# EMPIRICAL AND DATA-DRIVEN HARMONIZATION OF DIFFUSION WEIGHTED MRI

By

Colin Blake Hansen

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

August 31, 2021

Nashville, Tennessee

Approved:

Professor Bennett A. Landman, Ph.D.

Professor Catie Chang, Ph.D.

Professor Taylor Davis, M.D.

Professor Thomas Lasko, Ph.D.

Professor Jack Noble, Ph.D.

Professor Mikail Rubinov, Ph.D.

**ACKNOWLEDGMENTS**

I would like to thank my advisor Bennett Landman for teaching me how to conduct research and paving the way for me to finish my dissertation. Dr. Landman's experience and vision were invaluable to my work, and his guidance was of great help during my graduate career. Kurt Schilling has acted as an unofficial second advisor to me throughout my dissertation. I would like to thank Dr. Schilling for all the advice he has dispensed and the time he took to discuss projects and review papers. His relentless enthusiasm for science has been a great inspiration. I would also like to thank Baxter Rogers whose knowledge and expertise were vital to many projects in this dissertation. Dr. Rogers has always had an open door when I had questions and would take the time to help me understand difficult concepts. Thank you to the members of my committee, Dr. Catie Chang, Dr. Mikail Rubinov, Dr. L. Taylor Davis, Dr. Jack Noble, and Dr. Thomas Lasko, who provided sage wisdom and guidance during my last year of dissertation work. I would like to thank all the past and current members of the MASI Lab. In addition to those who have collaborated on this work, the members of this lab have created a great community which brought life to a room of monitors and computers. I would also like to thank the many friends who were there to celebrate the victories, provide distraction from the defeats, and enjoy the peaceful moments in between. I could always rely on my best friend Nathan Rugroden to provide a laugh, listen to a work story, or just waste a few hours with nonsense.

I would like to thank Roza Gunes Bayrak who always saw the best in me and was a calm and collected presence when I was under pressure. She has spent many hours discussing science, improving figures, and critiquing presentations which has significantly improved my work. We have spent nearly all our graduate careers together, and her adventurous spirit ensured that I took the time to enjoy a little of what the world has to offer. Lastly, I would like to thank my family. My parents Curtis and Kathy Hansen have always supported me and encouraged me to pursue whatever I wanted. Both taught me by example that determination and work ethic were necessary to achieve in life. My sisters, Miranda, Teresa, and Jessica have always been excited to hear about my research and have been great friends throughout my life. My grandparents Gene Sovereign, Judy Sovereign, Carroll Hansen, Madelon Hansen, and James Boullion have always been interested in my education and have been supportive in any way possible. Though Madelon and James are no longer here to witness my dissertation, I know they would have been proud.

**Table of Contents**

**List of Tables**

**List of Figures**

## Chapter I. Introduction

### 1. Overview:

Clinically, the primary application of diffusion weighted magnetic resonance imaging (DW-MRI) has been for early detection and characterization of cerebral ischemia. This has led to DW-MRI becoming the primary modality for the management of stroke patients [10-13] . DW-MRI is also being increasingly used in managing cancer patients [13-15]. In research DW-MRI is known for mapping white matter fibers of the brain and serves as the only available technique to probe tissue structure at a microscopic level in-vivo [16]. This has opened up new investigations into cognitive neuroscience and brain dysfunction in aging, addiction, mental health disorders, and neurological disease [17].

Water diffusion is anisotropic in the white matter regions of the brain because axon membranes limit molecular movement perpendicular to the fibers [18]. Diffusion tensor imaging (DTI) produces micro-architectural detail of white matter tracts by exploiting this property, and this provides information about white matter integrity [19]. This sensitivity to microstructural changes led researchers to using DTI alongside other modalities in longitudinal and connectomic based studies with a focus on characterizing the effects of neurological disorders [20-23]. However, statistical analysis of DW-MRI is held back by bias introduced by many factors. Variability in DW-MRI measurements can result from a difference in the number of head coils used, the sensitivity of the coils, the imaging gradient non-linearity, the magnetic field homogeneity, the differences in the algorithms used to reconstruct the data, as well as changes made during software upgrades [24-28]. Harmonization approaches and methods aim to increase reproducibility and reduce error caused by variance and bias introduced by hardware and site effects. Reproducibility is a closeness measure between a pair or group of measurements. For DW-MRI reproducibility is often evaluated between measurements of a phantom or a subject acquired at multiple sites or with multiple sets of acquisition parameters. Phantoms often have the benefit of a ground truth, so reproducibility measurements would not be necessary. For in-vivo human acquisitions, reproducibility is the best measure of harmonization.

Empirically derived models have made substantial strides in correcting for hardware specific effects [29-33] and have laid the groundwork for the harmonization field. Statistical models have been shown to be effective at harmonizing scalar and vector values [34, 35], and phantom work is progressing on isotropic [36] and temperature-controlled arrays [37] and biological mimics [38-40]. Modern regularized machine learning pushes harmonization

efforts towards data driven models such as in the CDMRI MUSHAC challenge [41]. Many deep learning methods, including convolutional and residual networks, have been proposed to solve this problem [42]. Most choose to learn from the spherical harmonic representation of the diffusion signal over a sphere or rotationally invariant features of these functions [43]. The algorithms are generally successful in harmonizing data across scanners showing improvements over simple linear spatial interpolations.

The focus of this thesis is on the harmonization of DW-MRI through both empirical and statistical means and how methods developed in this pursuit can influence other domains. The key areas in which we seek to improve the existing literature are the characterization and correction of spatially varying effects specific to DW-MRI hardware, the development of white matter bundle atlases and models, the use of these atlases and models in the harmonization of DW-MRI, and the use of a semi-supervised framework developed for harmonization for general classification tasks.

The first section in this chapter covers information about the imaging modalities used in the performed contributions or will be used in proposed contributions for the dissertation. The next covers the derived metrics that can be extracted from the observed DW-MRI signal and what microstructural information can be obtained and how it can be applied. The third section covers the general deep learning paradigm and how it has been applied to DW-MRI. The following section covers the current standard preprocessing steps for DW-MRI data as well as an overview of harmonization methods. The fifth section covers brain parcellation strategies. Last in this chapter, the contributions and current challenges in the field contained in this dissertation proposal are outlined.

**2. Neuroimaging**

Magnetic Resonance Imaging (MRI) systems use a powerful magnet to produce a strong magnetic field ($B_0$) which forces protons in the body to align with that field. By pulsing a radiofrequency current through the patient, the protons spin out of equilibrium. Depending on the environment and chemical nature of the molecules, the time it takes for the protons to realign with the magnetic field once the radiofrequency field is off and the amount of energy released from the proton's changes. These properties allow for differentiation between various types of tissues. This section discusses the MRI sequences and resulting signals of T1 weighted acquisitions and diffusion weighted acquisitions.

### 2.1. T1 Weighted MRI

T1 weighted images measure how quickly the net magnetization vector recovers to its ground state in the direction of the $B_0$ field (T1 relaxation time). Spins aligned in the $B_0$ field are put into a transverse plane by a radiofrequency pulse, and then move back toward the original equilibrium. A tissue's T1 reflects the amount of time its protons' spins realign with the $B_0$ field. T1 weighting tend of have short echo times (TE) which refers to the time between the application of the radiofrequency pulse and the peak of the signal induced in the coil and repetition times (TR) which refers to the time from the pulse to the application of the next pulse. T1 imaging provides high contrast where the there is a relatively higher fat content such as in white matter regions of the brain, and acute hemorrhage and provides low contrast for cerebral spinal fluid, bone, and air. As they are quick to acquire, T1 does not suffer from the effects of patient movement, and it is a standard clinical procedure to acquire a T1 weighted image before advanced sequences are acquired. As they provide the best gray/white contrast , T1 images are ideal for segmenting gray matter, white matter and cerebrospinal fluid regions [44]. Figure 1 shows a T1 acquisition and a resulting segmentation. Brain segmentation is further discussed in section 5.

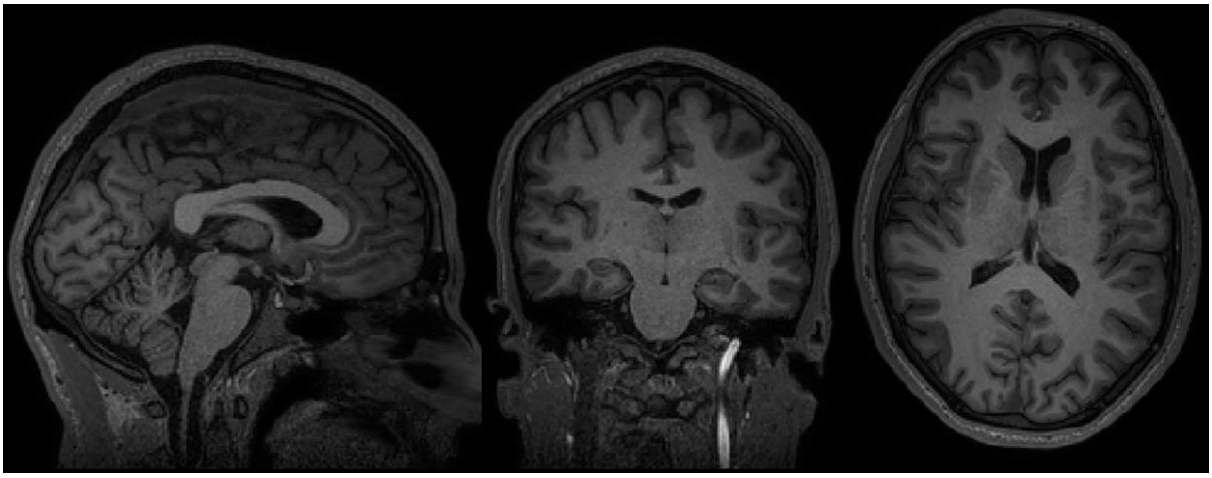

**Figure I-1. A sagittal, coronal, and axial slice (left to right) of a T1 weighted brain volume is shown. These acquisitions provide high resolution of anatomical structure (1mm isotropic in this image) with high signal to noise ratio. White matter regions tend to be brighter, cerebral spinal fluid regions have low intensity resembling the background, and gray matter regions tend to have mid-range intensities.**

### 2.2. Diffusion-Weighted MRI

The measurement of diffusion in the presence of a constant background gradient was outlined by Hahn [45]. Carr and Purcell further developed this [46], and the effects of diffusion in the presence of time varying magnetic field gradients was mathematically formalized by Torrey [47]. Today the most widely used diffusion pulse sequence is the

pulsed gradient spin echo proposed (PGSE) by Stejskal and Tanner [48]. For a PGSE sequence, the diffusion weighting imposed by the gradient pulses is determined by the gyromagnetic ratio ($\gamma$), gradient amplitude ($G$), duration ($\delta$), and separation ($\Delta$), with a b-value given by:

$$b = \gamma^2 \delta^2 G^2 \left( \Delta - \frac{\delta}{3} \right)$$

The diffusion weighted MRI signal is then related to the diffusion coefficient $D$ and the b-value by:

$$S = S_0 e^{-bD}$$

where $S_0$ is the signal in the absence of diffusion gradient pulses. The PGSE is widely used in the diffusion MRI community to probe diffusivity in a particular direction defined by a gradient vector and is employed in the diffusion weighted acquisitions in this thesis.

