 284 datasets acquired from 3T scanners across 6 different sites. Available data includes T1-weighted images, single-shell diffusion MRI (dMRI) acquisitions, spherical harmonics fitted to the dMRI signal, fiber ODFs, and reference streamlines for 30 delineated bundles generated using 4 tractography algorithms, as well as masks needed to run tractography algorithms. Manual quality control was additionally performed at multiple steps of the pipeline. We showcase *TractoInferno* by benchmarking the learn2track algorithm and 5 variations of the same recurrent neural network architecture. Creating the *TractoInferno* database required approximately 20,000 CPU-hours of processing power, 200 man-hours of manual QC, 3,000 GPU-hours of training baseline models, and 4 Tb of storage, to produce a final database of 350 Gb. By providing a standardized training dataset and evaluation protocol, *TractoInferno* is an excellent tool to address common issues in machine learning tractography.

4.1. INTRODUCTION

4.1 Introduction

Tractography is the computerized process of reconstructing brain white matter fibers from diffusion MRI (dMRI) data. It usually consists of four steps : i) pre-processing diffusion-weighted images (DWI) (e.g. denoising, eddy, motion, susceptibility corrections), ii) estimating local fiber directions, iii) reconstructing white matter pathways (i.e. tractography), and iv) delineating bundles (a group of similar streamlines connecting the same brain regions) [47, 82].

Current “traditional” tractography approaches (deterministic and probabilistic) mostly rely on making local point-wise decisions in the fiber ODF field, iterating until termination [120, 122]. Global methods have also been proposed [111, 84, 63, 79], but Rheault et al. mentions that “[...] global tractography methods ultimately rely on local information patched together” and “even global tractography algorithms struggle to correctly assemble a streamline” [117]. Tractogram filtering [30, 125, 126, 77] is also a popular post-processing method used to remove streamlines that do not fit anatomical constraints (such as explaining the underlying signal), but requires an over-complete tractogram as it does not create new streamlines, thus effectively “wasting” computing power. Finally, streamline clustering [49, 124] can be used to group streamlines based on similarity and remove possible outliers, but it suffers from the same drawback as tractogram filtering, as it requires an over-complete tractogram.

These approaches mostly rely on mathematical models or anatomical priors, and do not require histological ground truth to work. However, this is an issue for machine learning algorithms, where the training dataset is an integral part of the resulting model [103]. Machine learning methods need reference streamlines to train on. Unfortunately, on real datasets, streamlines can only be generated by traditional [and yet non-machine learning] tractography methods, which are imperfect by their very nature [82]. This is an issue for testing if the predictions made by these methods are reliable or not. Luckily, by combining streamlines (both true positives and false positives) generated by several tractography algorithms and using filtering and clustering to remove as much false positives as possible, it is possible to establish a *gold standard* reference dataset. Even without a histologically accurate ground truth, it would be desirable to have algorithms that can reproduce a gold standard reference

4.1. INTRODUCTION

while generating as little false positive streamlines as possible.

In the recent years, machine learning algorithms have been proposed to improve the tractography process by some combination of i) taking advantage of the full diffusion information or other modalities, ii) generating more reliable streamlines using a reference teacher dataset, or iii) integrating more spatial context to guide the tracking process (either neighbourhood or path information) [103, 93, 101, 156, 159]. For example, *TractSeg* [156] is a method that first identifies the volume of reference of a specific white matter bundle, and then generates a bundle-specific tractogram by running a traditional tractography algorithm inside the bundle mask only. To do so, convolutional neural networks [119] learn to map the diffusion volume to multiple binary bundle segmentation maps. *LearnToTrack* [101] and *DeepTract* [11] propose to use information along a streamline path to guide its generation process (instead of making point-wise decisions) using Recurrent Neural Networks [21, 60]. *Entrack* [159] proposes a Neural Network with a fixed context of 4 streamline steps, and models a probabilistic streamline direction using a von Mises-Fisher distribution trained with entropy regularization.

Unfortunately, these machine learning methods train and evaluate their models on different dataset which makes it difficult to compare their true generalization capabilities [103]. It is often a combination of the ISMRM2015 Tractography Challenge [82] and some subjects from the HCP Young Adults database [54]. Additionally, data pre-processing might vary between proposed methods, and different algorithms and protocols are used to generate the reference tracts. Finally, evaluating the generalizability of a model is almost impossible without diverse (aka multi-site) training and test sets. As a result, all those discrepancies in methodology make it very challenging to assess the reliability of a single approach, and make it almost impossible to fairly compare approaches.

We propose to address this problem by building *TractoInferno*: the largest publicly available, multi-site, dMRI and tractography database, which provides a new baseline for training and evaluating machine learning tractography methods. It provides 284 datasets acquired from 3T scanners across 6 different sites. *TractoInferno* includes T1-weighted images, single-shell diffusion MRI (dMRI) acquisitions, spherical harmonics fitted to the dMRI signal, fiber ODFs, and reference streamlines for 30 delineated

4.2. DATASETS

bundles generated by combining 4 different tractography algorithms, as well as masks needed to run tractography algorithms.

We use *TractoInferno* to benchmark the 4 tractography algorithms used to create the reference tractograms, along with the learn2track [101] algorithm and 5 variations of the same recurrent neural network architecture, inspired in part by the models of (Benou et Riklin Raviv) and (Wegmayr et Buhmann) [11, 159].

Creating the *TractoInferno* database required approximately 20,000 CPU-hours of processing power, 200 man-hours of manual QC, 3,000 GPU-hours of training baseline models, and 4 Tb of storage, to produce a final database of 350 Gb.

TractoInferno is a dataset intended to promote the development of machine learning tractography algorithms, which generally suffer from multiple issues, such as limited datasets or inconsistent training data. Its large-scale and multi-site aspect is an undeniable benefit to best evaluate the generalization capabilities of new ML algorithms. We consider *TractoInferno* to be by far the best available tool for training, evaluating, and comparing future machine learning tractography algorithms.

4.2 Datasets

The proposed dataset is made of a combination of six dMRI databases, either publicly available and free to redistribute or acquired through open-access data sharing agreements. Databases were chosen with the explicit goal of having a diversity of scanner manufacturers, models, and protocols. We chose to fix certain parameters for uniformity, such as having only 3T scanners, and b-values of around 1000 s/mm², as we don't know how they could affect machine learning models. The focus is effectively on assessing the reliability of algorithms under different scanner manufacturers and acquisition protocols. We obtained an initial number of data from 354 subjects, with the original metadata described in Table 4.1.

4.2.1 Mazoyer et. al - BIL & GIN

We retained 39 subjects from the BIL&GIN database [86], acquired on a 3T Philips Achieva, with the following dMRI protocol: TR = 8500 ms, TE = 81 ms, angle =

4.2. DATASETS

Name	<i>Mazoyer et. al [86]</i>	<i>Tsushida et. al [143]</i>	<i>DeLuca et. al [33]</i>	<i>Poldrack et. al [100]</i>	<i>Tamm et. al [133]</i>	<i>Tremblay et. al [142]</i>
Scanner	3T Philips Achieva	3T Siemens Prisma	3T Siemens Prisma	3T Siemens Trio	3T GE Discovery MR750	3T Siemens Magnetom TIM Trio
# subjects	39	20	64	130	86	15
Age avg	28.1	21.4	31.9	31.3	N/A	58.1
Age std	7.3	1.7	7.6	8.7	N/A	5.3
F/M	0/39	10/10	49/15	62/68	44/42	0/15
L/R	8/31	N/A	0/64	N/A	N/A	3/12
Resolution	2	1.75	2	2	2.3	2
b-value	1000	1000	1000	1000	1000	700
TR	8500	3540	1800	9000	7000	9200
TE	81	75	70	93	81	84
Nb dirs	21*	32	128**	64	45	30

Table 4.1 – Original datasets metadata. Not all metadata information was available from the original datasets. Missing metadata is reported as {N/A}. Resolution is in mm³ isotropic. b-value is in s/mm². TR and TE are in ms.

*: 21 directions acquired twice by reversing the gradient polarity, then averaged over another identical acquisition (total of 84 DWI volumes).

**: 64 directions acquired twice, not averaged.

90°, SENSE reduction factor = 2.5, FOV 224 mm, acquisition matrix 112 × 112, 2 mm³ isotropic voxel.

The dMRI acquisition consisted of 21 gradient directions at $b = 1000$ s/mm², acquired twice by reversing the polarity, and then repeated twice for a total of 84 DWI images, averaged down to a single volume with 21 directions. A single $b = 0$ s/mm² was also acquired alongside the DWI images. Subjects were all males, with age mean/std of 28.1 +- 7.3 (Min: 20, Max: 57). 8 subjects were left-handed and 31 right-handed.

4.2.2 Tsushida et. al - MRi-Share

We obtained 20 subjects from the MRi-Share database [143], acquired on a 3T Siemens Prisma, with a dMRI protocol designed to emulate the UKBioBank project [16], specifically: TR = 3540 ms, TE = 75 ms, 1.75 mm³ isotropic voxel.

We selected the $b = 1000$ s/mm² DWI images only, consisting of 32 gradient directions, and 3 provided $b = 0$ s/mm² images. Subjects were composed of 10

4.2. DATASETS

females, 10 males, with age mean/std of 21.4 ± 1.7 . Minimum/maximum age and handed-ness metadata were not available.

4.2.3 DeLuca et. al - Bilingualism and the brain

We have 64 subjects from the *Bilingualism and the Brain* database [33, 32], acquired on a 3T Siemens Prisma, with the following dMRI protocol: Echo planar imaging, TR = 1800 ms, TE = 70 ms, acquisition matrix 256 x 256, 2 mm³ isotropic voxel.

The dMRI acquisition consisted of 64 gradient directions at $b = 1000$ s/mm², acquired twice, and 4 $b = 0$ s/mm² images. Subjects were composed of 49 females and 15 males, with age mean/std of 31.9 ± 7.6 (Min: 18, Max: 52). All subjects were right-handed.

4.2.4 Poldrack et. al - UCLA CNP

We got 130 healthy subjects from the *UCLA Consortium for Neuropsychiatric Phenomics LA5c Study* [100], acquired on a 3T Siemens Trio, with the following dMRI protocol: echo planar imaging, TR = 9000 ms, TE = 93 ms, acquisition matrix 93 x 93, 90 degree flip angle, 2 mm³ isotropic voxel. DWI were corrected for eddy currents and head motion using the b0 images as reference.

The dMRI acquisition consisted of 64 gradient directions at $b = 1000$ s/mm², and 1 $b = 0$ s/mm² image. Subjects consisted of 62 females and 68 males, with age mean/std of 31.3 ± 8.7 (Min: 21, Max: 50). Handed-ness metadata was not available.

4.2.5 Tamm et. al - The Stockholm Sleepy Brain Study

We retained 86 subjects from the Stockholm Sleepy Brain Study database [133, 95], acquired on a 3T GE Discovery MR750, with the following dMRI protocol: Echo planar imaging, TR = 7000 ms, TE = 81 ms, 2.3 mm³ isotropic voxel.

The dMRI acquisition consisted of 45 gradient directions at $b = 1000$ s/mm², along with 5 $b = 0$ s/mm² images. Subjects were composed of 44 females and 42

4.3. METHODOLOGY

males, with 47 subjects in the [20-30] years old bracket and 39 subjects in the [65-75] years old bracket. Handedness was not available.

4.2.6 Tremblay et. al - mTBI and Aging study (controls)

We obtained 15 subjects from the mTBI and Aging Study [142], all controls from the “remote” group. they were acquired on a 3T Siemens Magnetom TIM Trio, with the following dMRI protocol: TR = 9200 ms, TE = 84 ms, 2 mm³ isotropic voxel.

The dMRI acquisition consisted of 30 gradient directions at $b = 700 \text{ s/mm}^2$. along with 1 $b = 0 \text{ s/mm}^2$ image. Subjects were all males, with age mean/std of 58.1 +- 5.3 (Min: 52, Max: 67). 3 subjects were left-handed and 12 were right-handed.

4.3 Methodology

We processed the original acquisition volumes of the 354 aforementioned subjects with the same pipeline to offer a uniform database of dMRI images, derivatives, and bundle tractograms. First, all original DWI went through a manual quality control (QC) step to remove any obvious errors prior to the processing pipeline. In this case, QC is done by a thorough visual inspection of all modalities, along with a spherical representation of the acquisition scheme. Then, the *TractoFlow* pipeline was run to process the data and compute necessary derivatives [134, 73, 38]. Another QC step was executed afterwards, to remove images with artifacts that could not be corrected automatically. Next, ensemble tractography was performed using four different algorithms to extract a diverse set of streamlines: deterministic tractography [7], probabilistic tractography [141], Particle-Filtered Tractography [53] and Surface-Enhanced Tractography [129]. RecoBundlesX (RBX) was used subsequently to perform bundle extraction on the whole-brain tractograms, using the default suggested bundle models [52, 118]. A final manual QC step was performed to examine the extracted bundles, and remove anything that contained obvious mistakes, or did not meet our criteria for bundle extraction. All manual quality control steps were done using dmriqc (https://github.com/scilus/dmriqc_flow). Figure 4.1 shows the processing steps of *TractoInferno*.

4.3. METHODOLOGY

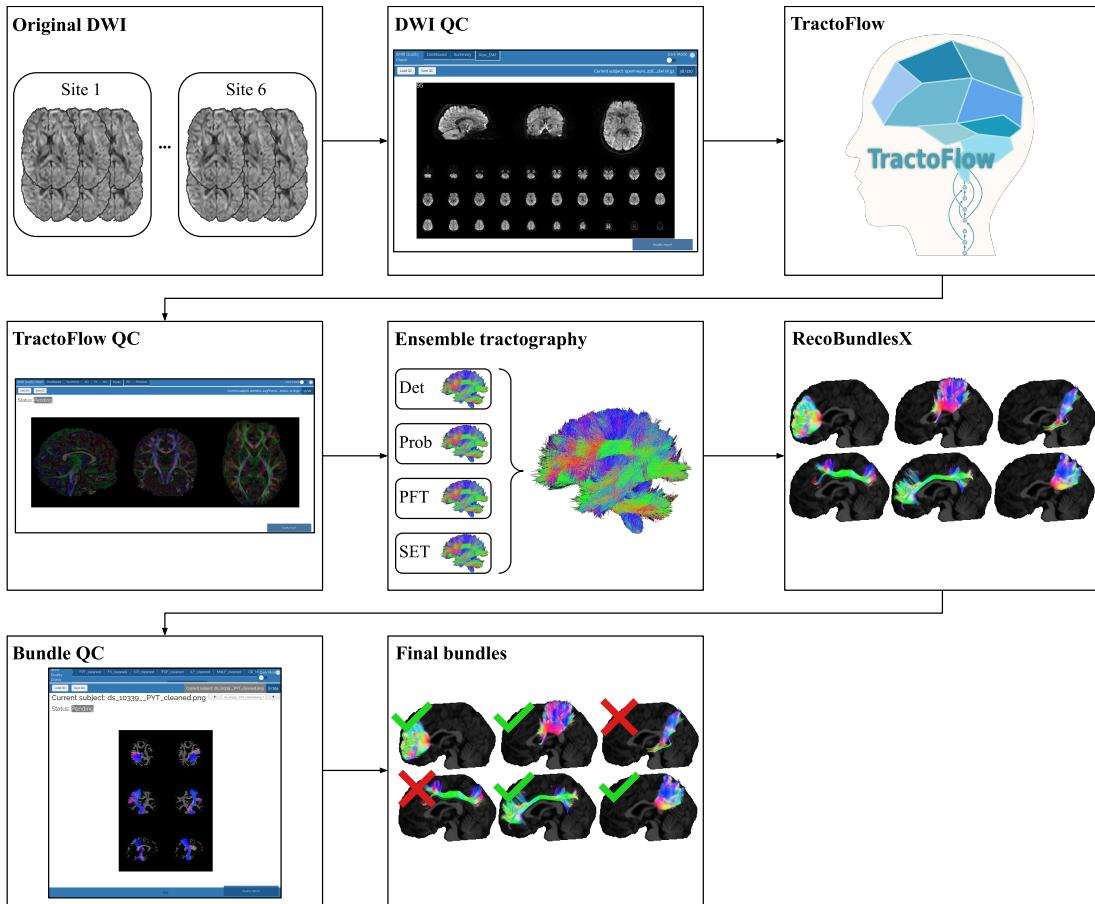


Figure 4.1 – *TractoInferno* processing pipeline, from original DWI images to final bundles.