Einstein's equation for diffusion assumes free or isotropic diffusion where the distribution of molecular displacement obeys a Gaussian law [49]. The self-diffusion coefficient of free water is approximately $3.0 \times 10^{-9}$ $m^2/s$ at 37°C [50], but in biological tissue water molecules encounter barriers such as cell membranes, fibers, and macromolecules causing molecular displacements to deviate from a Gaussian distribution. Because the derived diffusion coefficient from DWIs is no longer the free diffusion coefficient of water, the derived measure is the apparent diffusion coefficient (ADC). In brain tissue, the rate of diffusion for a given molecule may depend on the direction of diffusion which is termed anisotropic diffusion such as it is in white matter [51]. It was found that diffusion was typically fast in the direction of neuronal fibers and slower perpendicular to them as diffusion was hindered by the myelin sheath and axon membranes [52]. This led to the suggestion that diffusion directional specificity could be used to determine and map the orientation of white matter fibers in the brain.

DW-MRI is the only non-invasive modality to probe *in vivo* tissue micro-structure and macro-structure [16]. By sensitizing the MR signal to the Brownian motion of water molecules in directions on a sphere, the micro-architecture within the brain can be reconstructed from the signal attenuation across diffusion volumes [53]. Each diffusion volume is acquired with a specified b-value which reflects the strength and timing of the gradients and a unique gradient vector which defines the direction at which the pulsed gradient spin echo sequence is acquired. Figure 2 shows a volume with no diffusion weighting and a volume with a diffusion weighting of 1000 $s/mm^2$.

## 3. Microstructural Measures

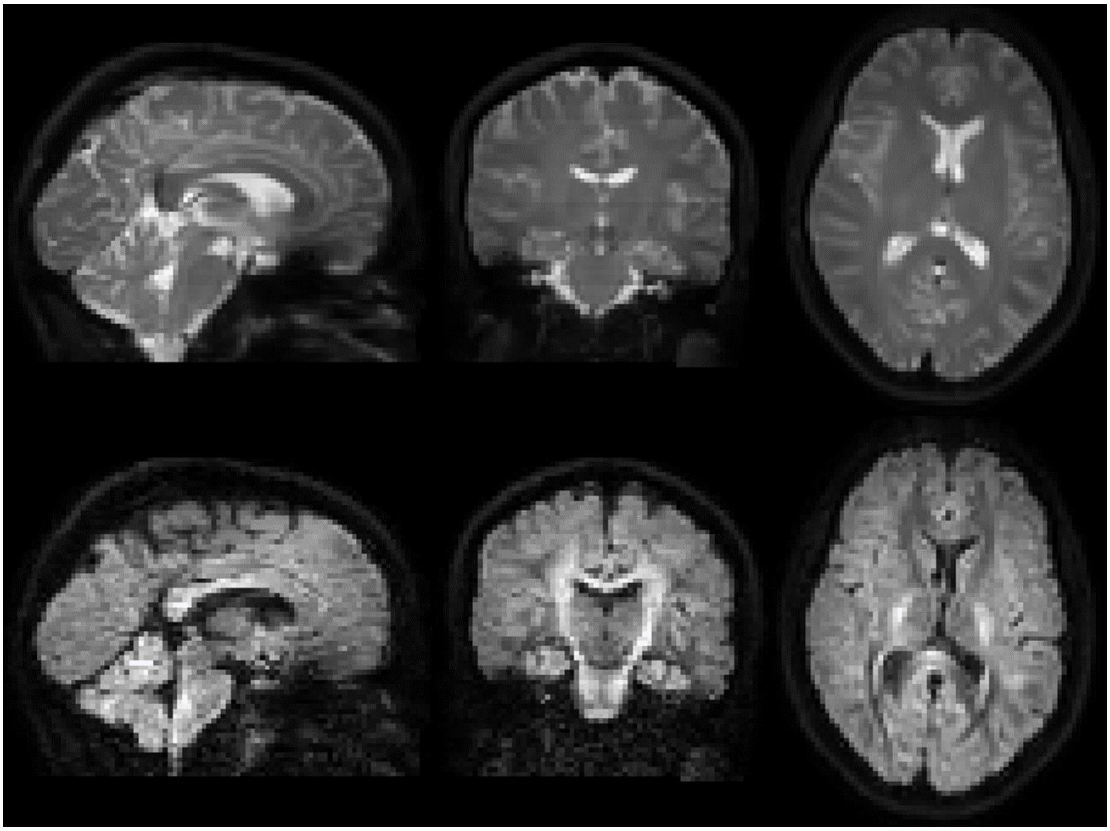

**Figure I-2. A sagittal, coronal, and axial slice (left to right) of a non-diffusion weighted brain volume (top) and a diffusion weighted volume at a b-value of 1000 s/mm$^2$ (bottom) are shown here. DW-MRI in a typical clinical acquisition have lower resolution than a typical T1 image. Here the resolution is 2mm isotropic. The non-diffusion weighted volume or b0 shows higher intensity for cerebral spinal fluid and lower intensity for white matter regions. In a diffusion weighted volume, certain white matter structures may be visible depending on the diffusion direction, but representations which consider all diffusion directions are more informative.**

### 3.1. Diffusion Tensor Imaging

Diffusion tensor imaging (DTI) is perhaps the most widely used model for estimating 3D white matter microstructure orientation within a voxel [53, 54]. DTI models diffusion as a zero-mean Gaussian distribution using a rank-2 symmetric positive definite tensor:

$$D = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{xy} & D_{yy} & D_{yz} \\ D_{xz} & D_{yz} & D_{zz} \end{bmatrix}$$

which replaces the 1D diffusion coefficient:

$$S = S_0 e^{-b:D}$$

Diagonalization of the diffusion tensor gives us the eigenvalues ($\lambda_1$, $\lambda_2$, $\lambda_3$) and corresponding eigenvectors ($v_1$, $v_2$, $v_3$) which describe the directions and apparent diffusivities along the axes of principal diffusion. The diffusion tensor can be visualized as an ellipsoid where the eigenvectors define the direction and the eigenvalues define the lengths of the semi-major axes. The diffusion tensor describes the magnitude, the degree of anisotropy, and the orientation of diffusion anisotropy. Estimates of white matter connectivity patterns in the brain from white matter tractography may be obtained using the diffusion anisotropy and the principal diffusion directions [55]. From DTI we get useful metrics such as mean diffusivity (MD):

$$MD = \frac{\lambda_1 + \lambda_2 + \lambda_3}{3}$$

which is a scalar measure of the total diffusion within a voxel and fractional anisotropy (FA):

$$FA = \sqrt{\frac{(\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_1 - \lambda_3)^2}{2\left({\lambda_1}^2 + {\lambda_2}^2 + {\lambda_3}^2\right)}}$$

which describes the total anisotropy within a voxel. The MD and FA for a subject is shown in Figure 3.

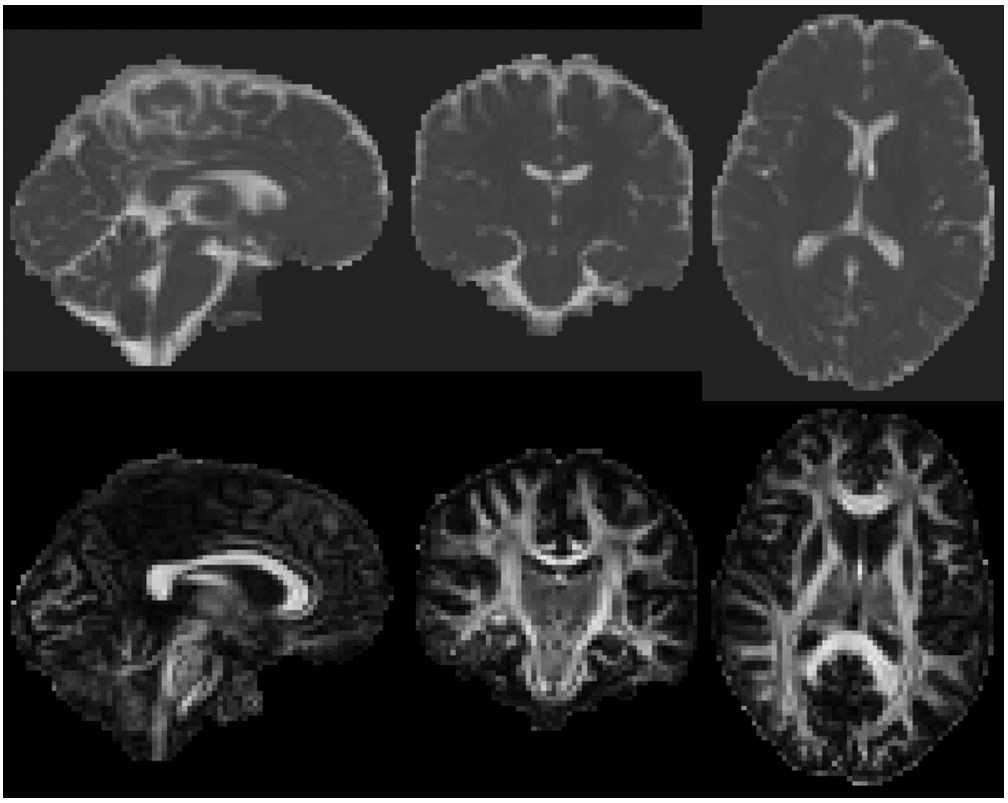

**Figure I-3. A sagittal, coronal, and axial slice (left to right) the estimated mean diffusivity (MD) (top) and fractional anisotropy (FA) (bottom). By constructing and diagonalizing the diffusion tensor for each voxel in a DW-MRI, the eigenvalues corresponding to the x, y, and z directions can be used to calculate MD and FA.**

While DTI is commonly used in clinical studies, it can only resolve the primary voxel orientation even when two or more differently oriented white matter bundles are in the same voxel. Known as a partial volume effect, when axons within a voxel do not all run parallel to each other, DTI can lead to incorrect estimations of fiber orientation. Crossing fibers have been shown to lead to ambiguous microstructural indices and also result in anatomically inaccurate tractography [56, 57]. There are many proposed methods which seek to estimate crossing fibers on a voxel-wise basis such as constrained spherical deconvolution (CSD) [58-60], diffusion orientation transform (DOT) [61], Q-ball imaging (QBI) [62], and Generalized q-space Imaging (GQI) [63]. These are termed as high angular resolution diffusion imaging (HARDI). Many of these approaches rely on spherical harmonics to model the directional diffusion information on a sphere according to the diffusion directions at which each volume is acquired. DTI and HARDI methods aim to provide an estimate of a spherical function called the diffusion orientation distribution function (dODF) or of the fiber orientation distribution (FOD). The dODF is the radial integration of the diffusion propagator:

$$ODF(\hat{x}) = \int_0^\infty P(r, \hat{x}) f(r) dr$$

where $\hat{x}$ is a unit vector in the direction of $x$, r is the radial distance from the origin, and the function $f$ weights the contribution to the integration along different radii. The dODF reflects the relative number of spins that have diffused in a given direction, $x$. The FOD is the fraction of fibers in each voxel that point in each direction and is also defined over a sphere. Daducci et al. used synthetically generated data to compare these reconstruction methods. Figure 4 shows the ellipsoids representing the orientations of five voxels in a human subject as estimated from three different reconstruction methods [64].

### 3.2. Tractography

Tractography delineates white matter pathways using the orientation information provided by DTI or HARDI methods. This “virtual dissection” technique produces streamlines which exhibit a strong resemblance with freeze-thaw brain white matter dissections [65], show a strong sensitivity for known white matter anatomy [66], and correlate well with disease phenotypes.

Deterministic tractography only considers the main direction of the estimated orientation distribution function which suits DTI well. The first attempt at tractography was introduced by Mori et al. [67] and was called Fiber Assignment by Continuous Tracking (FACT). FACT simply followed the primary eigenvector of the diffusion

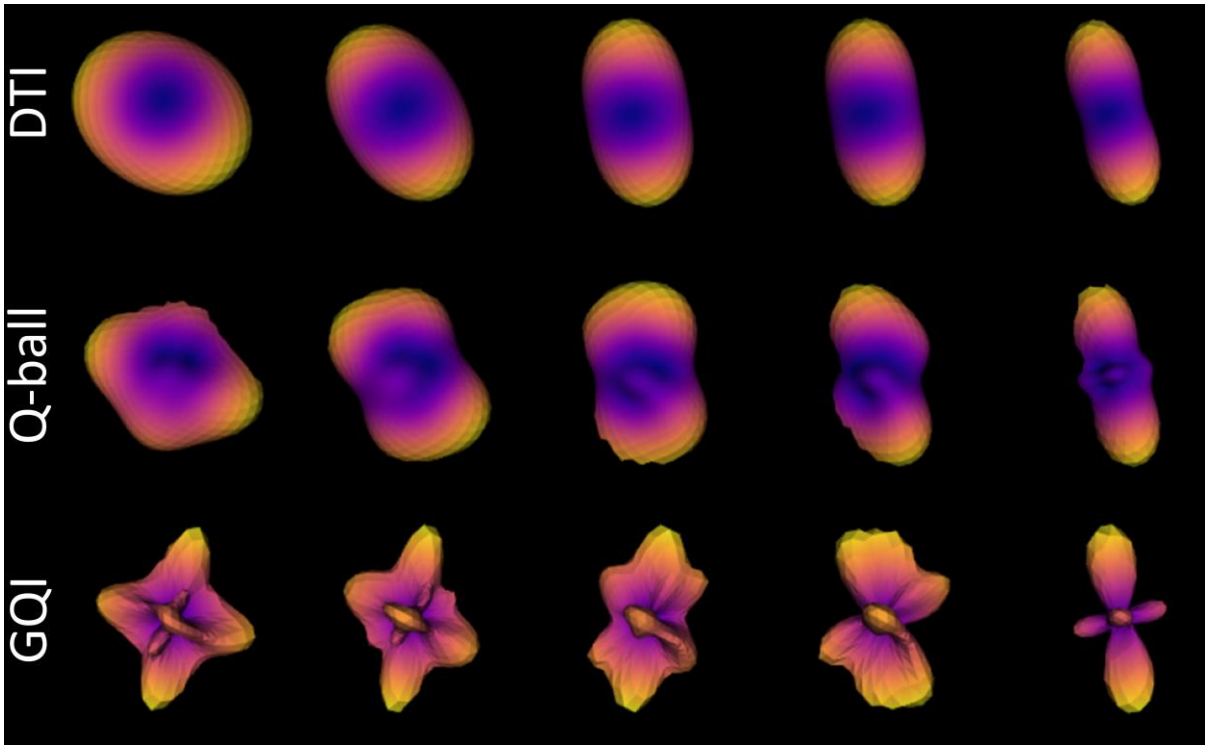


**Figure I-4. Estimated diffusion features is shown for five voxels from a human subject for reconstruction methods DTI, Q-ball, and GQI. The DW-MRI acquisition consisted of 384 diffusion gradient directions with a b-value of 1000 s/mm$^2$. Where DTI is limited to ODFs with a single direction, HARDI methods such as Q-ball or GQI are able to estimate ODFs capable of capturing crossing fibers within a voxel.**

tensor over the entire voxel. Despite its simplicity, this serves as the framework for state-of-the-art deterministic tractography methods which have introduce only minor variations in streamline propagation [18, 67, 68].

Those methods which resolve crossing fibers are more fitted for probabilistic tractography which additionally estimates a distribution representing how likely non-primary orientation is to lie along a fiber. Probabilistic tractography takes into account sources of uncertainty in orientation estimates. This can be accomplished by selecting a random sample from the orientation distribution for selecting the tracking direction for the next step in the tracking process. After many repetitions, the most visited pathways will be assigned a higher probability. Other methods characterize uncertainty through bootstrapping from multiple repetitions [69, 70] or use Bayesian approaches to infer uncertainties from the given parametric model [71, 72]. An example of ODFs and the tractography streamlines for the corpus callosum are shown in Figure 5.

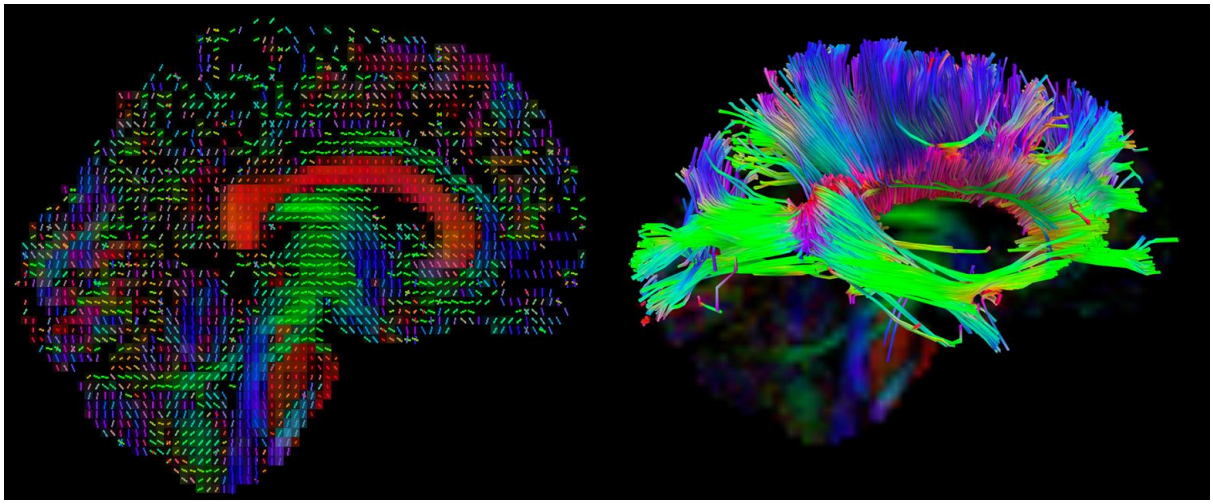

**Figure I-5. A sagittal slice visualizing the orientational distribution function (ODF) at each voxel (left) and the resulting tractography for the corpus callosum (right). The ODF within each voxel is estimated using Generalized q-space Imaging, and deterministic tractography is used to delineate the tracts. Here this was accomplished using a software which allows defines regions of interest such as seed regions and regions of exclusion which guide and restrict how and where the tract is estimated.**

## 4. Deep Learning in Medical Imaging

To understand some of the concepts and methods that are discussed in the following sections, deep learning and its development in medical imaging should be understood. This section will cover some basic machine learning concepts, how deep learning is applied to certain tasks, and how deep learning network architecture has been designed for medical imaging tasks relevant to this thesis.

### 4.1. Deep Learning

The most common form of machine learning is described as supervised learning in which some task is learned using a ground truth to guide the model during training. Machine learning uses statistical models to find patterns in large datasets, and these methods have been found to be powerful tools in regression and classification. In regression the ground truth can be a continuous value while in classification the ground truth is a discrete value or label. Traditional machine learning is limited in its ability to process raw data and has typically relied on domain expertise to design feature extractors which can transform data into a learnable representation or feature vector from which a classifier or regressor could detect patterns [73].