From the initial 354 volumes, after all the processing steps and quality control, we were left with 284 volumes and associated bundles. The final volumes were split into training, validation and test sets with a 70%/20%/10% split for reproducibility across future experiments. The specific commands for the whole pipeline are available in [4.A](#). For a final dataset size of 350Gb, we needed approximately 20,000 CPU-hours of processing time (using a cluster of nodes, each with 40 cores across 2 Intel Gold 6148 Skylake CPUs at 2.4 GHz), 200 man-hours of manual QC, and 4 Tb of storage. The models benchmarked in section [4.5](#) also required an additional 3,000 GPU-hours (using NVidia V100SXM2 GPUs with 16Gb VRAM) for training and generating candidate tractograms. In the next sub-sections, we detail the TractoInferno processing

4.3. METHODOLOGY

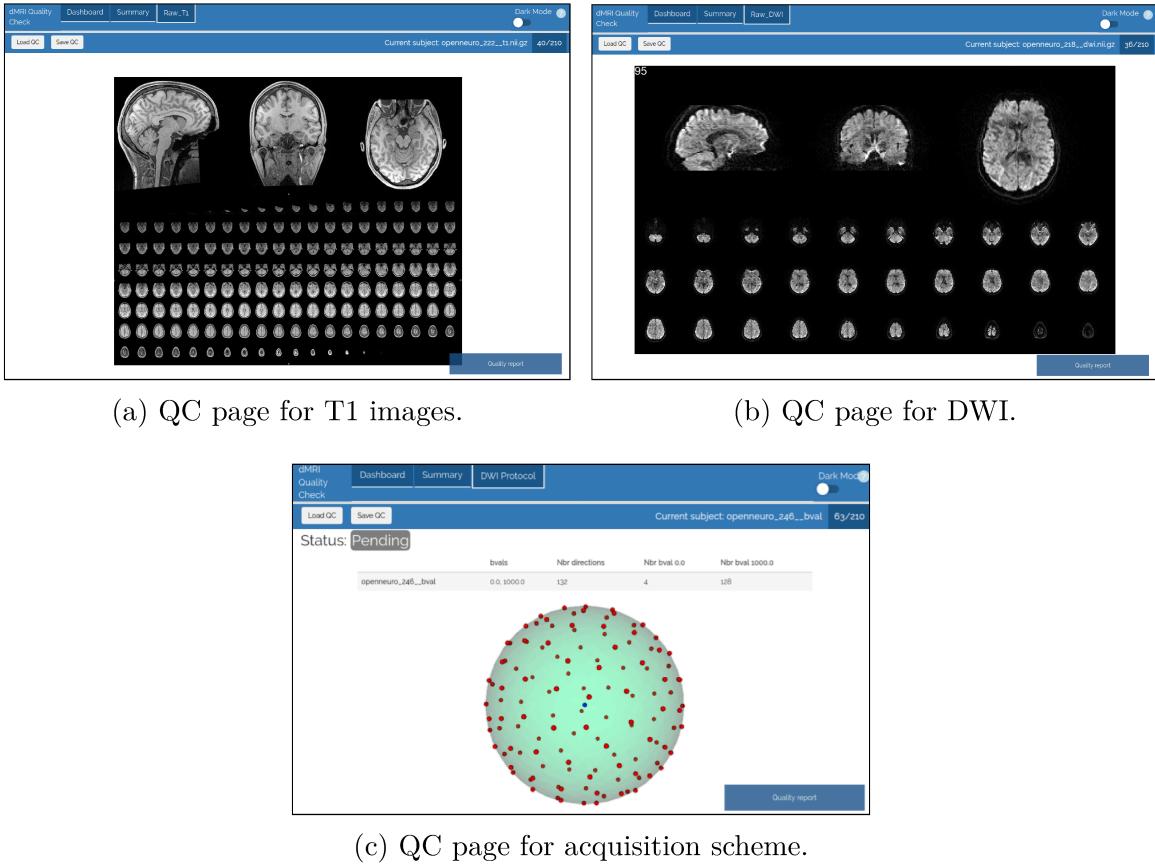


Figure 4.2 – Examples of HTML pages generated by *dmriqcipy* for data QC. (a) 3 slices of the T1 image (one for each axis), plus a mosaic of multiple axial slices. (b) 3 GIFs of the dMRI (one slice in each axis), plus a mosaic of multiple axial slices; (c) The gradient directions represented on a sphere.

steps.

4.3.1 Raw data QC

We used *dmriqcipy* to generate QC reports. These reports are in HTML format so it is easily assessed and annotated by multiple people. The raw data reports contain multiple tabs with complementary information, as shown in Figure 4.2. Three different raters went through the QC reports and individually rated every acquisition with a “score” (either *pass*, *fail*, or *warning*) and comment if necessary. Specifically, failure cases included the presence of visual artifacts (e.g. missing slices, low signal-to-noise

4.3. METHODOLOGY

ratio, corrupted data, high spatial distortion) and other artifacts harder to identify (such as a “broken” gradient acquisition scheme). Afterwards, all subjects tagged as “fail” were removed, and considered as impossible to repair with our available tools. All subjects tagged as “pass” or “warning” were passed on for TractoFlow, the next step in the pipeline. Subjects tagged as “warning” were re-examined after the TractoFlow processing to examine if any issues remained, or if they were compensated for by the pipeline.

4.3.2 TractoFlow pipeline

We used TractoFlow 2.1.1 [134] to process the raw DWI. To make sure that every processing step was traceable and reproducible, a Singularity [73] image was used along with the Nextflow pipeline [38]. Note however that some results may not be 100% reproducible due to the uncertain nature of registration, parallel processing, and floating point precision. We ran the full pipeline except for the *Topup* process, as not all reverse b0 images were available [3]. Specifically, the pipeline executed the following steps:

- DWI brain extraction [128], denoising [153], eddy current correction [2], N4 bias field correction [146], cropping, normalization [108, 88], and resampling [42];
- T1 denoising [24], N4 bias field correction [146], registration [4] and tissue segmentation [170] maps for Particle-Filtered Tractography [53, 127];
- DTI fitting and metrics extraction [50];
- fODF fitting using constrained spherical deconvolution [36, 139, 51], with a fiber response function fixed manually to [0.0015, 0.0004, 0.0004].

4.3.3 TractoFlow results QC

Outputs from TractoFlow went through a manual QC pass to identify failure cases. Using *dmriqcipy*, we were able to easily and quickly look at all maps derived from DTI and fODF metrics, along with T1 registration overlay. For example, RGB maps extracted from DTI metrics allowed us to quickly identify if tensor peaks were well-aligned or if a flip was needed, and T1 registration overlays showed whether too

4.3. METHODOLOGY

much deformation was present.

4.3.4 Ensemble tractography

Using a single tractography method as reference for a machine learning algorithm might induce unwanted biases. To avoid this, we chose to use ensemble tractography by combining 4 different algorithms to generate reference streamlines, namely deterministic [7], probabilistic [141], particle-filtered [53], and surface-enhanced [129] tractography. We fixed the tracking parameters to the standard default values:

- WM + WM/GM interface seeding
- 10 seeds per voxel (Det, Prob, PFT) or 10,000,000 surface seeds (SET)
- Step size 0.2mm (Det, Prob, SET) or 0.5mm (PFT)
- WM tracking mask (Det, Prob) or WM/GM/CSF probability maps (PFT, SET)

After tracking, we used streamline compression [106, 114, 62] in order to save space, which means that streamlines have a variable step size that need to be taken into account by ML tractography algorithms. We detail each algorithm in the following three subsections.

Deterministic tracking

Deterministic tracking [7] chooses the fODF peak most aligned with the previous direction as the next streamline step. It seems better suited to connectomics studies [120], mainly on account of the low number of false positives it produces. While it may be inadequate for spatial exploration and bundle reconstruction, deterministic tracking essentially produces smooth streamlines that follow the easiest path through the fODF field. Smooth streamlines are likely more desirable for machine learning algorithms rather than chaotic streamlines that often change directions locally.

Probabilistic tracking and Particle-Filtered Tractography

Probabilistic tracking [141] samples a new streamline direction inside a cone of evaluation aligned with the previous direction, with a probability distribution proportional to the shape of the fiber ODF within the cone.

4.3. METHODOLOGY

Particle-Filtered Tractography [53] is an improvement over probabilistic tracking. It takes as input probability maps for streamline continuation/stopping criteria, and allows to “go back” a few steps when a streamline terminates in a region not included in the “termination-allowed” map.

Both algorithms are better suited for spatial exploration, at the cost of producing much more false positives. They are especially effective for bundle reconstruction, in which case there are anatomical priors about both the endpoints that should be connected and the pathway that should be followed by the bundle.

Surface-Enhanced Tracking

Finally, Surface-Enhanced Tracking [129] is a state-of-the-art tractography algorithm that relies on initializing streamlines in an anatomically plausible way at the cortex, then running a PFT tracking algorithm. Indeed, gyri have been shown to be problematic regions for tractography, where low dMRI resolution can lead to a gyral bias in streamline terminations [113].

To this end, we computed the WM-GM boundary surface from the T1w image using the *CIVET* [69] tool and the CBRAIN [123] platform. Then, SET uses a geometric flow method, based on surface orthogonality, to reconstruct the fanning structure of the superficial white matter streamlines. The output of this flow is used to initialize and terminate a PFT tractography algorithm. The result is a tractogram with improved cortex coverage, improved fanning structure in gyri, and reduced gyral bias.

4.3.5 Bundle segmentation with RBX

We used RBX [52, 118] to automatically extract WM bundles. The algorithm works by matching streamlines to an atlas of reference bundles. First, a quick registration step brings the atlas into native space using the atlas FA image. Then, a whole-brain tractogram is compared against the bundles atlas using multiple sets of parameters to extract a fixed set of bundles, listed in Table 4.2. Finally, a majority voting step (label fusion) extracts the final streamlines for each bundle.

The whole pipeline was run using a Singularity container [73] and Nextflow [38]

4.3. METHODOLOGY

AC	Anterior commissure
AF	Arcuate fasciculus
CC_Fr_1	Corpus callosum, Frontal lobe (most anterior part)
CC_Fr_2	Corpus callosum, Frontal lobe (most posterior part)
CC_Oc	Corpus callosum, Occipital lobe
CC_Pa	Corpus callosum, Parietal lobe
CC_Pr_Po	Corpus callosum, Pre/Post central gyri
CC_Te	Corpus callosum, Temporal lobe
CG	Cingulum
FAT	Frontal aslant tract
FPT	Fronto-pontine tract
FX	Fornix
ICP	Inferior cerebellar peduncle
IFOF	Inferior fronto-occipital fasciculus
ILF	Inferior longitudinal fasciculus
MCP	Middle cerebellar peduncle
MdLF	Middle longitudinal fascicle
OR_ML	Optic radiation and Meyer's loop
PC	Posterior commissure
POPT	parieto-occipito pontine tract
PYT	Pyramidal tract
SCP	Superior cerebellar peduncle
SLF	Superior longitudinal fasciculus
UF	Uncinate fasciculus

Table 4.2 – List of bundles in the default RBX atlas.

for reproducibility. It is freely available online (https://github.com/scilus/rbx_flow/), along with a suggested bundles atlas (<https://zenodo.org/record/4630660#.YJvmwXVKhdU>).

4.3.6 Bundle segmentation QC

Automated pre-QC

To facilitate the QC procedure, we ran a pre-QC analysis to automatically rate bundles according to pre-defined criteria before manual inspection. These criteria are detailed in Table 4.3. Afterwards, all bundles were looked at manually through an

4.3. METHODOLOGY

easier procedure that consists in confirming an already assigned rating rather than rating from scratch.

Rating	Criteria
Fail	$x < 50$
Warning	$x == 0$ in either hemisphere (if symmetric bundle).
Pass	$x \notin [\mu - 1.5\sigma, \mu + 3.5\sigma]$

Table 4.3 – Automatic rating criteria, in order of priority.

x is the number of streamlines of the bundle of interest;

μ and σ are the average and the standard deviation, respectively, of the number of streamlines for the bundle of interest, across all subjects;

Manual quality control using *dmriqcipy*

A bundle was removed if it looked visually incomplete or if it deviated from the expected pathway. A poor bundle reconstruction might have an algorithmic cause, such as sub-optimal tracking parameters or improper registration in RBX. It might also have an anatomical cause, such as unknown or undisclosed neurological conditions. Furthermore, visually evaluating a bundle reconstruction is very subjective, and a rater's evaluation can be affected by the time of day, duration of QC, or even the angle of visualization in the QC tool [116]. For all those reasons, and with the goal of establishing a gold standard for machine learning tractography methods, we chose to be somewhat severe in the rating of bundles, in order to minimize the number of false positives, even if that meant missing out some true positive data. After QC, we chose to ignore the following bundles from the atlas due to generalized reconstruction errors : AC, CC_Te, Fx, ICP, PC, SCP. From the initial 354 volumes, after all the processing steps and quality control, we were left with 284 volumes and associated bundles.

4.4. EVALUATION PIPELINE FOR CANDIDATE TRACTOGRAMS

4.4 Evaluation pipeline for candidate tractograms

When evaluating machine learning tractography algorithms, we focus on the volume covered by the recognized bundles (compared to the gold standard bundles). We make no assumptions about the ability to “explore” the brain outside the scope of the *TractoInferno* dataset. Consequently, we ignore anything that is not recognized as a candidate bundle, and do not try to categorize streamlines as valid or invalid connections.

Candidate bundles are extracted in the same way that we defined the gold standard bundles. First, we run RBX to extract candidate bundles from the candidate whole-brain tractogram. Candidate bundles are then converted to binary volume coverage masks. Finally, each candidate mask is compared against its corresponding gold standard bundle mask to compute evaluation metrics.

For each subject in the testset, and for each available bundle of the given subject, we extract the following evaluation metrics: Dice score, overlap and overreach. The scores are averaged over all subjects of the testset to provide final scores. Altogether, these metrics help better understand the performance of a candidate tractography algorithm.

The evaluation pipeline is available online (<https://github.com/scil-vital/TractoInferno/>) and should be used with the provided *TractoInferno* testset, along with the default RBX-flow models.

4.5 RNN-based tractography

To gauge the performances of ML models trained on the *TractoInferno* dataset, we implemented an RNN model and the necessary framework to train it on a large-scale tractography database, which was used multiple times in published papers in the last few years, such as *Learn2Track* [101], *DeepTract* [11], and *Entrack* [159]. Using the base implementation, we can easily modify the last layer of the model and its loss function to mimic the mentioned RNN models, and a few more.

We chose the stacked Long Short-Term Memory (LSTM) network as the recurrent building block for conditional streamline prediction. The LSTM is a type of RNN

4.5. RNN-BASED TRACTOGRAPHY

designed specifically to handle long-term dependencies, with the ability to deal with exploding and vanishing gradient problems [60].

4.5.1 Learn2track

Learn2track[101] proposed an RNN model for tractography, where the output of the model at each timestep is a 3D vector, used as the next direction of the streamline. The predicted vector is then scaled to the chosen step size, in order to match the lengths of the target and prediction.

From the same idea, we implemented an LSTM for deterministic tractography. As in the original *learn2track* paper, we used the squared error loss function between the target and prediction. The loss for a single streamline S composed of T steps is the following squared error:

$$\mathcal{L}(S) = - \sum_{t=1}^T \| \mathbf{d}_t - \hat{\mathbf{d}}_t \|^2$$

where \mathbf{d}_t and $\hat{\mathbf{d}}_t$ are the target and predicted directions. This model is noted as *Det-SE*.

However, to accurately reflect that only the direction of the predicted vector is important (not the magnitude), we also performed an experiment where we minimized the negative cosine similarity between the target and predicted directions:

$$\mathcal{L}(S) = - \sum_{t=1}^T \cos(\theta_t) = - \sum_{t=1}^T \frac{\mathbf{d}_t \cdot \hat{\mathbf{d}}_t}{\|\mathbf{d}_t\| \|\hat{\mathbf{d}}_t\|}$$

where θ_t is the angle between \mathbf{d}_t and $\hat{\mathbf{d}}_t$. This model is noted as *Det-Cosine*.

4.5.2 DeepTract

In the same spirit as *learn2track*, *DeepTract* [11] is a recurrent model for probabilistic tractography. In this case, the model output is a distribution over classes, where each class corresponds to a direction on the unit sphere, i.e. a discrete conditional fiber ODF.

4.5. RNN-BASED TRACTOGRAPHY

As in the original paper, we implemented a cross-entropy loss function:

$$\mathcal{L}(S) = - \sum_{t=1}^T \sum_{m=1}^M y_{tm} \log(\hat{y}_{tm})$$

where M is the number of classes, and y_t and \hat{y}_t are vectors of target and predicted class probabilities. Note that we did not use label smoothing as in the original paper, nor entropy-based tracking termination. This model is noted as *Prob-Sphere*.

4.5.3 Entrack

Entrack[159] is a non-recurrent artificial neural network for probabilistic tractography. The model is instead a feed-forward neural network, but includes the previous streamline direction as prior information to guide the tracking process. The model outputs the parameters for a von Mises-Fisher distribution, i.e. a 3D unit-length vector for the mean, and a scalar concentration parameter. The distribution is analogous to a Gaussian distribution, but defined on the unit sphere instead of euclidean space.