As a form of representation learning, deep learning methods are fed raw data and discover representations or

features that are needed for the defined task automatically. This class of methods is characterized as "deep" due to having multiple levels of representation obtained by composing non-linear modules which transform features at one level into a higher, more abstract representation [73]. This ability to automatically detect features has enabled deep learning methods to outperform "traditional" machine learning models in image recognition [74-77], speech recognition [78-80], and various problems in chemistry, physics, and biology [81-86]. A deep neural network is a model architecture which is defined by sequential layers of hidden neurons or nodes and weighted connections between these layers. A fully connected layer is a hidden layer in which every neuron will have a weighted connection to every neuron in the previous layer. However, data or features must be in vector form to be used as input for a fully connected layer. Convolutional neural networks (CNNs) were designed to extract features from image data. Rather than having a neuron be fully connected, each neuron is connected to a small region in the image or a local receptive field. A simple CNN is shown in Figure 6. These are typically defined by $h \times w$ weights and a bias term, but these networks have been extended to 3D images as well where the weights for the receptive field are $h \times w \times d$. 3D CNNs are prevalent in medical imaging as many in-vivo acquisitions result in 3D images. Regardless of the architecture, neural networks are fed some input and the resulting output is evaluated against the ground truth through a loss function. The weights are then updated with the resulting loss through backpropagation. This process computes the gradient of the loss function with respect to the weights in the model through the application of the chain rule for derivatives [73].

When training deep learning models, data typically is divided into a training set from which features are learned, a validation set which serves as a stopping criterion during training, and a withheld testing set on which the trained model can be evaluated. It is possible for a large enough model to learn too much from a set of training data. Overfitting occurs when the extracted features are specific to the training data and do not generalize to unseen data. Deep neural networks tend to have millions of parameters, so regularization techniques are an important asset during training to prevent overfitting. Regularization can occur in many ways such as adding a regularization term to the loss function, applying dropout to temporarily disable a node [87], and data augmentation which artificially increases the size of the dataset [88]. Overfitting becomes harder to prevent with smaller datasets which is common for tasks where labeled data is costly to obtain. Semi-supervised learning describes methods which attempt to address this issue by extracting features from labeled data and unlabeled data simultaneously [89]. Recent semi-supervised methods constrain the model through an additional term in the loss function that is computed over unlabeled data [90-94].

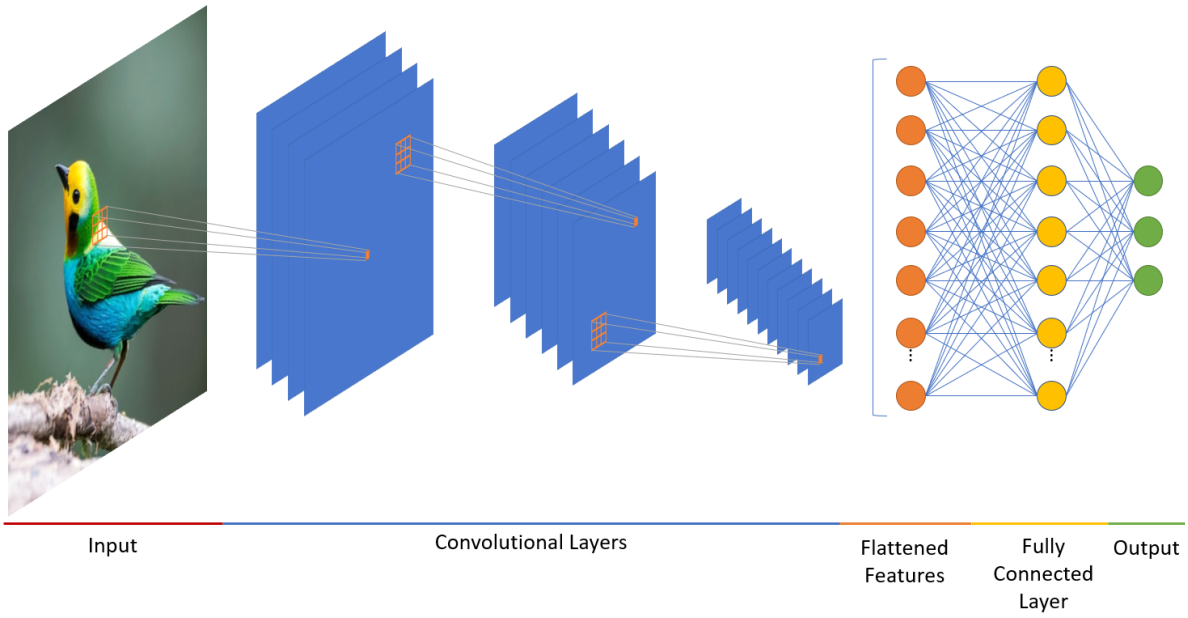


**Figure I-6. A simple 2D convolutional neural network is shown here where a 3×3 convolutional kernel is used at each convolutional layer to extract features from an input image. The resulting feature maps in earlier layers capture fine, local features while the feature maps from later layers will capture more global features. Though not all network architectures need a fully connected layer, many use them to produce the final output given all of the features extracted from convolutional layers.**

### 4.2. Segmentation

In segmentation tasks, the goal is to assign a label to every pixel or voxel in an image rather than a single label to the image itself. In medical image segmentation, the most widely successful deep CNN method has been the U-net [95]. The primary characteristics of this architecture are an encoding phase which extract high level features and downsamples the input image, a decoding phase which upsamples the low level features back to the original shape of the input image, and skip connections which allow features from the encoding layers to be considered within decoding layers and prevent gradients from becoming so small that they have little impact on the weights which define the encoding convolutional kernels [96].

### 4.3. Network Design for DW-MRI

Most CNN methods and implementation only consider 2D or 3D data when DW-MRI acquisitions are 4D. The simplest approach to address this has been to feed the spherical harmonic representation of the DW-MRI signal rather than the signal itself [6, 8] to the model. In this way a slice or volume of spherical harmonics can be passed

through a 2D or 3D convolutional layer while considering each spherical harmonic basis as an independent feature channel. However, some methods learn directly from the signal itself. In these cases spherical harmonics may be used to interpolate data such that all data fed to the model has the same number of gradient directions [9, 97].

## 5. Brain Parcellation

Image segmentation is a largely studied problem in computer vision and is fundamental in biomedical image analysis. Though manual segmentations are highly used and often considered the gold standard, they often are not highly reproducible and are very time consuming. Due to this, researchers tend to rely on predefined atlases and automated parcellation methods for statistical analysis.

### 5.1. Atlas-based Segmentation

In the medical field, an atlas is generally a set of two images: an intensity image or a template image and a segmentation image or label image. Single-subject atlases or deterministic atlases are constructed based on a single subject. An example of this is the 1988 Talairach Atlas [98]. Population-based atlases or probabilistic atlases are constructed using multiple subjects to better represent anatomical diversity. The MNI structural atlas is an example of a population based atlas and uses popular MNI152 template [99, 100]. Many atlases have been created varying in number of labels and populations used [101]. Most atlases emphasize cortical or sub-cortical gray matter regions [102-109], and some of these atlases include a homogenous label or left and right hemisphere labels for white matter [110, 111]. There also exists region based white matter atlases [98, 112, 113]. However, we feel the literature is missing a population based white matter bundle atlas which considers crossing fibers and the probabilistic nature of white matter bundles. To use an atlas in a study, labels need to be propagated to image space in which they are intended to be used. This relies on image registration which has been extensively studied [114]. Multi-atlas segmentation and label fusion extends this concept to using multiple atlases to obtain more robust segmentation [115-118].

### 5.2. Fully-automated Segmentation

Fuzzy c-mean methods became a popular fully-automated method for segmenting a brain in to gray matter, white matter, and cerebrospinal fluid regions in the 1990s [119]. Advanced whole brain segmentation methods such as region growing, clustering, and deformation models have been proposed since then and most recently deep learning methods such as the U-Net have become popular [95]. Because it is impractical to fit clinical, high-resolution MRI to state-of-the-art 3D CNN methods due to memory limitations of prevalent GPUs, many patch-based methods have been proposed to tackle whole brain segmentation [120-122]. Spatially localized atlas network tiles (SLANT) improves the patch based approach through the use of multiple independents 3D CNNs each of which are only responsible for a particular spatial location in the brain [1]. SLANT uses the BrainCOLOR labeling protocol [123] as its ground truth for whole brain segmentation. An example of the resulting BrainCOLOR segmentation from SLANT is shown in Figure 7.

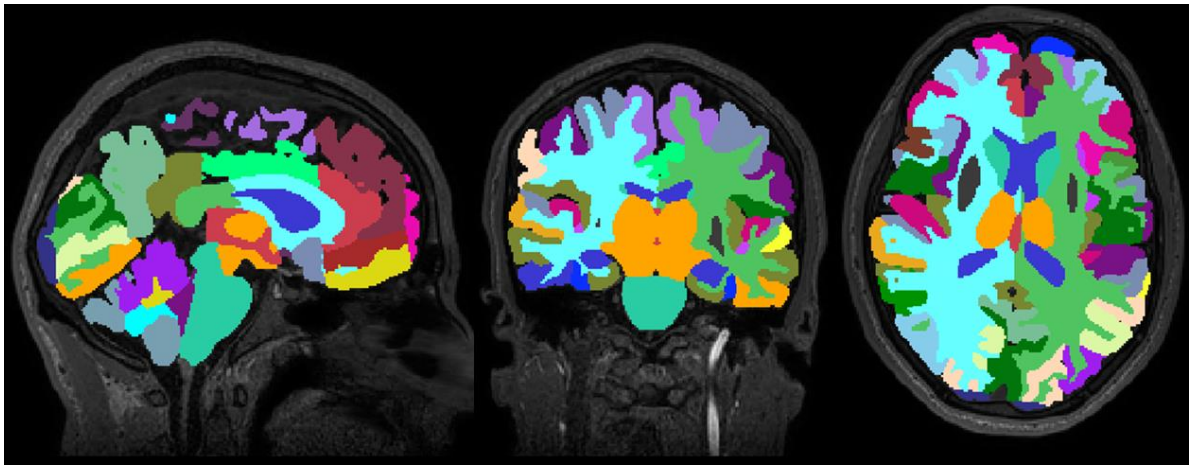

**Figure I-7. A sagittal, coronal, and axial slice (left to right) of a T1 weighted brain volume and the overlaid BrainCOLOR segmentation are shown for a single subject. The parcellation was generated using SLANT [1] and segments the brain in to different gray matter, white matter, and cerebral spinal fluid regions. Segmentations such as these allow for region based statistics and analyses.**

## 6. Harmonization & Preprocessing of DW-MRI

Variability in DW-MRI measurements can result from a difference in the number of head coils used, the sensitivity of the coils, the imaging gradient non-linearity, the magnetic field homogeneity, the differences in the algorithms used to reconstruct the data, as well as changes made during software upgrades [24-28]. Aggregating DW-MRI data from different sites or scanners becomes difficult as these can cause non-linear effects in the acquired images as well as the resulting FA or MD [124, 125]. These effects can be reduced by using similar scanners with similar

pulse sequence parameters and field strengths [126-128], but studies have shown that large differences still exist in diffusion measures from different sites [4, 129-131]. The variability introduced by differences in acquisition parameters or hardware affect the reproducibility of DW-MRI based microstructure models [132, 133]. Reproducibility is a closeness measure between a pair or group of measurements. Harmonization approaches and methods aim to increase reproducibility and reduce error caused by variance and bias introduced by hardware and site effects. This section discusses the preprocessing methods used to correct well studied imaging artifacts and harmonization methods used to correct for scanner and site effects.

### 6.1. Preprocessing

After acquiring a DW-MRI at a site, the DICOM data is typically converted to the NIfTI standard neuroimaging format before preprocessing. Diffusion weighted spin-echo EPI images are very sensitive to non-zero off-resonance fields caused by the susceptibility distribution of the subjects head and by eddy currents from rapid switching of the diffusion weighting gradients [134]. An effective method of correcting the susceptibility is to use two or more acquisitions such that the mapping field are different. This is commonly done by acquiring two images with opposite phase encoding directions. Given these images and the acquisition parameters the susceptibility field can be estimated by finding the field that when applied to the two volumes will maximize their similarity of the unwarped volumes. The similarity can be gauged by the sum-of-squared differences between the unwarped images, and this metric allows use of Gauss-Newton for jointly finding field and any movement that may have occurred between the two acquisitions [134].

The effects of eddy current distortions can be corrected through modeling the diffusion signal using a Gaussian Process. By assuming the signal from two acquisitions acquired with diffusion weighting along two vectors with a small angle between them is more similar than for two acquisitions with a large angle between them and assuming the signal from two acquisitions along vectors $v$ and $-v$ is identical, it can be assumed that if $v_1$ and $v_2$ are two vectors with a "small" angle between them so that it can be assumed that the signal from the corresponding acquisitions is "similar" then $v_1$ and $-v_2$ are equally similar [135].

It is also typical to perform skull stripping leaving only brain tissue in the data. MR data can contain a considerable amount of non-brain tissue such as eyeballs, skin, fat, and muscle, so a robust method is necessary to automatically extract the brain tissue. One such method uses a deformable model that is fit to the brain's surface by the application of a set of locally adaptive model forces [136]. Before any models are fit to the DW-MRI, the diffusion

data is normalized to a non-diffusion weighted b0 volume. Dividing the diffusion signal by the b0 signal removes intensity variations due to T2 weighting and radiofrequency inhomogeneity [137].

It is usually the case a structural T1 is acquired in the scanning session along with the DW-MRI data. Affine registration can be used to align a subject's diffusion data to the T1 [138]. This is useful for a few reasons. The first is that this allows anatomical structure-based segmentations that are estimated from a subject's T1 to be applied to the diffusion data. The second is that non-rigid registrations to a standard template such as the MNI template is most successfully done using T1 images, and this is necessary for most inter-subject analyses [139, 140]. Even after these preprocessing steps, differences in diffusion derived metrics exist in data acquired of the same subject across sites using the same acquisition parameters on hardware from the same manufacturer. Figure 8 shows the differences in FA for a single subject.

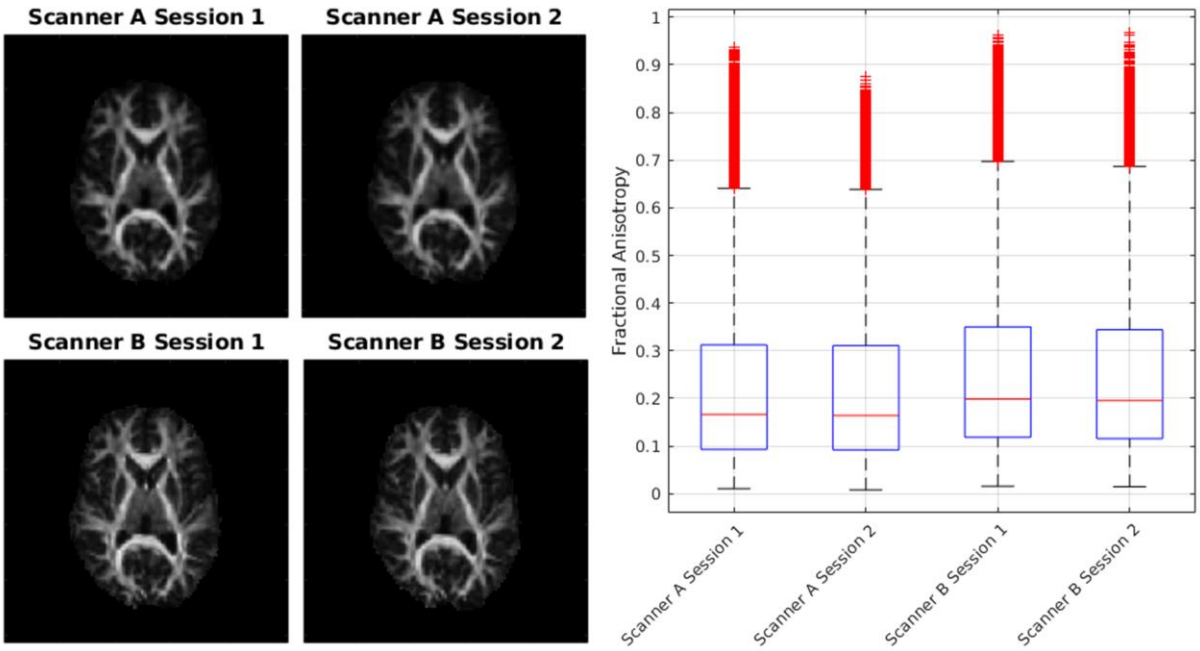


**Figure I-8. The FA for a single subject scanned at two sites with repeat acquisitions at each site can show differences even after standard preprocessing including susceptibility and eddy distortion correction. A medial axial slice of FA for each acquisition is shown (left) as well as the FA across the entire brain volume (right). All acquisitions were acquired with the same parameters and with 96 gradient directions of each of the following b-values: 1000 s/mm$^2$, 1500 s/mm$^2$, 2000 s/mm$^2$, 2500 s/mm$^2$. Affine registration was performed between the b0 of each acquisition and a single T1 of the subject and then applied to the resulting FA images.**

### 6.2. Harmonization Methods

Harmonization across scanners, studies, and patient populations can be approached in multiple ways. Empirically measuring and correcting for system performance through identifying systemic issues is one approach. One example of this is correcting for gradient non-linearity in the magnetic fields within a MR system [30, 32, 33]. Bias introduced by nonlinearity of the scanner's gradient fields causes spatially varying error sin the applied diffusion gradients and have been empirically confirmed as a major source of error for multisite studies in derived diffusion metrics. If the gradient coil fields are known, and their spatial derivatives can be used to compute the spatially varying properties of the diffusion gradients that are actually obtained from a given pulse sequence. Some preliminary work has been done to show that these fields can be obtained through empirical measurements of a large oil phantom if the manufacturer fields are not readily available [29, 141]. Another example of harmonization through empirical means is correcting for signal drift in DW-MRI [142]. This effect has been described as a temporal instability can be caused by a high load on the system through a combination of diffusion gradients and echo-planar imaging readout. It has been shown that this can be modeled as a linear or quadratic decrease in the diffusion signal across time using interspersed b0 images throughout a diffusion acquisition.

Statistical models have proven to be effective for multi-site harmonization. A DTI harmonization technique proposed by Mirzaalian et al. [4] utilizes rotation invariant spherical harmonic (RISH) features and combines the unprocessed DTI images across scanners. A major drawback of this method is that it requires DTI data to have similar acquisition parameters which is often unfeasible in multi-site studies. By analyzing the effectiveness of several statistical approaches that were developed for other data types, Fortin et al. [35] found that ComBat [2] achieved the best performance. Originally developed for genomics data, ComBat uses an empirical Bayes framework for adjusting data for batch effects that is robust to outliers in small sample sizes. The term batch effects here refers to non-biological differences that make samples in different batches not directly comparable.

Many deep learning approaches have been employed for diffusion harmonization as well. Nath et al. utilized a dual network to incorporate unlabeled paired in-vivo DW-MRI of human subjects along with labeled squirrel monkey DW-MRI with histology ground truth [8]. Koppers et al. designed a residual network specifically for spherical harmonic representations of DW-MRI which predicts the spherical representation at one scanner given the spherical harmonics of another scanner [6]. Given DW-MRI from multiple sources, Moyer et al. uses a method based on variational auto-encoders to learn an intermediate representation that is invariant to site and protocol specific effects

[7]. None of these techniques, however, incorporate information from other imaging modalities in their models. Specifically, T1 structural information is of particular interest to this thesis and how resulting segmentations can be used to improve harmonization models. Also of interest are the Nullspace Tuning methods developed by Nath et al. [8], and how they can be applied to other domains.