We chose to apply the same general idea, using a recurrent network that predicts the parameters for a von Mises-Fisher distribution on a 3D sphere. We used the negative log-likelihood of the von Mises-Fisher distribution as the loss function:

$$\mathcal{L}(S) = - \sum_{t=1}^T \log[C(\hat{\kappa}_t) \exp(\hat{\kappa}_t \hat{\mu}_t^\top \mathbf{d}_t)]$$

where the predicted parameters of the distribution are $\hat{\mu}_t$ (a unit-length vector) and $\hat{\kappa}_t$ (a scalar concentration parameter), and \mathbf{d}_t is the target unit-length vector at step t . $C(\hat{\kappa}_t)$ abbreviates the normalization constant associated with the distribution, defined as following in the 3-dimensional case:

$$C_3(\kappa) = \frac{\kappa}{2\pi(e^\kappa - e^{-\kappa})}$$

Note that unlike the original method, we didn't use an entropy maximization scheme to regularize the predicted distribution. This implementation is noted as *Prob-vMF*.

4.5. RNN-BASED TRACTOGRAPHY

4.5.4 Gaussian distribution output

Following *Entrack* and the idea of predicting the parameters of a continuous probability distribution, we implemented another model, using a multivariate Gaussian distribution instead of a von Mises-Fisher distribution. This model outputs a 3D vector for the mean, and 3 scalars for the variance, (one in each dimension). We choose to use a diagonal covariance matrix, for stability, and do not output any values for covariance.

In the 3-dimensional case, the negative log-likelihood loss function is:

$$\mathcal{L}(S) = -\sum_{t=1}^T \log\left[\frac{1}{\sqrt{(2\pi)^3 |\hat{\Sigma}_t|}} \exp\left(-\frac{1}{2}(\mathbf{d}_t - \hat{\mu}_t)^\top \hat{\Sigma}_t^{-1} (\mathbf{d}_t - \hat{\mu}_t)\right)\right]$$

where $\hat{\Sigma}_t = \begin{bmatrix} \sigma_{xt}^2 & 0 & 0 \\ 0 & \sigma_{yt}^2 & 0 \\ 0 & 0 & \sigma_{zt}^2 \end{bmatrix}$ is the predicted diagonal covariance matrix at streamline step t . This model is noted as *Prob-Gaussian*.

4.5.5 Gaussian mixture distribution output

The previous Gaussian model outputs a single average direction which is appropriated in most cases. However, there may be cases of bundle fanning or forking where the single-mode assumption may be an issue. This is because the Gaussian probability density can only be spread over a large area.

As such, some regions may be better modelled with more than one location of higher density. To this end, we implemented a mixture density network[14] using a mixture of 3 Gaussian distributions. For each Gaussian, the model outputs 1 mixture weight, a 3D vector for the mean, and 3 scalars for the variances (again, we fix the covariances to zero).

In the 3-dimensional case, using a mixture of 3 Gaussians, the negative log-

4.5. RNN-BASED TRACTOGRAPHY

likelihood loss function is:

$$\begin{aligned}\mathcal{L}(S) &= -\sum_{t=1}^T \log \left[\sum_{k=1}^3 \phi_{kt} \mathcal{N}(d_t | \hat{\mu}_{kt}, \hat{\Sigma}_{kt}) \right] \\ &= -\sum_{t=1}^T \log \left[\sum_{k=1}^3 \phi_{kt} \frac{1}{\sqrt{(2\pi)^3 |\hat{\Sigma}_{kt}|}} \exp\left(-\frac{1}{2}(\mathbf{d}_t - \hat{\mu}_{kt})^\top \hat{\Sigma}_{kt}^{-1} (\mathbf{d}_t - \hat{\mu}_{kt}))\right) \right]\end{aligned}$$

where k denotes the number of Gaussians in the mixture, and ϕ_{kt} is the mixture parameter for the Gaussian k at streamline step t . This model is noted as *Prob-Mixture*.

4.5.6 Implementation details

All models were composed of 5 hidden layers of 500 units, used dropout with a rate of 0.1, and a batch size of 50 000 streamline steps. We added skip connections from the input layer to all hidden layers, and from all hidden layers to the output layer, inspired by [57]. We applied layer normalization [78] between all hidden layers, in order to stabilize the hidden state dynamics in recurrent neural networks. We used the Adam optimizer with the default parameters.

For all experiments, we used the spherical harmonics (SH) coefficients of order 6 fitted to the TractoFlow-processed DWI signal as the input signal, without any other pre-processing. In all cases, the models were trained using the exact same training and validation datasets, with streamlines resampled to a fixed step size of 1.0 mm. To help guide the model, we also included as input the diffusion signal in a neighbourhood of 6 directions (two for each axis, positive and negative) at a distance of 1.2 mm.

All models were trained for a maximum of 30 epochs (corresponding to around 2 weeks of training time on a 16Gb NVidia V100SXM2), but early stopping was used to stop training when the loss has not improved after 5 epochs. Each epoch was capped to 10 000 updates, as the sheer size of the dataset would otherwise require multiple days of training for a single epoch.

4.6. RESULTS & DISCUSSION

4.6 Results & Discussion

Machine learning models were trained using the *TractoInferno* database, with a training set of 198 volumes and a validation set of 58 volumes. We report in Table 4.4 the results of the *TractoInferno* evaluation pipeline on the testing set of 28 volumes. Results include each individual tractography algorithm used to build the reference bundles, and every model detailed in Section 4.5 after the training procedure.

Of all the base algorithms used to build the reference tractograms, PFT performed the best in terms of Dice score and overlap. This is consistent with the fact that it is a state-of-the-art algorithm, and works best when trying to fill the space with streamlines. However, we show that no algorithm can single-handedly account for the gold standard, and using the union of all methods provides a more complete reconstruction.

In both traditional and RNN-based variants, models with the best Dice/ overlap results also had the worst overreach score. However, in the case of bundle reconstruction, it is less of a concern, because there is always a possibility of applying post-processing techniques to filter streamlines. Also, since our gold standard is not

	Dice	Overlap	Overreach
<i>Reference methods</i>			
Deterministic	0.397	0.267	0.029
Probabilistic	0.553	0.433	0.068
PFT	0.680	0.688	0.266
SET	0.624	0.570	0.184
Ensemble (Det+Prob+PFT+SET)	1.000	1.000	0.000
<i>RNN-based methods</i>			
Det-SE (Learn2track)	0.580	0.495	0.172
Det-Cosine	0.606	0.535	0.204
Prob-Sphere (DeepTract)	0.601	0.534	0.202
Prob-vMF (Entrack)	N/A	N/A	N/A
Prob-Gaussian	0.624	0.585	0.264
Prob-Mixture	0.407	0.284	0.053

Table 4.4 – Tractography evaluation results on the *TractoInferno* dataset. The Prob-vMF model did not produce any results, and is noted as {N/A}.

4.6. RESULTS & DISCUSSION

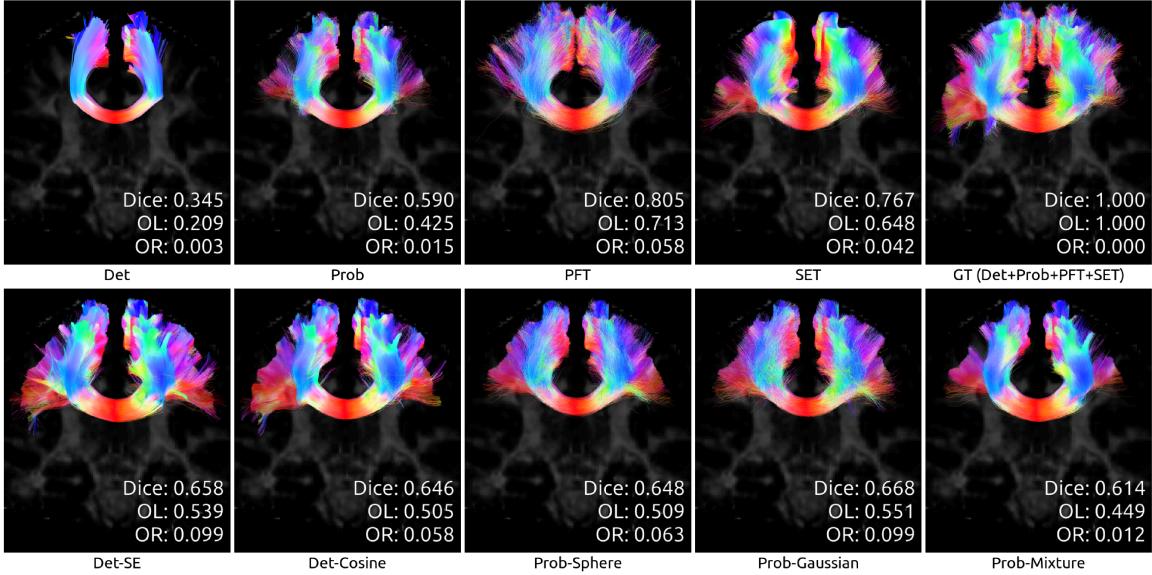


Figure 4.3 – Reconstruction of the *Corpus Callosum* (medium difficulty) by all algorithms, for test subject sub-1006.

perfect, it might not cover the whole possible space as delineated by the RBX algorithm. Furthermore, because the scores are evaluated using binary bundle masks, a small number of streamlines can easily cross a high number of overreaching voxels. Ultimately, the goal is to find a model that can cover as much space as possible, so the overreach score is an interesting information to have, but is not the best indicator of performance in our case.

Of all the RNN-based methods, the Gaussian output model obtained the best Dice score and overlap, hinting that a probabilistic model works best. This is in line with traditional probabilistic algorithms being more suited to bundle reconstruction than deterministic approaches.

Given the worse performance of other probabilistic models, it seems that adding complexity is not beneficial. Training an RNN with a more complex distribution like the mixture of Gaussians might require a different architecture, or more model capacity, to achieve better results. Unfortunately, the RNN with a von Mises-Fisher output had a hard time training, and produced erratic streamlines that mostly did not survive the evaluation pipeline. It would seem that training the vMF distribution is too unstable when using a likelihood loss function, and performing an entropy

4.6. RESULTS & DISCUSSION

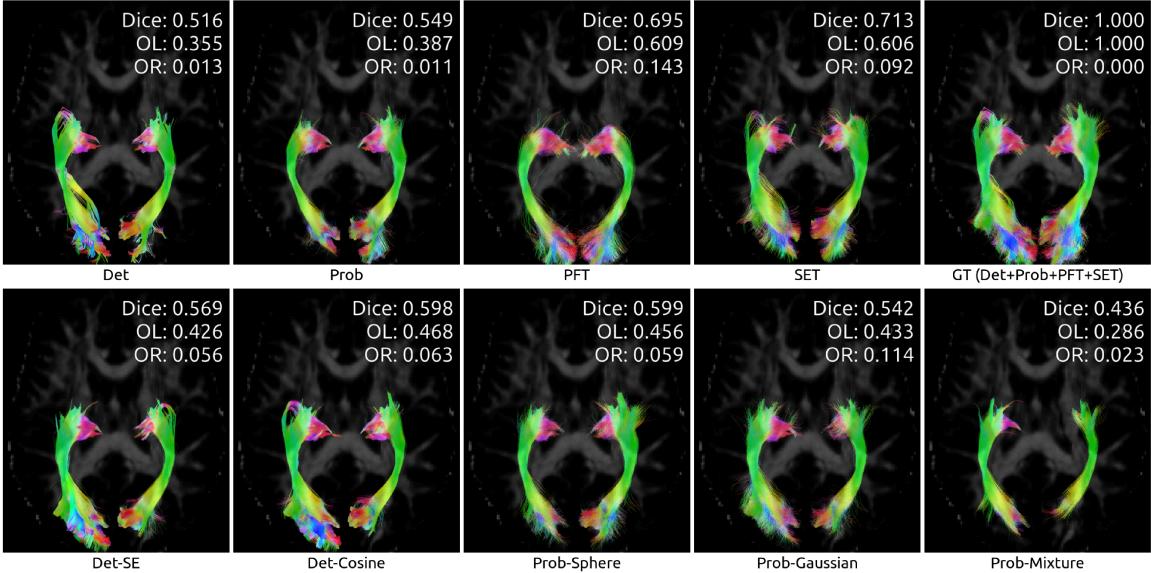


Figure 4.4 – Reconstruction of the *Optic Radiation* (hard difficulty) by all algorithms, for test subject sub-1006.

maximization procedure like the original authors might be required to have a stable training procedure.

Across all results (both reference algorithms and RNN-based methods), the general trend holds that with a better Dice score and overlap, there is also more overreach. This indicates that there is still work to be done to limit the production of false positive streamlines.

To illustrate the differences between algorithms, we showcase the reconstructions of three bundles taken from a random test subject. We chose bundles of both *medium* and *hard* difficulty for tractography, as reported in [82]. Figure 4.3 shows a part of the *Corpus Callosum* (medium difficulty), while Figures 4.4 and 4.5 show the *Optic Radiation* and the *Pyramidal Tract* (hard difficulty). Note that in all cases, as mentioned before, the Prob-vMF method did not produce any meaningful results, which explains why no results are shown.

Also of note, RNN-based models seem to get results on par with traditional algorithms, but not quite as good as the state-of-the-art Particle-Filtered Tractography. However, Poulin et al. produced results far beyond even PFT using an RNN approach trained on a single-database, using a single-bundle per model [102]. While

4.7. CONCLUSION

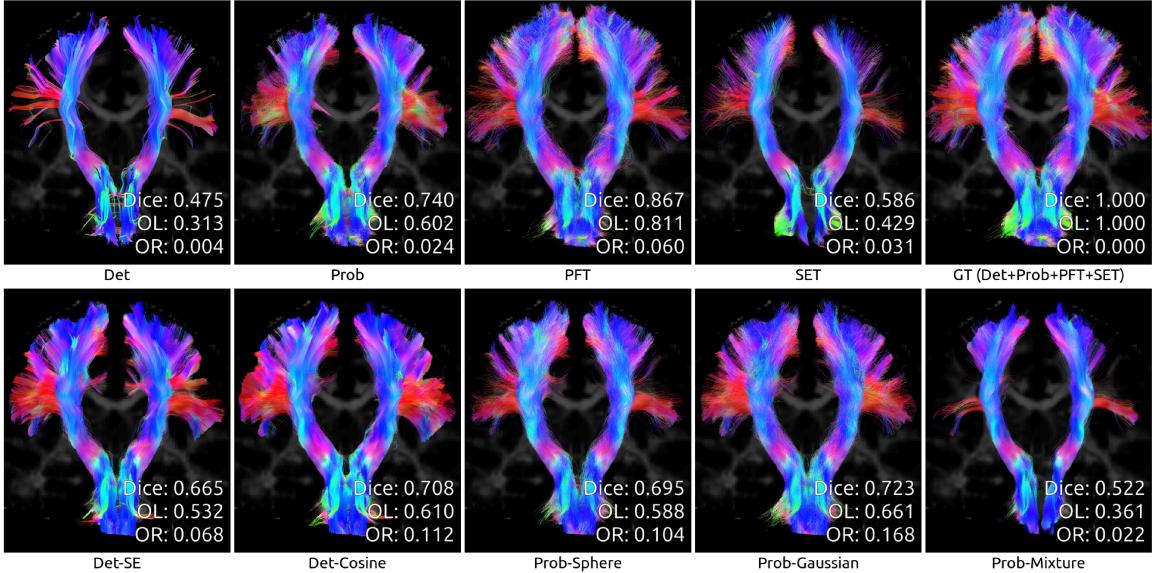


Figure 4.5 – Reconstruction of the *Pyramidal Tract* (hard difficulty) by all algorithms, for test subject sub-1006.

we did not train any model with the single-bundle approach on *TractoInferno*, both results hint that there is a need for more data, more model capacity, or for specialization of algorithms, in order to outperform currently-used methods. We advocate that *TractoInferno* is one way to investigate this problem further.

4.7 Conclusion

We provide an open-access, multi-site dMRI and tractography database aimed at training and evaluating machine learning tractography models. It combines data from multiple datasets, and applies the same processing and QC steps for a uniform database. We also produce results using the available evaluation pipeline for both traditional algorithms and machine learning models based on a recurrent architecture.

We offer *TractoInferno* as a solution to the multiple issues already reported in the literature for machine learning tractography. Indeed, while such algorithms have been proposed in the last few years with promising results, none has been shown to be the fundamental solution to classical tractography. They commonly suffer from variable training data, dissimilar evaluation method, and limited dataset size, among others.

4.8. DATA ACCESS

To this end, a uniform, large-scale, and multi-site database such as *TractoInferno* is an essential tool, paving the way for reproducible and comparable research among machine learning tractography researchers.