## 7. Contributed Work

We pursue two novel aspects of diffusion harmonization. First, we have developed empirically driven preprocessing techniques (**Contribution 1**). Second, we pursue the use of nullspace tuning as a general machine learning method. Though it was first used in diffusion harmonization, we show that it can be used generally as a semi-supervised classification framework where partially labeled data is available as well as in other medical imaging domains (**Contribution 2**). Third, we explored the use of anatomical information in DW-MRI. This included building white matter bundle atlases which are missing from literature and building models which can predict fiber tract probabilities from T1 structural brain data. Additionally, this included using nullspace tuning along with anatomical segmentations to improve deep learning approaches to diffusion harmonization (**Contribution 3**).

### 7.1. Contribution 1: Empirical characterization of system-based variation

Cerebral spinal fluid (CSF) exists throughout the brain and is known to be isotropic. However, the observed diffusion signal shows a large degree of variability in the CSF regions of the brain. We investigate the variation in a large longitudinal dataset and models that potentially explain this variability. We also evaluate if these models correlate with variation in the white matter. This contribution is covered in more detail in chapter II.

MRI systems in some cases have a stabilization period upon beginning a session during which the resulting signal between diffusion weighted volumes is either increasing or decreasing. This may be due to external factors of the parameters of the acquisition itself. Previously this was characterized as a global signal drift across the entire volumes. However, we investigate the use of spatially varying signal drift correction models and compare them to the global model using interspersed non-diffusion weighted volumes. Additionally, we compare spatially varying models which are parameterized at a voxel-wise level to those parameterized at global level which have a smaller parameter space. This contribution is covered in greater detail in chapter III.

It is understood that gradient fields impart scanner-dependent spatial variation in the applied DW-MRI. These can be corrected if the gradient nonlinearities are known but retrieving the manufacturer specifications is not well

supported. We propose an empirical approach to mapping the gradient nonlinearities with sequences that are supported across the major scanner vendors. These estimated fields are evaluated against the manufacturer specifications for our scanners, and we investigate the benefit in a diffusion phantom as well as in-vivo in a human subject. This contribution is covered in detail in chapter IV.

### 7.2. Contribution 2: Investigating the effectiveness of Nullspace Tuning across domains

Previous work has introduced the idea of Nullspace Tuning which performs semi-supervised learning through unlabeled paired data. The key idea is given a pair of unlabeled samples which belong to the same class but the class itself is unknown, a model can learn to minimize the distance between them, essentially tuning the nullspace. Nullspace Tuning was successfully used in estimating diffusion orientation distribution functions using DW-MRI and histology data. We formalize the theory and show the value it has in general classification tasks over and alongside popular semi-supervised learning methods. This contribution is covered in detail in chapter V. We go on to show the use of nullspace tuning in two other medical imaging domains: skin lesion diagnosis and lung cancer detection. We show in simulated and real paired data, nullspace tuning can improve a model's ability in the medical image processing tasks as labeled data become limited. This is covered in chapter VI.

### 7.3. Contribution 3: Exploring Anatomical Information in DW-MRI

Current literature lacks a white matter atlas which reflects the crossing tracts in the brain. Using current state-of-the-art automated tractography protocols, we construct a 4D white matter atlas for each definition where each volume corresponds to a specific bundle. The details behind the construction and the analysis of these atlases are covered in chapter VII. Additionally, using the U-Net architecture and the SLANT training approach, we develop a model which can estimate probabilistic bundle segmentations given a T1 structural image. We compare this model to registration to our atlases. This is covered in chapter VIII.

Lastly using the anatomical priors and the nullspace tuning concept, we develop a DW-MRI harmonization approach which can harmonize multiple datasets where paired data is available. By relying on the SHORE representation of the diffusion signal, we account for multiple acquisition parameters and relate a single shell input space to a multi-shell target space. This is covered in chapter IX.

The work of this dissertation is outlined with the following goals:

1.) Characterization and correlation of signal drift in DW-MRI

2.) Consideration of cerebrospinal fluid intensity variation in DW-MRI

3.) Gradient non-linearity correction of multi-site DW-MRI with empirical field maps

4.) The value of Nullspace Tuning using partial label information

5.) The use of Nullspace Tuning in other medical domains

6.) Predicting tract densities from structural T1

7.) DW-MRI harmonization using SHORE, anatomical segmentations, and nullspace tuning

## Chapter II. Consideration of Cerebrospinal Fluid Intensity Variation in Diffusion Weighted MRI

### 1. Introduction

Vos et al. recently reviewed the effectiveness of using minimally weighted images ("b0's") interspersed throughout a scan to correct temporal instability in scanner systems [142-144] and this was extended to spatially temporal models [37]. In addition, correcting for non-linearities in the gradient fields of the magnetic coils can be accomplished through empirical field mapping techniques [30, 141]. However, these corrections are not viable for datasets that were acquired before these techniques became widely available. Cerebral spinal fluid (CSF) can be used to observe trends in signal intensity as it is known to be isotropic. We present a case study on a large longitudinal dataset and examine variation in the CSF regions of the human brain (Figure 1, Figure 2, and Table 1).

**Table II-1. The median standard deviation for all volumes within each session, across all sessions for all subjects, and across all subjects. The percentage of that value with regards to the median signal for all data is also shown. The 3rd column shows the size of the data, and the last column shows the p-value of the data against the intra-session data.**

| Data | Median SD | Percent of Median Signal | N | p-value against Intra-session |
|---|---|---|---|---|
| Intra-session CSF | 0.0029 | 3.59 | 1954 | N/A |
| Inter-session CSF | 0.0086 | 10.60 | 542 | < 0.001 |
| Inter-subject CSF | 0.0216 | 26.59 | 3949 | < 0.001 |

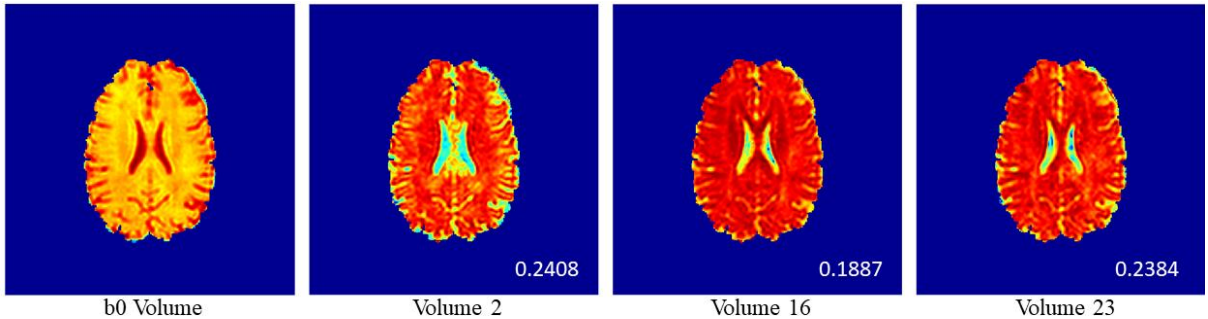


**Figure II-1. A slice from the b0 and three diffusion weighted direction of a single scan are shown with logarithmic intensity (a.u.). In the lower right-hand corner of the median intensity of the diffusion weighted volumes within the left lateral ventricle is shown. Note the variation of up to 28% in absolute intensity.**

## 2. Data

Herein, we consider a large longitudinal dataset comprised of 3949 MRI brain acquisitions of 918 subjects. Subjects have repeated DW-MRI acquisitions in each session, and 542 subjects have repeat sessions at later dates. All data were acquired after informed consent under institutional review board and accessed in de-identified form. Each session included a T1-weighted structural MP-RAGE (number of slices=170, voxel size=1mm×1mm×1.2mm, reconstruction matrix=256×256, flip angle=8 degrees and TR/TE=6.5ms/3.1ms) and two diffusion acquisitions. Each diffusion acquisition consists of an initial b0 image and thirty-two diffusion weighted volumes all with a b-value of 700 s/mm2 (number of gradients=32, number of b0 images=1, TR/TE=7454/75 ms, number of slices=70, voxel size=0.81×0.81×2.2 mm3, reconstruction matrix=320×320, acquisition matrix=116×115, field of view=260×260 mm, flip angle=90°). Susceptibility correction [134] and eddy current correction [135] techniques are applied to the diffusion data as a preprocessing step as well as b0 signal normalization. The MP-RAGE was segmented with the BrainCOLOR protocol (Neuromorphometrics, Inc., Somerville, MA) using hierarchical non-local spatial STAPLE [145]. For each 3D volume in a scan, the median signal is computed within the co-registered (FSL flirt[146]) regions of interest (ROI) from the BrainCOLOR segmentation defined over three CSF filled regions in the brain: the right lateral ventricle, left lateral ventricle, and third ventricle.

### 2.1. Variation

Figure 1 shows qualitatively the variation in the left and right lateral ventricles in a single scan. In Figure 2, the median normalized signal (i.e., diffusion weighted intensity divided by the minimally weighed reference) for all scans in the three ROIs is shown over the course of the thirty-two acquired diffusion volumes after the first median value of each scan has been subtracted to ensure all timeseries have the same starting position of zero. From this the variation across all scans can be seen at certain volumes especially in the left and right lateral ventricles. In Table 1, the average standard deviation for all ROIs is shown for intra-session data, inter-session data, and for inter-subject data. The standard deviation nearly triples from intra-session to inter-session and again at least doubles from inter-session to inter-subject. The relatively low standard deviation within in a session and higher standard deviation across all sessions for a subject indicates that the variation is not only an effect of anatomical differences. The steady increase in standard deviation from intra-session to inter-session to inter-subject indicates that the effect is static.