4.8 Data access

The *TractoInferno* database is freely available online on the OpenNeuro platform: <https://openneuro.org/datasets/ds003900>. The evaluation pipeline is available on GitHub: <https://github.com/scil-vital/TractoInferno/>.

4.9 Special thanks

Thanks to Laurent Petit for his help with the BIL & GIN database.

4.10 Acknowledgements

P.P. was supported by the FQRNT, grant #206270. This research was enabled in part by Compute Canada (www.computecanada.ca) and CBRAIN (www.cbrain.ca) [123].

4.A *TractoInferno* pipeline execution commands

All commands used to process the *TractoInferno* dataset are reported here. The input files and directories for each command might need to be reorganized between steps; refer to the specific package documentation for more details.

4.A.1 QC DWI

URLs:

- <https://github.com/scilus/dmriqcpy>
- https://github.com/scilus/dmriqc_flow

Command:

4.A. TRACTOINFERNO PIPELINE EXECUTION COMMANDS

```
nextflow run dmriqc-flow-0.1.2/main.nf -profile input_qc  
--root input/  
-with-singularity singularity_dmriqc_0.1.2.img -resume  
--raw_dwi_nb_threads 10
```

4.A.2 TractoFlow

URL: <https://github.com/scilus/tractoflow/>

Command:

```
nextflow run tractoflow-2.1.1/main.nf --root input/  
--dti_shells "0 700 1000 1200"  
--fodf_shells "0 700 1000 1200"  
-with-singularity tractoflow_2.1.1_650f776_2020-07-15.img  
-resume -profile fully_reproducible --mean_frf false  
--set_frf true --nbr_seeds 1
```

4.A.3 QC TractoFlow

URLs:

- <https://github.com/scilus/dmriqcpy>
- https://github.com/scilus/dmriqc_flow

Command:

```
nextflow run dmriqc-flow-0.1.2/main.nf  
-profile tractoflow_qc_light  
--root / ..../TractoFlow/results  
-with-singularity singularity_dmriqc_0.1.2.img -resume
```

4.A.4 SH signal fitting

URL: https://github.com/ppoulin91/tractoinferno_compute_sh_flow

Command:

4.A. TRACTOINFERNO PIPELINE EXECUTION COMMANDS

```
nextflow run code/tractoinferno_compute_sh_flow/main.nf  
--input input/ --use_attenuation --sh_order 6  
-with-singularity scilus-1.2.1.img -resume
```

4.A.5 Tractography (Det, Prob, PFT)

URL: https://github.com/ppoulin91/tractoinferno_tracking_flow

Command:

```
nextflow run code/tractoinferno_tracking_flow/main.nf  
--input ../TractoFlow/results  
-with-singularity tractoflow_2.1.1_650f776_2020-07-15.img  
-resume
```

4.A.6 Tractography (SET)

URLs:

- <https://github.com/StongEtienne/set-nf>
- https://github.com/scilus/convert_set_flow

Commands:

```
nextflow run code/set-nf/main.nf  
--tractoflow ../TractoFlow/results  
--surfaces ../civet/results -profile civet2_dkt  
-with-singularity set_lv1.img -resume  
  
nextflow run code/convert_set_flow/main.nf  
--root_set ../SET/results  
--root_tractoflow ../TractoFlow/results  
-with-singularity scilus-1.2.1.img  
-resume
```

4.A.7 RecoBundlesX (RBX)

URL: https://github.com/scilus/rbx_flow/

4.A. TRACTOINFERNO PIPELINE EXECUTION COMMANDS

Command:

```
nextflow run code/rbx_flow/main.nf -resume  
-with-singularity scilus-1.2.0_rbxflow-1.1.0.img  
-profile large_dataset --input input/  
--atlas_config code/rbx-atlas/config.json  
--atlas_anat code/rbx-atlas/mni_masked.nii.gz  
--atlas_directory code/rbx-atlas/atlas/  
--run_average_bundles false
```

4.A.8 QC RBX

URLs:

- <https://github.com/scilus/dmriqcipy>
- https://github.com/scilus/dmriqc_flow

Command:

```
nextflow run code/dmriqc_flow/main.nf -resume  
-with-singularity  
singularity_dmriqcflow_hotfix_scilpy_1.2.0.img  
-profile rbx_qc --input ../RBX/results/
```

Conclusion

La tractographie traditionnelle comporte plusieurs défis, qui ne peuvent être résolus uniquement par une amélioration de la résolution des images d'IRM de diffusion. L'intervention des experts via des méthodes manuelles ou semi-automatiques est encore requise pour filtrer et valider les résultats de la tractographie. L'intelligence artificielle permet d'utiliser des données expertes afin d'apprendre leur structure et de les reproduire à l'aide d'un modèle entraîné. Il s'agit donc d'un outil tout indiqué pour s'attaquer aux problèmes de la tractographie. Cette thèse présente les premières étapes de l'utilisation des réseaux de neurones récurrents pour la tractographie, ainsi que les problèmes liés au domaine et les solutions proposées.

Le modèle *Learn to Track* [101] a donc été la première tentative d'utiliser un réseau de neurones récurrent pour apprendre la structure d'une tracte de matière blanche valide. Cette publication présentait des résultats quantitatifs sur une acquisition synthétique, ainsi que des résultats qualitatifs sur une acquisition réelle. Elle a mené à de nouvelles publications inspirées de la même approche, soit le RNN à prédiction probabiliste [11], ainsi que les modèles présentés dans le chapitre 4.

Afin de produire des résultats plus crédibles — et de répondre aux critiques concernant le fait que les données d'entraînement étaient simulées et trop peu nombreuses — l'idée de base a été appliquée à la tractographie spécifique aux faisceaux, produisant le modèle *Bundle-Wise Deep Tracker* [102]. D'autres approches non récurrentes ont aussi été proposées pour la tractographie, comme la forêt d'arbres décisionnels [93], le perceptron multicouche [160, 159], les filtres harmoniques hiérarchiques [112], ou encore le réseau à convolutions (pour prédire le volume de faisceaux d'intérêt) [156, 158].

À travers toutes les méthodes proposées, certains problèmes sont devenus apparents, et ont été mis en évidence dans la publication *Tractography and machine*

CONCLUSION

learning : Current state and open challenges [103].

Les solutions mentionnées ont été mises en pratiques lors du développement d'une base de données — *TractoInferno* — construite spécifiquement pour la tractographie par apprentissage, incluant un outil d'évaluation pour obtenir des résultats comparables. En même temps, de nouveaux résultats ont été produits pour montrer le potentiel des réseaux de neurones récurrents dans le domaine de la tractographie.

La tractographie par apprentissage comporte un lot de problèmes à résoudre, et plus spécifiquement avec l'utilisation des réseaux de neurones récurrents. Les résultats présentés ici sont préliminaires, et bien qu'il reste plusieurs défis à résoudre, l'apprentissage machine a un potentiel énorme de répondre à plusieurs problèmes des algorithmes traditionnels.

Cette thèse est l'aboutissement de presque 5 ans de travail. Elle a mené à l'écriture de plusieurs articles. «*Learn to Track : Deep Learning for Tractography*» [101], publié à la conférence MICCAI 2017, présente le premier modèle récurrent appliqué à la tractographie. «*Bundle-Wise Deep Tracker : Learning to track bundle-specific streamline paths*» [102], publié en tant qu'article court à la conférence ISMRM 2018, propose l'utilisation d'un réseau récurrent dans la cadre de la tractographie spécifique aux faisceaux (une version plus complète, mais non publiée, est présentée dans l'[annexe A](#)). «*Tractography and machine learning : Current state and open challenges*» [103], publié dans le journal *Magnetic Resonance Imaging* en 2019, expose plusieurs problèmes qui se sont développés à travers les années dans le domaine de la tractographie par apprentissage. «*TractoInferno : A large-scale, open-source, multi-site database for machine learning dMRI tractography*», soumis à *Nature Scientific Data*, détaille la création d'une nouvelle base de données dont la but est de répondre aux problèmes de comparaison d'algorithmes en tractographie par apprentissage.

CONCLUSION

Directions futures pour la tractographie par apprentissage

Autres approches pour la tractographie par apprentissage

D'autres approches encore inexplorées seraient intéressantes pour la tractographie par apprentissage. On peut s'inspirer des modèles pour la génération musicale, comme le *WaveNet* [147]. Il ne s'agit pas d'un modèle récurrent, puisque la taille du contexte séquentiel est fixé. Par contre, la structure hiérarchique du modèle permet justement d'utiliser un très long contexte sans nécessiter autant de mémoire qu'une réseau de neurones artificiels standard. Un tel modèle permettrait d'expérimenter pour comprendre quelle taille de contexte est nécessaire pour guider de façon appropriée la tractographie.

On peut également s'inspirer de la traduction de textes, en utilisant un modèle comme le *Transformer* [151]. Celui-ci utilise le concept d'attention afin de concentrer l'information utilisée par le modèle sur certaines parties de la séquence. Cette attention est prédite par le modèle en fonction de la position de la séquence, et aurait le potentiel de mieux utiliser le chemin passé d'une tracte pour sortir d'une région de chevauchement, par exemple.

Aussi, d'autres approches non-supervisées existent également pour la tractographie, comme l'apprentissage par renforcement [137]. Dans ce contexte, on n'utilise pas de tractes de référence. On fixe plutôt une fonction de récompense qui respecte certains critères désirables. Par exemple, dans le cadre de la tractographie, une grande récompense serait associée à une tracte qui i) est bien alignée avec l'ODF de fibre, ii) est aussi longue que possible, et iii) demeure aussi lisse que possible. Aussi, une approche par auto-encodeurs [77] existe, où on «compresse» l'information d'une tracte dans un espace latent. Un modèle entraîné peut ensuite générer de nouvelles tractes en explorant son espace latent.

Améliorations possibles au modèle récurrent

Diverses améliorations seraient possibles pour le modèle récurrent. Par exemple, comme le modèle de forêt d'arbres décisionnels de [93], on peut ajouter une valeur de

CONCLUSION

sortie qui prédit si la tracte est terminée, et que le processus de tractographie devrait s'arrêter (p. ex. si la tracte est hors de la matière blanche. Ainsi, il ne serait plus requis d'utiliser un masque de matière blanche lors de l'utilisation du modèle après l'entraînement.

Aussi, la force de l'apprentissage machine est sa capacité d'utiliser n'importe quelles données en entrée. En ajoutant diverses informations, comme une image T1, ou bien des cartes de segmentations de tissus probabilistes, le modèle pourrait tirer profit de ces informations supplémentaires pour faire de meilleures prédictions.

Finalement, les nouvelles séquences d'acquisition par tenseur-b [45, 162] peuvent fournir des informations anatomiques, autrement impossible à avoir avec une acquisition linéaire standard par facteur-b. Ce signal est difficile à utiliser cependant pour les méthodes classiques, puisqu'il ne vit pas dans le même espace, c-à-d sur une sphère de rayon constant en espace-q. L'utiliser en entrée d'un modèle d'apprentissage machine permettrait d'en profiter et d'extraire l'information pertinente à la tractographie.

Validation des données et vérité terrain

Il est encore difficile d'obtenir une vérité terrain pour la tractographie. Les définitions de faisceaux reposent encore historiquement sur des études histologiques et des dissections *post-mortem*. Elles sont parfois basées sur des descriptions textuelles, et utilisent des points de référence anatomiques qui peuvent varier d'une personne à l'autre. Ces définitions ne sont pas nécessairement adaptées à l'apprentissage machine. On commence à voir de nouvelles définitions plus logiques, comme dans [152]. Les données validées à l'aide de traceurs sont aussi sujettes à une certaine variabilité [121] et sont extrêmement limitées pour une utilisation en tractographie par apprentissage.

Il y a un besoin criant de données pour l'utilisation de l'apprentissage machine en tractographie. Or, le travail des experts ne suffit à produire les quantités nécessaires ; il faut donc trouver des façons de générer et valider des données de façon automatique. Revoir la définition des faisceaux serait une première étape, de sorte à obtenir des définitions plus claires et reproductibles que les définitions utilisées présentement. Ainsi, les méthodes de segmentation automatique de faisceaux, similaires à la méthode

CONCLUSION

proposée dans le [chapitre 4](#), seraient probablement plus fiables, sans nécessiter de vérification manuelle subséquente.

Harmonisation de données de diffusion

Un des problèmes que nous avons parfois remarqué est la difficulté du modèle récurrent à généraliser à de nouvelles données (provenant d'un nouveau site d'acquisition) — même comparativement à d'autres modèles d'intelligence artificielle non récurrents. Une hypothèse est le manque d'harmonisation des données lorsque plusieurs sites d'acquisition sont impliqués. En effet, comme le réseau récurrent dépend de longues séquences de données de diffusion pour faire ses prédictions, il peut être plus sensible aux variations de signal entre les sites d'acquisition. L'harmonisation de données est un domaine d'intérêt pour la recherche depuis déjà plusieurs années [96], et le sera d'autant plus maintenant que l'intelligence artificielle prend une place de plus en plus importante dans le domaine de l'imagerie médicale.

L'utilisation d'une base de données multi-site comme *TractoInferno* [104] est une solution partielle, puisque même si on permet au modèle de voir des données provenant de plusieurs sites, rien ne garantit son comportement face à un site d'acquisition inconnu. Une prochaine version de *TractoInferno* pourrait proposer un site d'acquisition réservé au jeu de test, afin de mesurer la capacité de généralisation à un site inconnu.

Exploration ou reconstruction ?

Une autre question importante pour la tractographie par apprentissage est la capacité du modèle à «explorer» le cerveau, par opposition à la reconstruction de structure vues en entraînement. Les prédictions d'un modèle peuvent être remises en question si elles concernent des structures absentes des données d'entraînement. Aussi, le comportement des algorithmes d'apprentissage peut être imprévisible en présence de tumeurs ou de lésions. En effet, on ne connaît pas entièrement le comportement des fibres de matière blanche en présence de différentes pathologies. Les fibres peuvent être déplacées, coupées, nécrosées, et le comportement attendu d'un algorithme de tractographie est encore incertain. Il s'agit d'une étape importante à franchir avant

CONCLUSION

de pouvoir utiliser la tractographie par apprentissage avec des sujets pathologiques.

CONCLUSION

Publications résultant de cette thèse

Articles de journaux

Premier auteur

- **Poulin, Philippe**, Daniel Jörgens, Pierre-Marc Jodoin, and Maxime Descoteaux. "Tractography and machine learning : Current state and open challenges." *Magnetic resonance imaging* 64 (2019) : 37-48.
- **Poulin, Philippe**, Guillaume Theaud, Francois Rheault, Etienne St-Onge, Arnaud Bore, Emmanuelle Renauld, Louis de Beaumont, Samuel Guay, Pierre-Marc Jodoin, and Maxime Descoteaux. "TractoInferno : A Large-Scale, Open-Source, Multi-Site Database for Machine Learning DMRI Tractography." *BiorXiv* : 470422 (2021).

Co-auteur

- Rheault, Francois, **Philippe Poulin**, Alex Valcourt Caron, Etienne St-Onge, and Maxime Descoteaux. "Common misconceptions, hidden biases and modern challenges of dMRI tractography." *Journal of neural engineering* 17, no. 1 (2020) : 011001.
- Rheault, Francois, Alessandro De Benedictis, Alessandro Daducci, ..., **Philippe Poulin**, ..., Chiara Maffei, Chantal MW Tax, David Romascano, Eduardo Caverzasi et al. "Tractostorm : The what, why, and how of tractography dissection reproducibility." *Human brain mapping* 41, no. 7 (2020) : 1859-1874.

Articles de conférences

- **Poulin, Philippe**, Marc-Alexandre Cote, Jean-Christophe Houde, Laurent Petit, Peter F. Neher, Klaus H. Maier-Hein, Hugo Larochelle, and Maxime Descoteaux. "Learn to track : Deep learning for tractography." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 540-547. Springer, Cham, 2017.

CONCLUSION

Résumés de conférences

- **Poulin, Philippe**, Francois Rheault, Etienne St-Onge, Pierre-Marc Jodoin, and Maxime Descoteaux. "Bundle-Wise Deep Tracker : Learning to track bundle-specific streamline paths." Proceedings of the International Society for Magnetic Resonance in medicine ISMRM-ESMRMB (2018).
- Jörgens, Daniel, **Philippe Poulin**, Rodrigo Moreno, Pierre-Marc Jodoin, and Maxime Descoteaux. "Towards a deep learning model for diffusion-aware tractogram filtering." In ISMRM 27th Annual Meeting & Exhibition, 11-16 May 2019, Montréal, QC, Canada. 2019.

Stages et collaborations

- 2021 - Stage MITACS de huit mois chez Imeka (Sherbrooke, QC).
 - Étude de l'harmonisation de données d'imagerie de diffusion.
 - Stage réalisé sous la supervision de Jean-Christophe Houde.