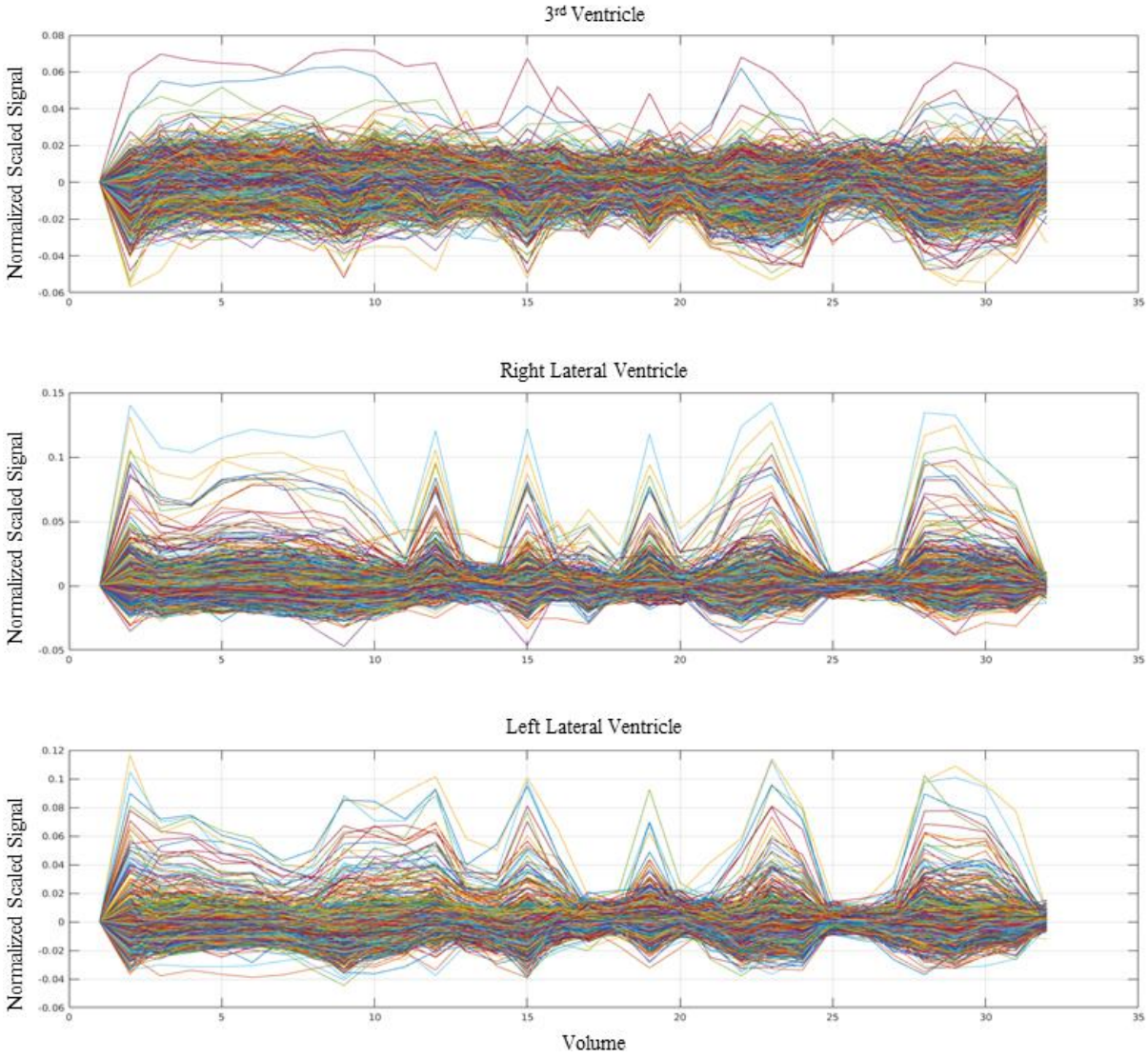


**Figure II-2. All median signals for three CSF regions in the brain for 3949 scans, with each line corresponding to a single scan. From top to bottom the rows correspond to the 3rd ventricle, the right lateral ventricle, and the left lateral ventricle. The median signal has been normalized and scaled so that all start at the same point. Note the wide range of signal variation and the visually clear dependence on gradient direction.**

### 2.2. Modes of Variation

Figure 3 shows the contribution of variation from the principle components of the median signals as well as the cumulative variation from the most contribution component to the least for each ROI. It can be seen that almost ninety percent of the variance can be attributed to the first three components indicating that an appropriate correction model would be able to reduce this variance. In Figure 4, the first three components are normalized and plotted. Figure 5 shows the same data from Figure 2 represented as a scatter plot, but now the color of the point represents the value

of the corresponding gradient direction. In the right and left lateral ventricles, it can be seen that the volumes with the most variation are either acquired with the gradient taken along the y or z direction and the volumes with the least variation are acquired with the gradient taken along the x direction. In addition, it seems that the sharp decreases in the plot of component one in Figure 4 correlates with the volumes at which the gradient is taken in the y direction as seen in Figure 5.

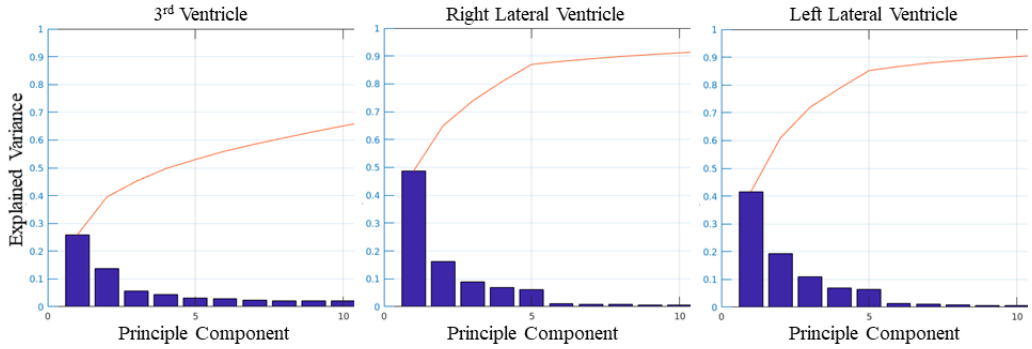


**Figure II-3. Here we see the explained variance for the principle components for the median signal for all scans. From left to right the plot corresponds to the 3rd ventricle, right lateral ventricle, and left lateral ventricle. This shows that most of the variance is explained by the first 3 modes of variation.**

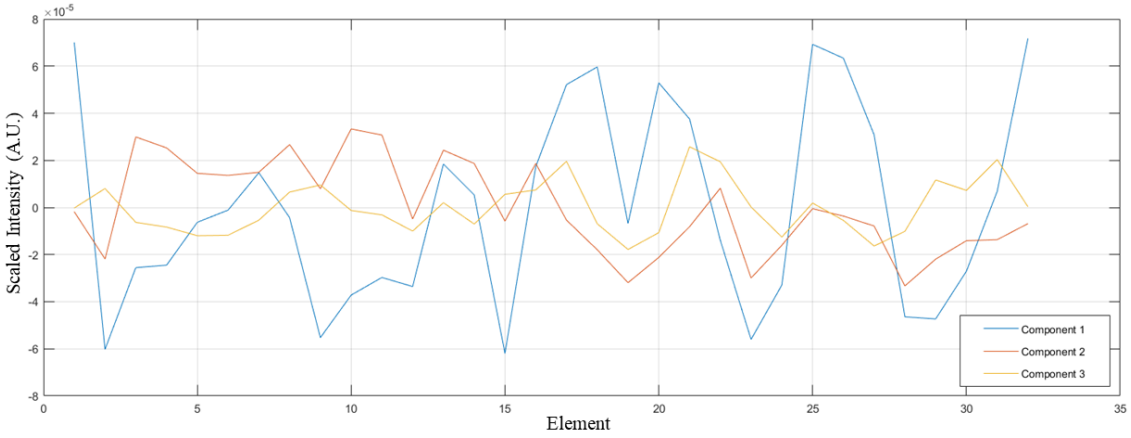


**Figure II-4. This plot shows the first three principle components. Note the lack of low frequency temporal drift.**

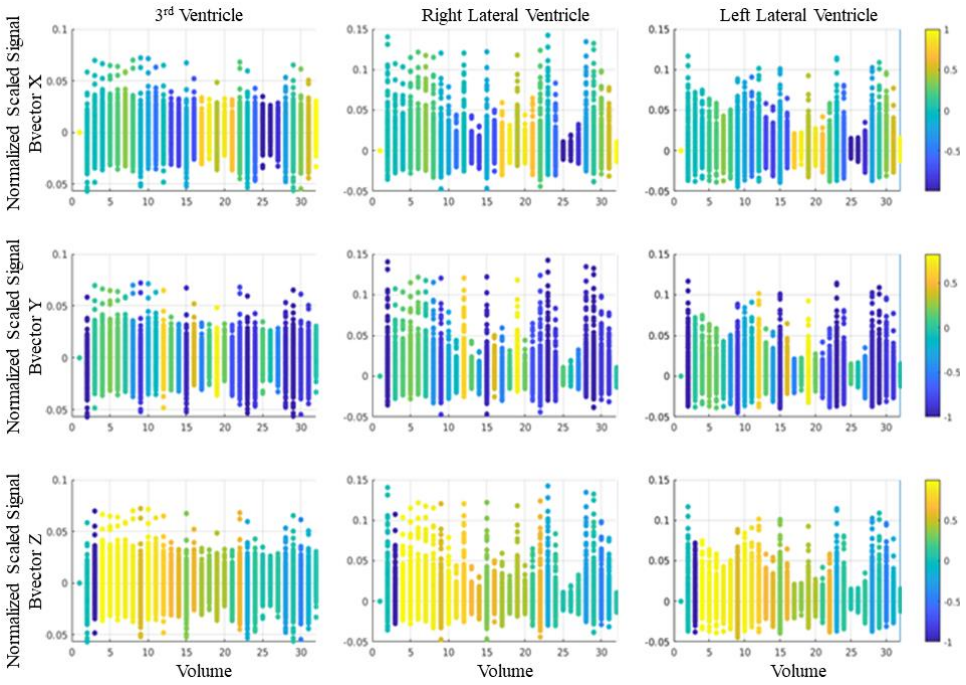


**Figure II-5. Each row corresponds to a b-vector (x, y, and z from top to bottom) and each column corresponds to a CSF region (3rd, right lateral, left lateral, from left to right). Each represents the same data from figure 1, but the color represents the value of the b-vector at that volume. This shows that as the variation in the data increases as the gradient is taken in the y and z directions.**

## 3. Experiments

To capture the variability in the signal, we examine five models: two linear models and three non-linear models, which will be referred to as models one through five. This section outlines the basis function used to approximate the signal through regression.