Annexe A

Tractographie par apprentissage spécifique aux faisceaux

Résumé

La tractographie par apprentissage à l'aide de réseaux récurrents a beaucoup de potentiel. Cet article propose une amélioration du modèle original *learn2track*, en plus d'utiliser une base de données de sujets réels au lieu de données synthétiques. De plus, en utilisant un modèle spécifique à chaque faisceau, il est possible de produire trois à dix fois plus de tractes qu'une méthode de pointe, ainsi que de couvrir un volume plus large. Ces résultats indiquent que l'utilisation de modèles récurrents a beaucoup de potentiel, et que de plus grosses bases de données et des modèles avec plus de capacité sont nécessaires pour accomplir une tâche complexe comme la tractographie.

Contributions de la publication

- Amélioration du modèle récurrent proposé par *learn to track* (voir [chapitre 2](#)).

- Production de trois à dix fois plus de tractes valides qu'une méthode de pointe.
- Présentation de résultats quantitatifs impressionnantes pour un réseau de neurones récurrent entraîné sur une base de données de 37 sujets réels.

Contributions des auteurs

- Entraînement du modèle, écriture du manuscrit, analyse des résultats, création des figures (Philippe Poulin)
- Création de l'outil d'évaluation semi-automatique (François Rheault et Laurent Petit)
- Génération des tractogrammes comparatifs (*baselines*) (François Rheault et Étienne St-Onge)
- Supervision du projet et révision du manuscrit (Maxime Descoteaux et Pierre-Marc Jodoin)

Commentaires

Suite à l'article *Learn to track*, publié à la conférence MICCAI en 2017, nous avons réalisé le potentiel des réseaux de neurones récurrents pour la tractographie, mais également les failles de la méthode originale. La tractographie spécifique aux faisceaux de [115] a été développée au même moment, et il s'agissait de l'occasion parfaite pour collaborer et développer une version spécifique aux faisceaux de *Learn to Track*. Nous avons pu montrer que le modèle est non seulement encore performant sur des données réelles, mais a surpassé nos ententes dans le contexte de la tractographie spécifique aux faisceaux. Il s'agissait donc d'un indice que la tâche originale de tractographie globale du cerveau était peut-être trop complexe comme défi initial, notamment en raison de la quantité limitée de données (qui a conséquemment mené au développement de la base de données du [chapitre 4](#)).

Cet article a été soumis à la conférence MICCAI en 2018, mais a été refusé, principalement en raison du manque de nouveauté du modèle, qui était un réseau récurrent comme *Learn to track*. Malheureusement, l'article n'a pas été soumis dans une autre conférence ou journal par la suite, nos efforts étant plutôt dirigés vers l'article présenté au [chapitre 3](#). Je considère que les résultats présentés sont tout de même essentiels pour comprendre la progression de la méthode, ainsi que pour mettre en perspective le [chapitre 4](#), où les résultats obtenus par le modèle récurrent sont comparables aux méthodes traditionnelles. Une future approche spécifique aux faisceaux appliquée à *TractoInferno* serait donc intéressante pour aider à comprendre le potentiel et les limites d'un modèle récurrent sur des données réelles et variées.

Bundle-Wise Deep Tracker: Recurrent networks tracking bundle-specific streamline paths

Philippe Poulin^{1,2}, Francois Rheault¹, Etienne St-Onge¹, Laurent Petit³, Pierre-Marc Jodoin², Maxime Descoteaux¹

¹ Sherbrooke Connectivity Imaging Laboratory (SCIL), Computer Science Department, Université de Sherbrooke, Sherbrooke, Canada

² Videos & Images Theory and Analytics Laboratory (VITAL), Computer Science department, Université de Sherbrooke, Sherbrooke, Canada

³ Groupe d’Imagerie Neurofonctionnelle, IMN, CNRS, CEA, Université de Bordeaux, Bordeaux, France

Abstract

We propose a novel bundle-wise tracking algorithm based on deep learning and recurrent neural networks. This allows bundle-specific features to be learned from the diffusion signal without the need to reconstruct a fiber orientation distribution. We study multiple bundles of interest: the left-right arcuate fasciculus (AF), the sensory-motor subsection of the corpus callosum (CC), and the left-right pyramidal tracts (PyT). Using a dataset of 37 *in vivo* human dMRI acquisitions, our proposed solution has 3 to 10 times more valid streamlines than state-of-the-art methods. Overall, our model consistently produces more valid streamlines, covers more volume and is more efficient than other methods.

A.1. INTRODUCTION

A.1 Introduction

Classic tractography methods are known to make systematic errors, such as producing false positive streamlines or failing to recover the full extent of bundles [82]. Part of the problem is that local modeling and tracking approaches usually cannot untangle common complex axonal configurations, thus leading to frequent false positives in “bottleneck” areas, i.e. where several bundles are intermingled. Furthermore, reconstructing the full extent of a bundle is a hard task, in part due to partial volume effects. Addressing those problems is important: tractometry is usually concerned with bundle extent, and connectomics studies are particularly affected by false positives. Data-driven, non-local methods are some of the innovative ways in which those problems can be tackled. Machine learning is another promising avenue with recent works using diffusion MRI for various applications (more details in section A.2). For tractography, however, results suggest that learning methods are prone to make the same mistakes as classic methods, i.e. generating a prohibitive number of false positive streamlines.

In this work, we choose to explore bundle-wise tractography. This is a constrained version of the tracking problem where a specific bundle of interest has been identified. It is particularly useful in domains like neurological disease studies [31] and neuro-surgical planning [46]. However, even if tracking is restrained to a single bundle, the diffusion signal is tinged with fiber crossing from other bundles, and tracking is prone to follow those pathways. In such a case, machine learning methods are fundamentally relevant as they can learn discriminant features for making predictions, i.e. learn the right diffusion features to properly explain the bundle.

We propose a 3-layer deep recurrent neural network that computes an internal representation of diffusion features tailored to bundle-wise tracking predictions. We name our method the Bundle-Wise Deep Tracker (BW-DeepTracker). We use *in vivo* human dMRI acquisitions to train and test our method. Specifically, we study five bundles of interest: the left & right arcuate fasciculus (AF), the sensory-motor subsection of the corpus callosum (CC), and the left & right pyramidal tracts (PyT). We train our method on 25 subjects using reference streamlines, and test generalization performance on 12 subjects. We measure multiple metrics to compare tracking meth-

A.2. RELATED WORK

ods, and the proposed model consistently obtains the best results. In some cases, our method even drastically outperforms state-of-the-art deterministic and probabilistic tracking methods.

A.2 Related work

A series of deep learning models for tractography-related problems were recently published. For example, Wasserthal *et al.* [156] proposed a U-Net approach for direct bundle segmentation based on fiber ODFs while Koppers *et al.* [71] use a convolutional neural network to reconstruct fiber ODFs from diffusion MRI. In contrast to these two methods, our proposed solution does not force a representation on the signal and the model is free to learn from the data.

Neher *et al.* showed that machine learning can be used specifically for tractography [93]. While their method is non-local (neighborhood is considered through multiple predictions, and the previous tracking direction is added as input), it does not take into account streamline evolution. In our case, we compute a tracking-specific representation conditioned on the whole streamline being tracked, and the neighborhood is included as input instead of using multiple predictions.

Recurrent neural networks are well-suited for a variety of problems involving a sequence of observations such as machine translation [131], time series prediction [55], and handwriting generation [57]. As such, RNNs offer an interesting footing for tractography which, like handwriting generation, involves the generation of a series of points in space. In that perspective, Graves *et al.* [57] underlined that RNNs can handle long-range structures similar to streamlines (handwritten lines were composed of 700 steps on average in the paper).

To our knowledge, the only RNN model used in tractography was proposed in 2017 by Poulin *et al.* [101]. Unfortunately, their model did not include any neighborhood information or regularization, and results were focused on whole-brain tracking of a single subject. Although the authors showed good results, the fact that the model was trained on a single synthetic subject (and qualitatively tested on only one real subject) leads us to believe that its real generalization capabilities on real data has yet to be shown.

A.3. METHODOLOGY

A.3 Methodology

A.3.1 Dataset

We trained and tested our method with diffusion-weighted images (DWI) from 37 participants gathered from the BIL&GIN database [86]. The only preprocessing applied to the images was an eddy current correction, done with the FMRIB software library. Images are composed of $2 \times 2 \times 2 \text{ mm}^3$ isotropic voxels. A template was computed using the FA maps of all acquisitions in order to register the final streamlines onto a common space for segmentation and evaluation. The template was also used to define seeding and tracking masks for all methods. Finally, the 37 participants were randomly divided into training (20), validation (5) and test (12) sets.

A.3.2 Model

In the wake of the *learn to track* architecture of Poulin *et al.* [101], we propose the Bundle-Wise Deep Tracker (BW-DeepTracker), a Recurrent Neural Network (RNN) trained to predict tracking directions using bundle-wise input data. The main idea is to dismiss fiber ODFs, and instead use a non-local, data-driven approach to map the input signal to a conditional tracking direction. Figure A.1 details the model architecture. The core model is a Gated Recurrent Unit, composed of three hidden layers of size 512, 256 and 128, respectively. We apply layer normalization on each layer. Skip connections connect the input to all hidden layers, and all hidden layers to the output layer. Optimization is done with the Adam optimizer, using an initial learning rate of 0.0001 and gradient clipping of norm 5. During training, an L2 reconstruction error is minimized between normalized predictions and targets, and early stopping is used with a lookahead of 15 epochs. Models are regularized using a variant of Dropout called Zoneout [72] with a drop probability of 0.3. Code will be made available upon acceptance of this paper.

A.3.3 Model inputs

Each diffusion-weighted image is normalized by the $b=0$ image, then fitted with spherical harmonics of order 4 [35]. We normalize the data for each subject in non-

A.3. METHODOLOGY

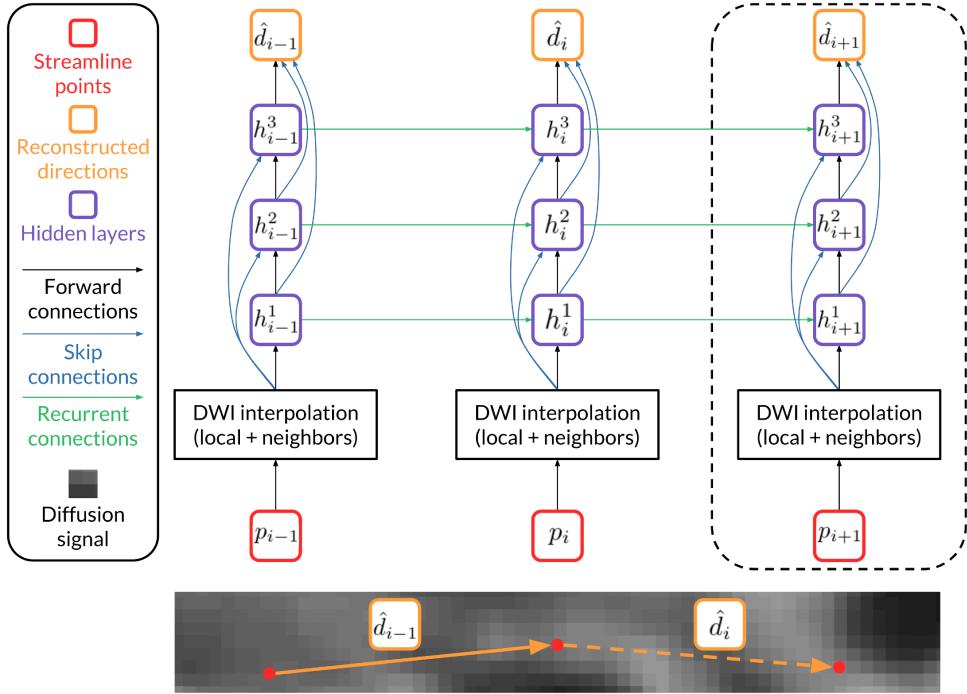


Figure A.1 – BW-DeepTracker model, from streamline coordinates and diffusion information to predicted directions. Dotted line represents future computations. (Figure best viewed in color)

zero voxels, so that each dimension has zero mean. Model inputs are interpolated at reference streamline positions, along with neighborhood information. This neighborhood information is added by concatenating the interpolated coefficients from 6 neighboring positions at a distance of 1mm.

A.3.4 Reference streamlines

Reference streamlines our system is trained on are generated using probabilistic bundle-specific tracking (BST) with particle filtering [115]. They are then automatically segmented by the same procedure used for evaluation (detailed in section A.3.5). Final valid streamlines are resampled to a constant step size of 1.0mm. In order to reduce the dataset size and accelerate training, we use a downsampling procedure to remove near-identical streamlines. The resulting bundles are less densely packed

A.3. METHODOLOGY

but still fill the original volume. To offset this loss of information, we add random zero-mean Gaussian noise with 0.05mm variance to the input coordinates of each streamline at each epoch of the training process. Doing so can be seen as a data augmentation technique allowing the model to see a wider variety of streamlines while having the advantage of accelerating the training process. After downsampling, training streamlines across all 20 subjects amount to the following numbers: 341,514 (Left AF), 167,854 (Right AF), 223,865 (CC), 31,674 (Left PyT) and 19,693 (Right PyT).

A.3.5 Evaluation

We compare our model with classic deterministic (Det) and probabilistic (Prob) algorithms, along with a state-of-the-art bundle-specific tracking algorithm, which is a fairer comparison since our model is also bundle-specific. The baseline tracking methods are: 1) Det [50], 2) Prob [50], 3) Det-BST [115], 4) Prob-BST with particle filtering (PFT) [115], and 5) *learn to track* [101] (Det). Baseline algorithms 1-4 use the constrained spherical deconvolution algorithm to compute fiber ODFs for tracking. Note that method 4 (Prob-BST+PFT) is the one used to generate our model’s reference streamlines. We trained a *learn to track* recurrent model using the code provided by the authors. In this case, we formed a training set from all 5 bundles (to reflect the whole-brain approach of the original method), and used no downsampling procedure. We used the hyperparameters provided in the original paper.

All methods track five bundles of interest in the randomly chosen 12 test subjects: AF (left/right), CC, and PyT (left/right). Note that [82] identified the tracking difficulty of the AF and CC as medium, and the CST/PyT as hard. Tracking is initialized with 5 seeds per voxel in the white matter (WM) - gray matter (GM) interface, using a dilated WM tracking mask (template registration is used to provide seeding and tracking masks). Methods 1-4 use a step size of 0.2mm while the recurrent methods use a step size of 1.0mm.

In order to assess performances, valid streamlines are automatically extracted from the resulting tractogram of each method. This is done with ROIs defined by an expert neuroanatomist in template space. Across all test subjects, 5 metrics are computed:

A.4. RESULTS

number of valid streamlines, volume coverage, valid streamline ratio, efficiency ratio (number of valid tracking steps over total number of tracking steps), and average weighted Dice score (a modification to the Dice similarity measure taking into account the number of streamlines going through each voxel). The weighted Dice score is computed on pairwise subjects, and measures how much the volume covered by the generated streamlines varies between subjects, or how “reliable” tracking is. Note that the efficiency ratio is a particularly interesting metric, because it measures how much computing power is “wasted” on tracking invalid streamlines.

A.4 Results

One BW-DeepTracker model was trained on each bundle using an Nvidia TITAN Xp GPU, which took between 12 and 24 hours per model (depending on the number of streamlines in a bundle’s training set and the number of epochs before validation error stopped improving). Testing results are detailed in Table A.1.

The proposed method scored best across every metric for all five bundles (in some cases, two methods are identified as best, because no statistically significant difference was found with a p-value of 5%). Across all bundles, less streamlines are rejected, and the rejected streamlines are generally shorter (as seen in the efficiency ratio), which leads to faster tracking when generating the same number of streamlines.

The most striking result is our model’s valid streamline ratio, where absolute scores are improved between 15.21% and 23.6% over the second best method (resulting in 3x to 10x more valid streamlines). This suggests that we produce less false positive streamlines, i.e. streamlines that branch off on another bundle “pathway”. The high weighted Dice score shows that BW-DeepTracker is reliable across subjects. Combined with the highest volume coverage scores, this means that our method can consistently reach bundle areas missed by other methods. Also note that volume coverage is overall better than a state-of-the-art probabilistic tracking with particle filtering even though it is a deterministic model, and is much better than other deterministic methods.

High efficiency scores are explained to some extent by the step size used by recurrent models (1.0mm for *learn to track* and BW-DeepTracker compared to 0.2mm

A.4. RESULTS

Bundle	Method	Number of valid streamlines	Volume (mm ³)	Valid ratio (%)	Efficiency ratio (%)	Average Weighted Dice
Left AF	Det	1,570 ± 381	17,044 ± 2,791	1.3 ± 0.3	11.0 ± 2.5	0.41 ± 0.09
	Prob	2,433 ± 608	48,642 ± 6,040	1.7 ± 0.5	9.9 ± 2.7	0.68 ± 0.07
	Det-BST	9,623 ± 1,586	37,847 ± 4,855	8.3 ± 1.2	45.4 ± 5.2	0.62 ± 0.07
	Prob-BST+PFT	17,743 ± 2,501	106,634 ± 8,460	9.6 ± 1.2	31.3 ± 3.7	0.85 ± 0.03
	Learn to track	196 ± 270	3454 ± 3416	0.1 ± 0.1	1.0 ± 1.5	0.11 ± 0.10
	BW-DeepTracker	73,510 ± 7,057	144,221 ± 5,391	33.2 ± 3.2	66.7 ± 5.2	0.93 ± 0.01
Right AF	Det	916 ± 326	10,313 ± 3,061	1.0 ± 0.3	8.1 ± 3.9	0.29 ± 0.13
	Prob	1,294 ± 354	32,889 ± 5,636	1.1 ± 0.3	6.1 ± 2.1	0.56 ± 0.12
	Det-BST	5,379 ± 1,905	24,156 ± 4,852	5.8 ± 2.0	34.9 ± 10.9	0.51 ± 0.12
	Prob-BST+PFT	9,557 ± 2,658	81,231 ± 11,008	6.1 ± 1.7	18.7 ± 5.1	0.80 ± 0.05
	Learn to track	82 ± 53	891 ± 931	0.0 ± 0.0	0.2 ± 0.3	0.06 ± 0.05
	BW-DeepTracker	53,860 ± 8,616	123,703 ± 11,427	27.5 ± 4.8	54.1 ± 9.3	0.90 ± 0.03
CC	Det	2,797 ± 916	19,433 ± 4,962	4.3 ± 1.3	27.6 ± 5.4	0.59 ± 0.09
	Prob	1,517 ± 467	34,333 ± 6,338	2.0 ± 0.6	13.6 ± 2.9	0.74 ± 0.07
	Det-BST	9,379 ± 2,281	33,419 ± 6,324	13.5 ± 2.6	44.0 ± 5.1	0.68 ± 0.08
	Prob-BST+PFT	12,230 ± 3,049	85,830 ± 7,199	11.3 ± 2.4	28.4 ± 5.6	0.86 ± 0.05
	Learn to track	11 ± 21	536 ± 864	0.0 ± 0.0	0.1 ± 0.3	0.01 ± 0.02
	BW-DeepTracker	40,262 ± 6,177	85,008 ± 6,307	31.7 ± 4.5	47.6 ± 5.6	0.88 ± 0.03
Left PyT	Det	419 ± 428	3,499 ± 1,844	1.1 ± 1.1	9.5 ± 7.8	0.41 ± 0.20
	Prob	269 ± 180	10,488 ± 3,175	0.6 ± 0.4	4.5 ± 2.8	0.69 ± 0.10
	Det-BST	1,580 ± 1,111	6,511 ± 2,258	4.3 ± 2.9	20.9 ± 11.6	0.58 ± 0.16
	Prob-BST+PFT	2,384 ± 1,103	28,468 ± 7,185	3.8 ± 1.7	11.6 ± 5.0	0.85 ± 0.06
	Learn to track	29 ± 20	1184 ± 531	0.0 ± 0.0	0.5 ± 0.4	0.14 ± 0.07
	BW-DeepTracker	17,292 ± 3,787	36,664 ± 5,430	20.7 ± 4.4	38.3 ± 8.2	0.86 ± 0.05
Right PyT	Det	396 ± 315	3,603 ± 1,815	1.0 ± 0.8	9.5 ± 7.5	0.35 ± 0.17
	Prob	198 ± 112	8,809 ± 3,294	0.4 ± 0.2	3.2 ± 1.9	0.60 ± 0.12
	Det-BST	1,498 ± 1,039	6,661 ± 2,755	3.9 ± 2.7	19.8 ± 13.6	0.48 ± 0.18
	Prob-BST+PFT	1,683 ± 749	25,243 ± 6,271	2.5 ± 1.1	7.5 ± 3.2	0.83 ± 0.06
	Learn to track	23 ± 13	960 ± 411	0.0 ± 0.0	0.4 ± 0.2	0.14 ± 0.10
	BW-DeepTracker	16,821 ± 8,745	36,258 ± 10,523	19.1 ± 10.3	32.4 ± 17.1	0.76 ± 0.17

Table A.1 – Evaluation metrics on the test set, with means and standard deviations computed across all 12 subjects for five previous methods [50, 115, 101] and our proposed deep learning solution (BW-DeepTracker). Best results are in bold (Welch’s unequal variances t-test [161] was used to determine results that were not significantly different with a p-value of 0.05).

for other baseline methods). In this case, 1.0mm was empirically determined to be a sweet spot for performance.

We provide a visual comparison for a random test subject in Figure A.2. Our method recovers most of the fanning structure in all bundles (especially visible in the PyT). It also produces smoother streamlines than probabilistic methods, while being able to navigate through narrow areas and reconstruct a fuller extent of the bundle.

We observe very poor results for the *learn to track* method [101] (although we rigorously used the same implementation than the authors). As shown in the supplementary materials, *learn to track* streamlines are visually plausible at a global scale (which is coherent with the original paper), but locally do not follow the bundle

A.4. RESULTS

definition. We hypothesize that trying to learn the structure of five bundles at the same time is a difficult problem without seeing more varied structures, as in true whole-brain tracking. Also, instead of using the original hyper-parameters, a full hyper-parameter search could potentially improve results. That said, the fact that good *learn to track* results in the original paper were obtained only on synthetic data cast a shadow on the true generalization capabilities of the model.

While we do not report the wall clock time for tracking, using a deep-learning

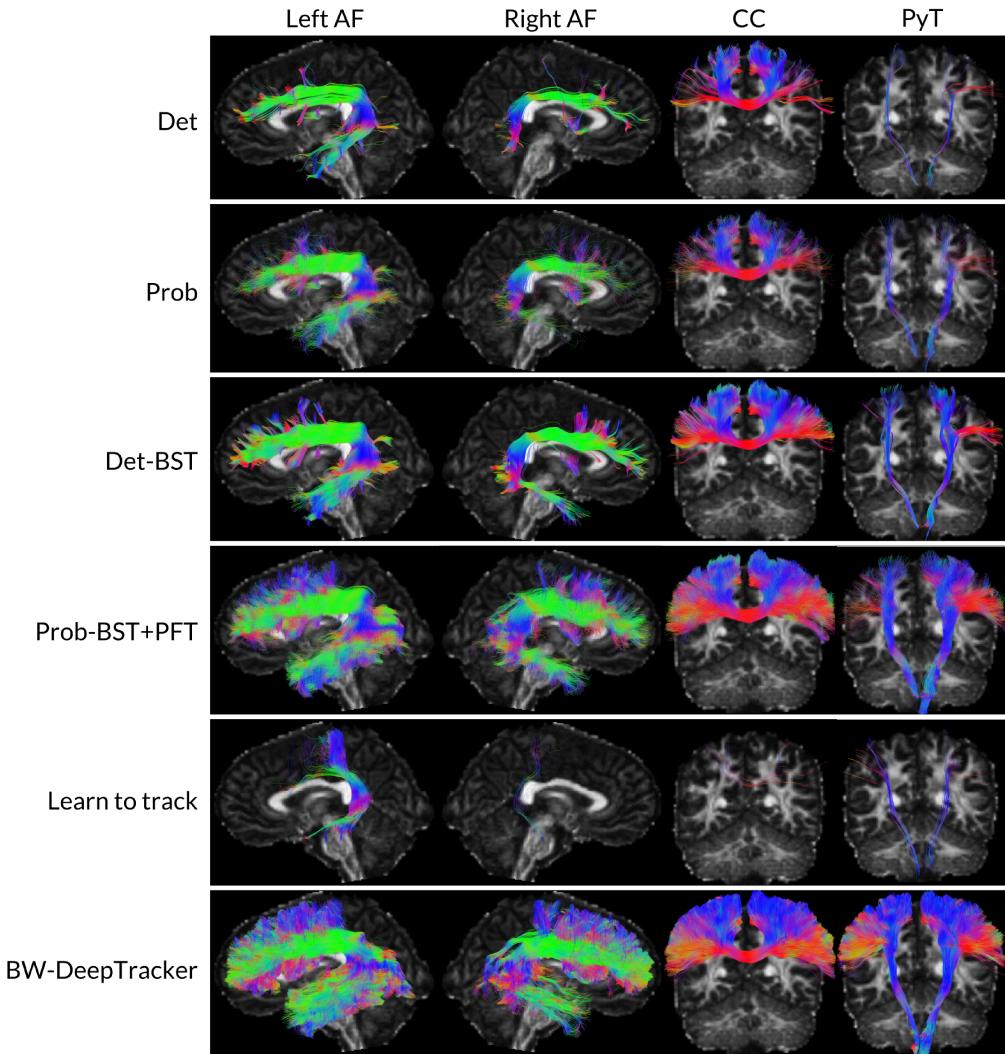


Figure A.2 – Tracked bundles on BIL&GIN test subject t0203 (best viewed in color).

A.5. CONCLUSION

GPU implementation results in faster tracking times than other baseline methods, which are implemented on CPU with parallelized seeding. For example, tracking in a single subject from around 200,000 seeds in the left AF takes about 4 hours for deterministic methods, 2 hours for probabilistic methods and 3 minutes for *learn to track* and BW-DeepTracker. We do not make any claim about the computing efficiency of each method, but the fact that a GPU implementation is available for the proposed model gives it a non-trivial advantage, especially for large-scale studies.

A.5 Conclusion

The proposed BW-DeepTracker model outperforms deterministic and probabilistic state-of-the-art methods on unseen *in vivo* human dMRI acquisitions. By focusing on bundle-wise tracking, we were able to get the best results compared to five baseline methods in all metrics across five bundles. Overall, our model consistently produces more valid streamlines, covers more volume and is more efficient than other methods. Given computational efficiency and GPU implementation of BW-DeepTracker, tracking times are reduced by an order of magnitude compared to baseline methods that only provide a parallelized CPU implementation, which is especially useful for clinical usage.

This is a promising way to address the biggest issues of tractography, where generating false positives and missing the full extent of bundles are still common problems. The reported results are encouraging for fields like tractometry, connectomics and neurosurgical planning.

Bibliographie

- [1] I. Aganj, C. Lenglet, et G. Sapiro, “ODF reconstruction in q-ball imaging with solid angle consideration,” dans *Biomedical Imaging: From Nano to Macro, 2009. ISBI’09. IEEE International Symposium.* IEEE, 2009, pp. 1398–1401.
- [2] J. L. R. Andersson et S. N. Sotropoulos, “An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging,” *NeuroImage*, vol. 125, pp. 1063–1078, janvier 2016.
- [3] J. L. R. Andersson, S. Skare, et J. Ashburner, “How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging,” *NeuroImage*, vol. 20, no. 2, pp. 870–888, octobre 2003.
- [4] B. B. Avants, C. L. Epstein, M. Grossman, et J. C. Gee, “Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain,” *Medical Image Analysis*, vol. 12, no. 1, pp. 26–41, février 2008.
- [5] P. J. Basser, J. Mattiello, et D. Lebihan, “Estimation of the Effective Self-Diffusion Tensor from the NMR Spin Echo,” *Journal of Magnetic Resonance, Series B*, vol. 103, no. 3, pp. 247–254, mars 1994.
- [6] P. J. Basser, J. Mattiello, et D. LeBihan, “MR diffusion tensor spectroscopy and imaging,” *Biophysical Journal*, vol. 66, no. 1, pp. 259–267, janvier 1994.
- [7] P. J. Basser, S. Pajevic, C. Pierpaoli, J. Duda, et A. Aldroubi, “In vivo fiber tractography using DT-MRI data,” *Magnetic Resonance in Medicine*, vol. 44, no. 4, pp. 625–632, 2000.

BIBLIOGRAPHIE

- [8] S. Basu, T. Fletcher, et R. Whitaker, “Rician noise removal in diffusion tensor MRI,” dans *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2006, pp. 117–125.
- [9] T. E. Behrens, H. J. Berg, S. Jbabdi, M. F. Rushworth, et M. W. Woolrich, “Probabilistic diffusion tractography with multiple fibre orientations: What can we gain?” *Neuroimage*, vol. 34, no. 1, pp. 144–155, 2007.
- [10] C. Beleites, U. Neugebauer, T. Bocklitz, C. Krafft, et J. Popp, “Sample size planning for classification models,” *Analytica Chimica Acta*, vol. 760, pp. 25–33, 2013.
- [11] I. Benou et T. Riklin Raviv, “DeepTract: A Probabilistic Deep Learning Framework for White Matter Fiber Tractography,” dans *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, série Lecture Notes in Computer Science, D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, et A. Khan, éditeurs. Springer International Publishing, 2019, pp. 626–635.
- [12] O. Bernard *et al.*, “Standardized evaluation system for left ventricular segmentation algorithms in 3D echocardiography,” *IEEE transactions on medical imaging*, vol. 35, no. 4, pp. 967–977, 2016.
- [13] T. D. Bird, “Alzheimer Disease Overview,” dans *GeneReviews®*, M. P. Adam, H. H. Ardinger, R. A. Pagon, S. E. Wallace, L. J. Bean, G. Mirzaa, et A. Amemiya, éditeurs. University of Washington, Seattle, 1998.
- [14] C. M. Bishop, “Mixture density networks,” Aston University, Birmingham, Technical Report, 1994.
- [15] M. Boccardi *et al.*, “Delphi definition of the EADC-ADNI Harmonized Protocol for hippocampal segmentation on magnetic resonance,” *Alzheimer’s & Demenia*, vol. 11, no. 2, pp. 126–138, 2015.
- [16] C. Bycroft *et al.*, “The UK Biobank resource with deep phenotyping and genomic data,” *Nature*, vol. 562, no. 7726, pp. 203–209, octobre 2018.

BIBLIOGRAPHIE

- [17] M. W. Caan, H. G. Khedoe, D. H. Poot, J. Arjan, S. D. Olabarriaga, K. A. Grimbergen, L. J. Van Vliet, et F. M. Vos, “Estimation of diffusion properties in crossing fiber bundles,” *IEEE transactions on medical imaging*, vol. 29, no. 8, pp. 1504–1515, 2010.
- [18] A. Calamuneri *et al.*, “White Matter Tissue Quantification at Low b-Values Within Constrained Spherical Deconvolution Framework,” *Frontiers in Neurology*, vol. 9, p. 716, 2018.
- [19] M. Catani et M. T. De Schotten, “A diffusion tensor imaging tractography atlas for virtual in vivo dissections,” *Cortex*, vol. 44, no. 8, pp. 1105–1132, 2008.
- [20] Q. Chenot *et al.*, “A population-based atlas of the human pyramidal tract in 410 healthy participants,” *Brain Structure and Function*, pp. 1–14, 2018.
- [21] K. Cho, B. van Merriënboer, D. Bahdanau, et Y. Bengio, “On the Properties of Neural Machine Translation: Encoder–Decoder Approaches,” dans *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 2014, pp. 103–111.
- [22] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, et Y. Bengio, “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation,” *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014.
- [23] D. Christiaens, M. Reisert, T. Dhollander, S. Sunaert, P. Suetens, et F. Maes, “Global tractography of multi-shell diffusion-weighted imaging data using a multi-tissue model,” *Neuroimage*, vol. 123, pp. 89–101, 2015.
- [24] P. Coupe, P. Yger, S. Prima, P. Hellier, C. Kervrann, et C. Barillot, “An Optimized Blockwise Nonlocal Means Denoising Filter for 3-D Magnetic Resonance Images,” *IEEE Transactions on Medical Imaging*, vol. 27, no. 4, pp. 425–441, avril 2008.

BIBLIOGRAPHIE

- [25] A. Cárdenas-Blanco, C. Tejos, P. Irarrazaval, et I. Cameron, “Noise in magnitude magnetic resonance images,” *Concepts in Magnetic Resonance Part A: An Educational Journal*, vol. 32, no. 6, pp. 409–416, 2008.
- [26] M.-A. Côté, “ISMRM 2015 Tractography challenge,” 2015. Disponible à tractometer.org
- [27] M.-A. Côté, G. Girard, A. Boré, E. Garyfallidis, J.-C. Houde, et M. Descoteaux, “Tractometer: Towards validation of tractography pipelines,” *Medical Image Analysis*, vol. 17, no. 7, pp. 844–857, octobre 2013.
- [28] A. Daducci, E. Caruyer, M. Descoteaux, J. Houde, et J. Thiran, “HARDI reconstruction challenge 2013,” dans *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI), San Francisco, CA*, 2013.
- [29] A. Daducci *et al.*, “Quantitative comparison of reconstruction methods for intra-voxel fiber recovery from diffusion MRI,” *IEEE transactions on medical imaging*, vol. 33, no. 2, pp. 384–399, février 2014.
- [30] A. Daducci, A. Dal Palù, A. Lemkadem, et J.-P. Thiran, “COMMIT: Convex Optimization Modeling for Microstructure Informed Tractography,” *IEEE Transactions on Medical Imaging*, vol. 34, no. 1, pp. 246–257, janvier 2015.
- [31] M. Dayan, E. Monohan, S. Pandya, A. Kuceyeski, T. D. Nguyen, A. Raj, et S. A. Gauthier, “Profilometry: a new statistical framework for the characterization of white matter pathways, with application to multiple sclerosis,” *Human brain mapping*, vol. 37, no. 3, pp. 989–1004, 2016.
- [32] V. DeLuca et C. Pliatsikas, “Bilingualism and the brain,” OpenNeuro, Dataset ds001796, version 1.4.1, 2020.
- [33] V. DeLuca, J. Rothman, E. Bialystok, et C. Pliatsikas, “Redefining bilingualism as a spectrum of experiences that differentially affects brain structure and function,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 15, pp. 7565–7574, avril 2019.

BIBLIOGRAPHIE

- [34] M. Descoteaux, “High angular resolution diffusion MRI: from local estimation to segmentation and tractography,” Thèse de doctorat, Université Nice Sophia Antipolis, 2008.
- [35] M. Descoteaux, E. Angelino, S. Fitzgibbons, et R. Deriche, “Apparent diffusion coefficients from high angular resolution diffusion imaging: Estimation and applications,” *Magnetic Resonance in Medicine*, vol. 56, no. 2, pp. 395–410, 2006.
- [36] M. Descoteaux, E. Angelino, S. Fitzgibbons, et R. Deriche, “Regularized, fast, and robust analytical Q-ball imaging,” *Magnetic Resonance in Medicine*, vol. 58, no. 3, pp. 497–510, 2007.
- [37] M. Descoteaux, R. Deriche, T. R. Knösche, et A. Anwander, “Deterministic and probabilistic tractography based on complex fibre orientation distributions.” *IEEE transactions on medical imaging*, vol. 28, no. 2, pp. 269–86, février 2009.
- [38] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, et C. Notredame, “Nextflow enables reproducible computational workflows,” *Nature Biotechnology*, vol. 35, no. 4, pp. 316–319, avril 2017.
- [39] L. R. Dice, “Measures of the Amount of Ecologic Association Between Species,” *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.
- [40] D. G. Duru et M. Ozkan, “Self-organizing maps for brain tractography in MRI,” dans *Neural Engineering (NER), 2013 6th International IEEE/EMBS Conference on*. IEEE, 2013, pp. 1509–1512.
- [41] D. G. Duru et M. Ozkan, “SOM Based Diffusion Tensor MR Analysis,” dans *Image and Signal Processing and Analysis, 2007. ISPA 2007. 5th International Symposium on*. IEEE, 2007, pp. 403–406.
- [42] T. B. Dyrby, H. Lundell, M. W. Burke, N. L. Reislev, O. B. Paulson, M. Ptito, et H. R. Siebner, “Interpolation of diffusion weighted imaging datasets,” *NeuroImage*, vol. 103, pp. 202–213, décembre 2014.
- [43] B. Efron et R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994.

BIBLIOGRAPHIE

- [44] J. J. Entis, P. Doerga, L. F. Barrett, et B. C. Dickerson, “A reliable protocol for the manual segmentation of the human amygdala and its subregions using ultra-high resolution MRI,” *Neuroimage*, vol. 60, no. 2, pp. 1226–1235, 2012.
- [45] S. Eriksson, S. Lasic, et D. Topgaard, “Isotropic diffusion weighting in PGSE NMR by magic-angle spinning of the q-vector.” *J Magn Reson*, vol. 226, 2013.
- [46] W. I. Essayed, F. Zhang, P. Unadkat, G. R. Cosgrove, A. J. Golby, et L. J. O’Donnell, “White matter tractography for neurosurgical planning: A topography-based review of the current state of the art,” *NeuroImage: Clinical*, vol. 15, pp. 659–672, 2017.
- [47] S. Farquharson, J.-D. Tournier, F. Calamante, G. Fabinyi, M. Schneider-Kolsky, G. D. Jackson, et A. Connelly, “White matter fiber tractography: why we need to move beyond DTI: Clinical article,” *Journal of Neurosurgery*, vol. 118, no. 6, pp. 1367–1377, juin 2013.
- [48] P. Fillard *et al.*, “Quantitative evaluation of 10 tractography algorithms on a realistic diffusion MR phantom,” *Neuroimage*, vol. 56, no. 1, pp. 220–234, 2011.
- [49] E. Garyfallidis, M. Brett, M. M. Correia, G. B. Williams, et I. Nimmo-Smith, “QuickBundles, a Method for Tractography Simplification,” *Frontiers in Neuroscience*, vol. 6, 2012.
- [50] E. Garyfallidis, M. Brett, B. Amirbekian, A. Rokem, S. Van Der Walt, M. Descoteaux, et I. Nimmo-Smith, “Dipy, a library for the analysis of diffusion MRI data,” *Frontiers in Neuroinformatics*, vol. 8, 2014.
- [51] E. Garyfallidis, M. Zucchelli, J. Houde, et M. Descoteaux, “How to perform best odf reconstruction from the human connectome project sampling scheme,” dans *Proc. Intl. Soc. Mag. Reson. Med*, 2014.
- [52] E. Garyfallidis *et al.*, “Recognition of white matter bundles using local and global streamline-based registration and clustering,” *NeuroImage*, vol. 170, pp. 283–295, avril 2018.

BIBLIOGRAPHIE

- [53] G. Girard, K. Whittingstall, R. Deriche, et M. Descoteaux, “Towards quantitative connectivity analysis: reducing tractography biases,” *NeuroImage*, vol. 98, pp. 266–278, septembre 2014.
- [54] M. F. Glasser *et al.*, “The minimal preprocessing pipelines for the Human Connectome Project,” *NeuroImage*, vol. 80, pp. 105–124, octobre 2013.
- [55] H. Goel, I. Melnyk, et A. Banerjee, “R2N2: Residual Recurrent Neural Networks for Multivariate Time Series Forecasting,” arXiv e-prints, p. arXiv:1709.03159, septembre 2017.
- [56] I. Goodfellow, Y. Bengio, et A. Courville, *Deep Learning*. MIT Press, 2016.
- [57] A. Graves, “Generating Sequences With Recurrent Neural Networks,” *arXiv e-prints*, p. arXiv:1308.0850, août 2013.
- [58] V. Gupta, S. I. Thomopoulos, F. M. Rashid, et P. M. Thompson, “FiberNET: An Ensemble Deep Learning Framework for Clustering White Matter Fibers,” dans *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, série Lecture Notes in Computer Science, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, et S. Duchesne, éditeurs. Springer International Publishing, 2017, pp. 548–555.
- [59] V. Gupta, S. I. Thomopoulos, C. K. Corbin, F. Rashid, et P. M. Thompson, “FIBERNET 2.0: An automatic neural network based tool for clustering white matter fibers in the brain,” dans *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, avril 2018, pp. 708–711.
- [60] S. Hochreiter et J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, novembre 1997.
- [61] S. Hochreiter, Y. Bengio, P. Frasconi, et J. Schmidhuber, “Gradient flow in recurrent nets: the difficulty of learning long-term dependencies,” dans *A field guide to dynamical recurrent neural networks*, J. F. Kolen et S. C. Kremer, éditeurs. IEEE Press, 2001.

BIBLIOGRAPHIE

- [62] J.-C. Houde, M.-A. Côté-Harnois, et M. Descoteaux, “How to avoid biased streamlines-based metrics for streamlines with variable step sizes,” *Proceedings of: International Society of Magnetic Resonance in Medicine (ISMRM), (Toronto, ON)*, 2015.
- [63] S. Jbabdi, M. W. Woolrich, J. L. R. Andersson, et T. E. J. Behrens, “A Bayesian framework for global tractography,” *NeuroImage*, vol. 37, no. 1, pp. 116–129, août 2007.
- [64] B. Jeurissen, J.-D. Tournier, T. Dhollander, A. Connelly, et J. Sijbers, “Multi-tissue constrained spherical deconvolution for improved analysis of multi-shell diffusion MRI data,” *NeuroImage*, vol. 103, pp. 411–426, décembre 2014.
- [65] B. Jeurissen, M. Descoteaux, S. Mori, et A. Leemans, “Diffusion MRI fiber tractography of the brain,” *NMR in Biomedicine*, vol. 32, no. 4, p. e3785, 2019.
- [66] R. R. Jha, S. Patil, A. Nigam, et A. Bhavsar, “FS2Net: Fiber Structural Similarity Network (FS2Net) for Rotation Invariant Brain Tractography Segmentation Using Stacked LSTM Based Siamese Network,” dans *Computer Analysis of Images and Patterns*, série Lecture Notes in Computer Science, M. Vento et G. Percannella, éditeurs. Springer International Publishing, 2019, pp. 459–469.
- [67] D. Jorgens, O. Smedby, et R. Moreno, “Learning a Single Step of Streamline Tractography Based on Neural Networks,” dans *Computational Diffusion MRI*. Springer, 2018, pp. 103–116.
- [68] S. Kaufman, S. Rosset, C. Perlich, et O. Stitelman, “Leakage in data mining: Formulation, detection, and avoidance,” *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 6, no. 4, p. 15, 2012.
- [69] J. S. Kim *et al.*, “Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a Laplacian map and partial volume effect classification,” *NeuroImage*, vol. 27, no. 1, pp. 210–221, août 2005.
- [70] J. Kleesiek *et al.*, “Virtual raters for reproducible and objective assessments in radiology,” *Scientific reports*, vol. 6, p. 25007, 2016.

BIBLIOGRAPHIE

- [71] S. Koppers et D. Merhof, “Direct Estimation of Fiber Orientations Using Deep Learning in Diffusion Imaging,” dans *Machine Learning in Medical Imaging*, série Lecture Notes in Computer Science, L. Wang, E. Adeli, Q. Wang, Y. Shi, et H.-I. Suk, éditeurs. Springer International Publishing, 2016, pp. 53–60.
- [72] D. Krueger *et al.*, “Zoneout: Regularizing RNNs by Randomly Preserving Hidden Activations,” dans *5th International Conference on Learning Representations, ICLR 2017*. OpenReview.net, 2017.
- [73] G. M. Kurtzer, V. Sochat, et M. W. Bauer, “Singularity: Scientific containers for mobility of compute,” *PLOS ONE*, vol. 12, no. 5, p. e0177459, mai 2017.
- [74] D. Le Bihan, “Molecular diffusion nuclear magnetic resonance imaging,” *Magnetic Resonance Quarterly*, vol. 7, no. 1, pp. 1–30, janvier 1991.
- [75] D. Le Bihan et E. Breton, “Imagerie de diffusion in-vivo par résonance magnétique nucléaire,” *Comptes-Rendus de l’Académie des Sciences*, vol. 93, no. 5, pp. 27–34, 1985.
- [76] Y. LeCun, Y. Bengio, et G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, 2015.
- [77] J. H. Legarreta, L. Petit, F. Rheault, G. Theaud, C. Lemaire, M. Descoteaux, et P.-M. Jodoin, “Filtering in tractography using autoencoders (FINTA),” *Medical Image Analysis*, vol. 72, p. 102126, août 2021.
- [78] J. Lei Ba, J. R. Kiros, et G. E. Hinton, “Layer Normalization,” arXiv e-prints, p. arXiv:1607.06450, juillet 2016.
- [79] A. Lemkaddem, D. Skiöldebrand, A. Dal Palú, J.-P. Thiran, et A. Daducci, “Global tractography with embedded anatomical priors for quantitative connectivity analysis,” *Frontiers in Neurology*, vol. 5, p. 232, 2014.
- [80] O. A. S. d. Lucena, “Deep learning for brain analysis in MR imaging,” Text, Université de Campinas, Campinas, Brésil, 2018.

BIBLIOGRAPHIE

- [81] K. Maier-Hein *et al.*, “Tractography-based connectomes are dominated by false-positive connections,” *bioRxiv*, 2016.
- [82] K. H. Maier-Hein *et al.*, “The challenge of mapping the human connectome based on diffusion tractography,” *Nature Communications*, vol. 8, no. 1, p. 1349, novembre 2017.
- [83] J. G. Malcolm, M. E. Shenton, et Y. Rathi, “Filtered multitensor tractography,” *IEEE transactions on medical imaging*, vol. 29, no. 9, pp. 1664–1675, 2010.
- [84] J. F. Mangin, P. Fillard, Y. Cointepas, D. Le Bihan, V. Frouin, et C. Poupon, “Toward global tractography,” *NeuroImage*, vol. 80, pp. 290–296, octobre 2013.
- [85] J. V. Manjón, P. Coupé, L. Concha, A. Buades, D. L. Collins, et M. Robles, “Diffusion Weighted Image Denoising Using Overcomplete Local PCA,” *PLOS ONE*, vol. 8, no. 9, pp. 1–12, septembre 2013.
- [86] B. Mazoyer *et al.*, “BIL&GIN: a neuroimaging, cognitive, behavioral, and genetic database for the study of human brain lateralization,” *Neuroimage*, vol. 124, pp. 1225–1231, 2016.
- [87] B. H. Menze *et al.*, “The multimodal brain tumor image segmentation benchmark (BRATS),” *IEEE transactions on medical imaging*, vol. 34, no. 10, p. 1993, 2015.
- [88] R. Mito *et al.*, “Fibre-specific white matter reductions in Alzheimer’s disease and mild cognitive impairment,” *Brain*, vol. 141, no. 3, pp. 888–902, mars 2018.
- [89] S. Mori, B. J. Crain, V. P. Chacko, et P. Van Zijl, “Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging,” *Annals of neurology*, vol. 45, no. 2, pp. 265–269, 1999.
- [90] S. Mori, S. Wakana, P. C. Van Zijl, et L. Nagae-Poetscher, *MRI atlas of human white matter*. Elsevier, 2005.
- [91] P. F. Neher, F. B. Laun, B. Stieltjes, et K. H. Maier-Hein, “Fiberfox: Facilitating the creation of realistic white matter software phantoms,” *Magnetic Resonance in Medicine*, vol. 72, no. 5, pp. 1460–1470, 2014.

BIBLIOGRAPHIE

- [92] P. F. Neher, G. Michael, T. Norajitra, C. Weber, et K. H. Maier-hein, “A Machine Learning Based Approach to Fiber Tractography Using Classifier Voting,” *Medical Image Computing and Computer-Assisted Intervention*, vol. 9349, pp. 45–52, 2015.
- [93] P. F. Neher, M.-A. Côté, J.-C. Houde, M. Descoteaux, et K. H. Maier-Hein, “Fiber tractography using machine learning,” *NeuroImage*, vol. 158, pp. 417–429, 2017.
- [94] P. D. Ngattai Lam, G. Belhomme, J. Ferrall, B. Patterson, M. Styner, et J. C. Prieto, “TRAFIC: Fiber Tract Classification Using Deep Learning.” *Proceedings of SPIE—the International Society for Optical Engineering*, vol. 10574, février 2018.
- [95] G. Nilsonne *et al.*, “The Stockholm Sleepy Brain Study: Effects of Sleep Deprivation on Cognitive and Emotional Processing in Young and Old,” *OpenNeuro*, 2020.
- [96] L. Ning *et al.*, “Cross-scanner and cross-protocol multi-shell diffusion MRI data harmonization: Algorithms and results,” *NeuroImage*, vol. 221, p. 117128, novembre 2020.
- [97] E. Ozarslan et T. H. Mareci, “Generalized diffusion tensor imaging and analytical relationships between diffusion tensor imaging and high angular resolution diffusion imaging,” *Magnetic Resonance in Medicine*, vol. 50, no. 5, pp. 955–965, novembre 2003.
- [98] F. Piccinini, A. Tesei, G. Paganelli, W. Zoli, et A. Bevilacqua, “Improving reliability of live/dead cell counting through automated image mosaicing,” *Computer methods and programs in biomedicine*, vol. 117, no. 3, pp. 448–463, 2014.
- [99] C. Pierpaoli, P. Jezzard, P. J. Basser, A. Barnett, et G. Di Chiro, “Diffusion tensor MR imaging of the human brain.” *Radiology*, vol. 201, no. 3, pp. 637–648, 1996.

BIBLIOGRAPHIE

- [100] R. A. Poldrack *et al.*, “A phenome-wide examination of neural and cognitive function,” *Scientific Data*, vol. 3, no. 1, p. 160110, décembre 2016.
- [101] P. Poulin, M.-A. Côté, J.-C. Houde, L. Petit, P. F. Neher, K. H. Maier-Hein, H. Larochelle, et M. Descoteaux, “Learn to Track: Deep Learning for Tractography,” dans *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, série Lecture Notes in Computer Science, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, et S. Duchesne, éditeurs. Springer International Publishing, 2017, pp. 540–547.
- [102] P. Poulin, F. Rheault, E. St-Onge, P.-M. Jodoin, et M. Descoteaux, “Bundle-Wise Deep Tracker: Learning to track bundle-specific streamline paths,” dans *Proceedings of the International Society for Magnetic Resonance in Medicine*. ISMRM-ESMRMB, 2018.
- [103] P. Poulin, D. Jörgens, P.-M. Jodoin, et M. Descoteaux, “Tractography and machine learning: Current state and open challenges,” *Magnetic Resonance Imaging*, vol. 64, pp. 37–48, décembre 2019.
- [104] P. Poulin *et al.*, “TractoInferno: A large-scale, open-source, multi-site database for machine learning dMRI tractography,” *bioRxiv*, p. 2021.11.29.470422, 2021.
- [105] C. Poupon, L. Laribiere, G. Tournier, J. Bernard, D. Fournier, P. Fillard, M. Descoteaux, et J. Mangin, “A diffusion hardware phantom looking like a coronal brain slice,” dans *Proceedings of the International Society for Magnetic Resonance in Medicine*, vol. 18, 2010, p. 581.
- [106] C. Presseau, P.-M. Jodoin, J.-C. Houde, et M. Descoteaux, “A new compression format for fiber tracking datasets,” *NeuroImage*, vol. 109, pp. 73–83, avril 2015.
- [107] S. Pujol *et al.*, “The DTI challenge: toward standardized evaluation of diffusion tensor imaging tractography for neurosurgery,” *Journal of Neuroimaging*, vol. 25, no. 6, pp. 875–882, 2015.
- [108] D. Raffelt, T. Dhollander, J.-D. Tournier, R. Tabbara, R. E. Smith, E. Pierre, et A. Connelly, “Bias field correction and intensity normalisation for quantitative

BIBLIOGRAPHIE

- analysis of apparent fibre density,” dans *Proc. Intl. Soc. Mag. Reson. Med*, vol. 25, 2017, p. 3541.
- [109] S. J. Raudys et A. K. Jain, “Small sample size effects in statistical pattern recognition: Recommendations for practitioners,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 3, pp. 252–264, 1991.
- [110] C. P. Reddy et Y. Rathi, “Joint multi-fiber NODDI parameter estimation and tractography using the unscented information filter,” *Frontiers in neuroscience*, vol. 10, p. 166, 2016.
- [111] M. Reisert, I. Mader, C. Anastasopoulos, M. Weigel, S. Schnell, et V. Kiselev, “Global fiber reconstruction becomes practical,” *NeuroImage*, vol. 54, no. 2, pp. 955–962, janvier 2011.
- [112] M. Reisert, V. A. Coenen, C. Kaller, K. Egger, et H. Skibbe, “HAMLET: Hierarchical Harmonic Filters for Learning Tracts from Diffusion MRI,” *arXiv e-prints*, p. arXiv:1807.01068, juillet 2018.
- [113] C. Reveley, A. K. Seth, C. Pierpaoli, A. C. Silva, D. Yu, R. C. Saunders, D. A. Leopold, et F. Q. Ye, “Superficial white matter fiber systems impede detection of long-range cortical connections in diffusion MR tractography,” *Proceedings of the National Academy of Sciences*, vol. 112, no. 21, pp. E2820–E2828, mai 2015.
- [114] F. Rheault, J.-C. Houde, et M. Descoteaux, “Real time interaction with millions of streamlines,” *Proceedings of: International Society of Magnetic Resonance in Medicine (ISMRM) (Toronto, ON)*, 2015.
- [115] F. Rheault, E. St-Onge, J. Sidhu, K. Maier-Hein, N. Tzourio-Mazoyer, L. Petit, et M. Descoteaux, “Bundle-specific tractography with incorporated anatomical and orientational priors,” *NeuroImage*, vol. 186, pp. 382–398, 2019.
- [116] F. Rheault *et al.*, “Tractostorm: The what, why, and how of tractography dissection reproducibility,” *Human Brain Mapping*, vol. 41, no. 7, pp. 1859–1874, 2020.

BIBLIOGRAPHIE

- [117] F. Rheault, P. Poulin, A. Valcourt Caron, E. St-Onge, et M. Descoteaux, “Common misconceptions, hidden biases and modern challenges of dMRI tractography,” *Journal of Neural Engineering*, vol. 17, no. 1, p. 011001, février 2020.
- [118] F. Rheault, “Analyse et reconstruction de faisceaux de la matière blanche,” Thèse de doctorat, Université de Sherbrooke, 2020.
- [119] O. Ronneberger, P. Fischer, et T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” dans *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, série Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. Wells, et A. F. Frangi, éditeurs. Springer International Publishing, 2015, pp. 234–241.
- [120] T. Sarwar, K. Ramamohanarao, et A. Zalesky, “Mapping connectomes with diffusion MRI: deterministic or probabilistic tractography?” *Magnetic Resonance in Medicine*, vol. 81, no. 2, pp. 1368–1384, février 2019.
- [121] T. Sarwar, K. Ramamohanarao, et A. Zalesky, “A critical review of connectome validation studies,” *NMR in Biomedicine*, vol. 34, no. 12, p. e4605, 2021, _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/nbm.4605>.
- [122] K. G. Schilling *et al.*, “Challenges in diffusion MRI tractography – Lessons learned from international benchmark competitions,” *Magnetic Resonance Imaging*, vol. 57, pp. 194–209, avril 2019.
- [123] T. Sherif *et al.*, “CBRAIN: a web-based, distributed computing platform for collaborative neuroimaging research,” *Frontiers in Neuroinformatics*, vol. 8, p. 54, 2014.
- [124] V. Siless, K. Chang, B. Fischl, A. Yendiki, et Feb, “AnatomiCuts: Hierarchical clustering of tractography streamlines based on anatomical similarity,” *NeuroImage*, vol. 166, pp. 32–45,, 2018.
- [125] R. Smith, J.-D. Tournier, F. Calamante, A. Connelly, et Feb, “SIFT: Spherical-deconvolution informed filtering of tractograms,” *NeuroImage*, vol. 67, pp. 298–312,, 2013.

BIBLIOGRAPHIE

- [126] R. Smith, J.-D. Tournier, F. Calamante, et A. Connelly, “SIFT2: Enabling dense quantitative assessment of brain white matter connectivity using streamlines tractography,” *NeuroImage*, vol. 119, pp. 338–351,, octobre 2015.
- [127] R. E. Smith, J.-D. Tournier, F. Calamante, et A. Connelly, “Anatomically-constrained tractography: Improved diffusion MRI streamlines tractography through effective use of anatomical information,” *NeuroImage*, vol. 62, no. 3, pp. 1924–1938, septembre 2012.
- [128] S. M. Smith, “Fast robust automated brain extraction,” *Human Brain Mapping*, vol. 17, no. 3, pp. 143–155, 2002.
- [129] E. St-Onge, A. Daducci, G. Girard, et M. Descoteaux, “Surface-enhanced tractography (SET),” *NeuroImage*, vol. 169, pp. 524–539, avril 2018.
- [130] B. Stieltjes, R. M. Brunner, K. Fritzsche, et F. Laun, *Diffusion tensor imaging: introduction and atlas.* Springer Science & Business Media, 2013.
- [131] I. Sutskever, O. Vinyals, et Q. V. Le, “Sequence to sequence learning with neural networks,” dans *Adv. in neural information processing systems*, 2014, pp. 3104–3112.
- [132] A. A. Taha et A. Hanbury, “Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool,” *BMC medical imaging*, vol. 15, no. 1, p. 29, 2015.
- [133] S. Tamm *et al.*, “The effect of sleep restriction on empathy for pain: An fMRI study in younger and older adults,” *Scientific Reports*, vol. 7, no. 1, p. 12236, septembre 2017.
- [134] G. Theaud, J.-C. Houde, A. Boré, F. Rheault, F. Morency, et M. Descoteaux, “TractoFlow: A robust, efficient and reproducible diffusion MRI pipeline leveraging Nextflow & Singularity,” *NeuroImage*, vol. 218, p. 116889, septembre 2020.

BIBLIOGRAPHIE

- [135] C. Thomas, Q. Y. Frank, M. O. Irfanoglu, P. Modi, K. S. Saleem, D. A. Leopold, et C. Pierpaoli, “Anatomical accuracy of brain connections derived from diffusion MRI tractography is inherently limited,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 46, pp. 16 574–16 579, 2014.
- [136] A. Thon, U. Teichgräber, C. Tennstedt-Schenk, S. Hadjidemetriou, S. Winzler, A. Malich, et I. Papageorgiou, “Computer aided detection in prostate cancer diagnostics: A promising alternative to biopsy? A retrospective study from 104 lesions with histological ground truth,” *PLOS ONE*, vol. 12, no. 10, pp. 1–21, octobre 2017.
- [137] A. Théberge, C. Desrosiers, M. Descoteaux, et P.-M. Jodoin, “Track-to-Learn: A general framework for tractography with deep reinforcement learning,” *Medical Image Analysis*, vol. 72, p. 102093, août 2021.
- [138] J. D. Tournier, F. Calamante, D. G. Gadian, et A. Connelly, “Direct estimation of the fiber orientation density function from diffusion-weighted MRI data using spherical deconvolution,” *NeuroImage*, vol. 23, no. 3, pp. 1176–1185, novembre 2004.
- [139] J.-D. Tournier, F. Calamante, et A. Connelly, “Robust determination of the fibre orientation distribution in diffusion MRI: Non-negativity constrained super-resolved spherical deconvolution,” *NeuroImage*, vol. 35, no. 4, pp. 1459–1472, mai 2007.
- [140] J. D. Tournier, F. Calamante, et A. Connelly, “Improved probabilistic streamlines tractography by 2nd order integration over fibre orientation distributions,” dans *Proceedings of the international society for magnetic resonance in medicine*, vol. 18, 2010, p. 1670.
- [141] J.-D. Tournier, F. Calamante, et A. Connelly, “MRtrix: Diffusion tractography in crossing fiber regions,” *International Journal of Imaging Systems and Technology*, vol. 22, no. 1, pp. 53–66, 2012.
- [142] S. Tremblay, M. Desjardins, P. Bermudez, Y. Iturria-Medina, A. C. Evans, P. Jolicoeur, et L. De Beaumont, “Mild traumatic brain injury: The effect of

BIBLIOGRAPHIE

- age at trauma onset on brain structure integrity,” *NeuroImage: Clinical*, vol. 23, p. 101907, janvier 2019.
- [143] A. Tsuchida *et al.*, “The MRi-Share database: brain imaging in a cross-sectional cohort of 1,870 university students,” *bioRxiv*, p. 2020.06.17.154666, juin 2020.
- [144] D. S. Tuch, “Q-ball imaging,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 52, no. 6, pp. 1358–1372, 2004.
- [145] D. S. Tuch, T. G. Reese, M. R. Wiegell, N. Makris, J. W. Belliveau, et V. J. Wedeen, “High angular resolution diffusion imaging reveals intravoxel white matter fiber heterogeneity,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 48, no. 4, pp. 577–582, 2002.
- [146] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, et J. C. Gee, “N4ITK: Improved N3 Bias Correction,” *IEEE Transactions on Medical Imaging*, vol. 29, no. 6, pp. 1310–1320, juin 2010.
- [147] A. van den Oord *et al.*, “WaveNet: A Generative Model for Raw Audio,” *arXiv e-prints*, p. arXiv:1609.03499, septembre 2016.
- [148] D. C. Van Essen *et al.*, “The Human Connectome Project: a data acquisition perspective,” *Neuroimage*, vol. 62, no. 4, pp. 2222–2231, 2012.
- [149] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. J. Behrens, E. Yacoub, et K. Ugurbil, “The WU-Minn Human Connectome Project: An overview,” *NeuroImage*, vol. 80, pp. 62–79, octobre 2013.
- [150] D. Varadarajan et J. P. Haldar, “A majorize-minimize framework for Rician and non-central chi MR images,” *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 2191–2202, 2015.
- [151] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, et I. Polosukhin, “Attention is All you Need,” dans *Advances in*

BIBLIOGRAPHIE

- Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et R. Garnett, éditeurs, vol. 30. Curran Associates, Inc., 2017.
- [152] L. Vavassori, S. Sarubbo, et L. Petit, “Hodology of the superior longitudinal system of the human brain: a historical perspective, the current controversies, and a proposal,” *Brain Structure and Function*, vol. 226, no. 5, pp. 1363–1384, juin 2021.
 - [153] J. Veraart, D. S. Novikov, D. Christiaens, B. Ades-aron, J. Sijbers, et E. Fieremans, “Denoising of diffusion MRI using random matrix theory,” *NeuroImage*, vol. 142, pp. 394–406, novembre 2016.
 - [154] D. Wassermann, N. Makris, Y. Rathi, M. Shenton, R. Kikinis, M. Kubicki, et C.-F. Westin, “The white matter query language: a novel approach for describing human white matter anatomy,” *Brain Structure and Function*, vol. 221, no. 9, pp. 4705–4721, 2016.
 - [155] J. Wasserthal, P. Neher, et K. H. Maier-Hein, “TractSeg-Fast and accurate white matter tract segmentation,” *NeuroImage*, vol. 183, pp. 239–253, 2018.
 - [156] J. Wasserthal, P. Neher, et K. H. Maier-Hein, “TractSeg - Fast and accurate white matter tract segmentation,” *NeuroImage*, vol. 183, pp. 239–253, décembre 2018.
 - [157] J. Wasserthal, P. F. Neher, et K. H. Maier-Hein, “Tract orientation mapping for bundle-specific tractography,” dans *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 36–44.
 - [158] J. Wasserthal, P. F. Neher, D. Hirjak, et K. H. Maier-Hein, “Combined tract segmentation and orientation mapping for bundle-specific tractography,” *Medical Image Analysis*, vol. 58, p. 101559, décembre 2019.
 - [159] V. Wegmayr et J. M. Buhmann, “Entrack: Probabilistic Spherical Regression with Entropy Regularization for Fiber Tractography,” *International Journal of Computer Vision*, novembre 2020.

BIBLIOGRAPHIE

- [160] V. Wegmayr, G. Giuliani, S. Holdener, et J. Buhmann, “Data-driven fiber tractography with neural networks,” dans *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, avril 2018, pp. 1030–1033.
- [161] B. L. Welch, “The generalization of ‘student’s’ problem when several different population variances are involved,” *Biometrika*, vol. 34, no. 1/2, pp. 28–35, 1947.
- [162] C.-F. Westin *et al.*, “Q-space trajectory imaging for multidimensional diffusion MRI of the human brain,” *NeuroImage*, vol. 135, pp. 345–362, juillet 2016.
- [163] T. White *et al.*, “Global White Matter Abnormalities in Schizophrenia: A Multisite Diffusion Tensor Imaging Study,” *Schizophrenia Bulletin*, vol. 37, no. 1, pp. 222–232, janvier 2011.
- [164] B. Wilkins, N. Lee, et M. Singh, “Development and evaluation of a simulated FiberCup phantom,” dans *International Symposium on Magnetic Resonance in Medicine (ISMRM’12)*, 2012, p. 1938.
- [165] I. Wolf *et al.*, “The medical imaging interaction toolkit (MITK): a toolkit facilitating the creation of interactive software by extending VTK and ITK,” dans *Medical Imaging 2004: Visualization, Image-Guided Procedures, and Display*, vol. 5367. SPIE, mai 2004, pp. 16–27.
- [166] F.-C. Yeh, T. D. Verstynen, Y. Wang, J. C. Fernández-Miranda, et W.-Y. I. Tseng, “Deterministic diffusion fiber tracking improved by quantitative anisotropy,” *PloS one*, vol. 8, no. 11, p. e80713, 2013.
- [167] F.-C. Yeh, S. Panesar, D. Fernandes, A. Meola, M. Yoshino, J. C. Fernandez-Miranda, J. M. Vettel, et T. Verstynen, “Population-averaged atlas of the macroscale human structural connectome and its network topology,” *NeuroImage*, vol. 178, pp. 57–68, 2018.
- [168] A. Yendiki *et al.*, “Automated probabilistic reconstruction of white-matter pathways in health and disease using an atlas of the underlying anatomy,” *Frontiers in neuroinformatics*, vol. 5, p. 23, 2011.

BIBLIOGRAPHIE

- [169] F. Zhang, Y. Wu, I. Norton, L. Rigolo, Y. Rathi, N. Makris, et L. J. O'Donnell, “An anatomically curated fiber clustering white matter atlas for consistent white matter tract parcellation across the lifespan,” *NeuroImage*, vol. 179, pp. 429–447, octobre 2018.
- [170] Y. Zhang, M. Brady, et S. Smith, “Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm,” *IEEE transactions on medical imaging*, vol. 20, no. 1, pp. 45–57, janvier 2001.