### 3.1. Linear Model

The linear model approximates the apparent diffusion coefficient (ADC) as a constant and attributes variation to temporal drift and baseline sensitivity to applied gradient direction. For example, these effects could be associated with directional flow related effects. The first linear model is defined by:

$$S(n,X) = d_1 s0 + d_2 n + d_3 n^2 + d_4 x + d_5 y + d_6 z + d_7 xy + d_8 xz + d_9 yz + d_{10} x^2 + d_{11} y^2 + d_{12} z^2 \quad (1)$$

where $S$ is the normalized signal, $s0$ is the normalized signal at the first volume, $n$ is the volume index, and $x$, $y$, and $z$ correspond to the vector $X$ that defines the gradient direction (b-vector). The coefficients are defined by $d_i$.

### 3.2. Simple Exponential Model

The third model is a simple concatenation of both prior models, while dropping temporal baseline drift (as it was not found to be significant, see below). The first non-linear model is defined by:

$$S_L = d_1 s0 + d_2 x + d_3 y + d_4 z \tag{3}$$

$$S_e = e^{-d_5 s0 - d_6 x - d_7 y - d_8 z} \tag{4}$$

$$S = S_L S_e \tag{5}$$

where $S_L$ is the linear portion of the model and $S_e$ is the exponential portion of the model.

### 3.3. Cross-term Exponential Model

The second non-linear model expands the third model to evaluate potential interactions between the gradients:

$$S_L = d_1 s0 + d_2 x + d_3 y + d_4 z + d_5 xy + d_6 xz + d_7 yz \tag{6}$$

$$S_e = e^{-d_8 s0 - d_9 x - d_{10} y - d_{11} z - d_{12} xy - d_{13} xz - d_{14} yz} \tag{7}$$

$$S = S_L S_e \tag{8}$$

### 3.4. Squared Exponential Model

The fourth non-linear model expands on the third model with non-linear terms for the x and y gradient direction but neglects cross terms due to limited statistical power:

$$S_L = d_1 s0 + d_2 x + d_3 y + d_4 z + d_5 x^2 + d_6 y^2 \tag{9}$$

$$S_e = e^{-d_7 s0 - d_8 x - d_9 y - d_{10} z - d_{11} x^2 - d_{12} y^2} \tag{10}$$

$$S = S_L S_e \tag{11}$$

## 4. Results

Each model was fit to the median signal in the left and right lateral ventricles from each scan. The significance values associated with each term of the models are visualized for each scan in Figure 6. In the linear and log model,

the $\boldsymbol{s0}$, $\boldsymbol{yz}$, and quadratic terms were the most significant. In the simple and cross-term model, the $\boldsymbol{s0}$, $\boldsymbol{y}$, $\boldsymbol{z}$, $\boldsymbol{e^{s0}}$, $\boldsymbol{e^{y}}$, and $\boldsymbol{e^{z}}$ terms were the most significant. In the squared exponential model, only the $\boldsymbol{s0}$, $\boldsymbol{z}$, $\boldsymbol{e^{s0}}$, and $\boldsymbol{e^{z}}$ terms were significant in most of the fits. In the linear model, we can see that the terms representing the index of the volume are insignificant which indicates that temporal affects are not causing the high variation. An estimation for the median signal of each scan was generated using the coefficients learned from each regression and the total root mean squared error (rmse) is shown in Table 2 for each method. In terms of accurately estimating the signal, the squared exponential model performed the best by a small margin over the other models. Table 2 also shows the mean $R^2$ and mean adjusted $R^2$ for each model. The cross-term exponential model had the highest $R^2$ while the squared exponential model had the highest adjusted $R^2$.

**Table II-2. The sum of the root mean squared error between the estimated median signal and the true median signal, the mean R2, and mean adjusted R2 for all models for all scans.**

| Model | Total RMSE | Mean $R^2$ | Mean Adjusted $R^2$ |
|---|---|---|---|
| Linear | 10.591 | 0.797 | 0.755 |
| Log | 10.666 | 0.794 | 0.759 |
| Simple Exponential | 11.991 | 0.753 | 0.722 |
| Cross-term Exponential | 10.199 | 0.818 | 0.770 |
| Squared Exponential | 9.859 | 0.817 | 0.779 |

**5. Conclusion**

The isotropic nature of CSF has allowed us to look at the variation of the signal which may be indicative of the variation within surrounding areas or even the whole brain. With few modes of variation, a viable model should be able to estimate the median signal with few variables being utilized as basis functions. Our results show that using the values of the b-vector to fit a model to each scan's signal over time allows for fairly accurate estimations. However, it is not clear if there is a strong correlation between the variation in the white matter of the brain and the variation in the CSF. If there is a strong positive correlation between the median signal in the CSF and the surrounding white matter regions in the brain, the estimated signal could be used to correct the variation in the same manner as the b0 drift correction as proposed by Vos et al [142]. Unfortunately, the CSF does not appear to provide a clear reference tissue. Yet, the variations are highly structured, dependent on diffusion weighting direction, and may provide useful anatomical metrics with additional biophysical modeling.

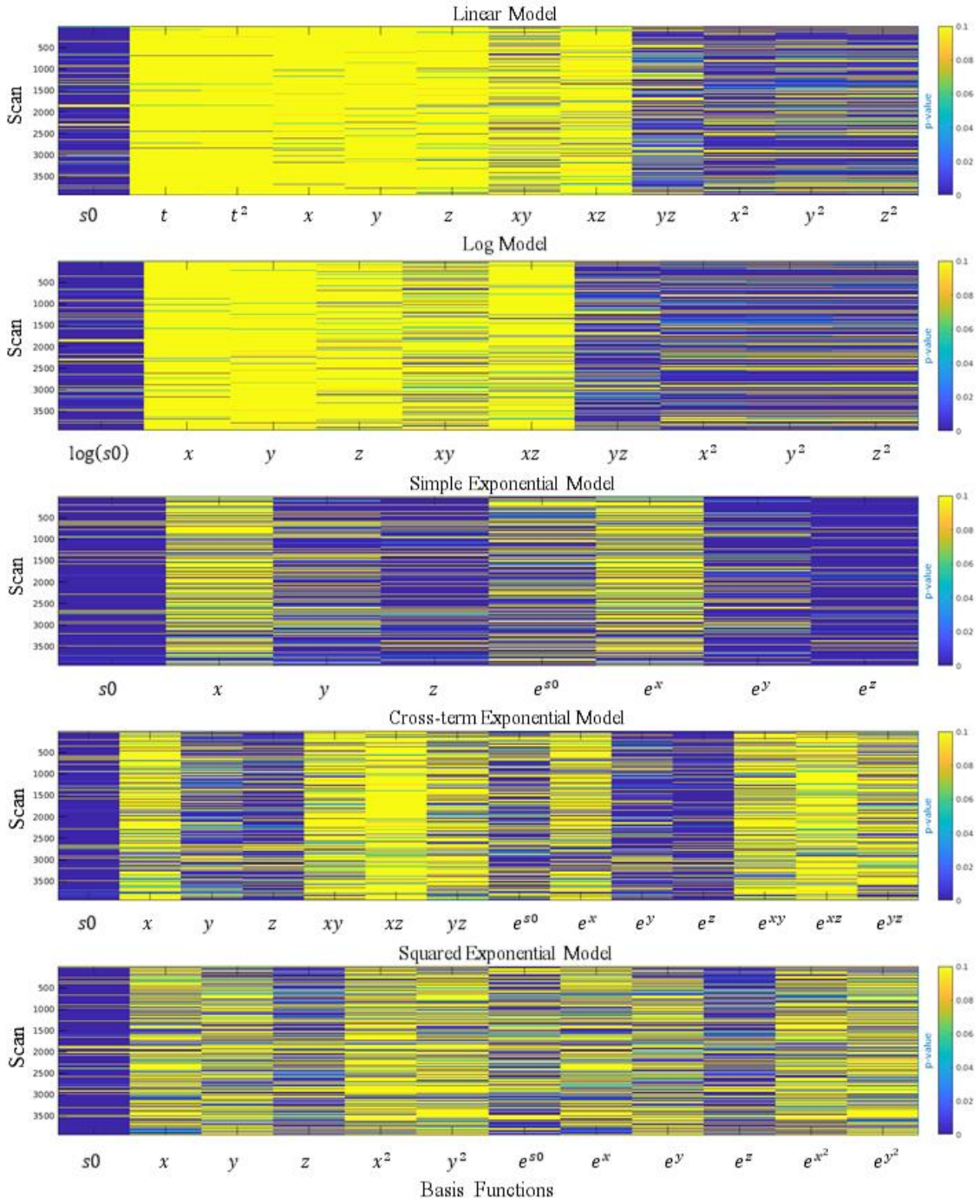


**Figure II-6. Each row corresponds to a model used to capture the variance in the left and right lateral ventricles. From top to bottom the models are the linear model, log model, simple exponential model, cross-term exponential model, and squared exponential model. Each row in the images represents p-values of the coefficients from the fitting the model to the median signal. Each column represents one of the terms that were used as the basis functions for the models.**

## Chapter III. Characterization and Correlation of Signal Drift in Diffusion Weighted MRI

### 1. Introduction

Diffusion weighted magnetic resonance imaging (DWMRI) enables non-invasive mapping of in vivo neural fiber architecture [66, 147]. It also provides important, albeit non-specific, markers for low anisotropy [148, 149] which may be indicative of the microstructure given a priori knowledge of the architecture of the pathway of interest [19, 150]. Quantitative accuracy of DWMRI derived metrics across scan sessions, scanners, and scanner manufacturers is critically important for broader application of DWMRI-derived metrics in the clinical setting, and substantial efforts have sought to map intra-[151, 152] and inter-[153]scanner variability while harmonizing DWMRI acquisitions and subsequent analyses [4, 35]. Recently, MRI scanner temporal instability has been shown to introduce systematic nonlinearities that can substantively impact observed apparent diffusion coefficients (ADCs) in a directionally dependent manner [143, 144], but fortunately these effects can be quantified and compensated through relatively standard modifications to traditional DWMRI protocols.

Vos et al. recently explored the effects of temporal instability in the scanner system on DWMRI data [142] and found that the effects were characterized as a decrease in global signal intensity. They proposed a temporal non-linearity model on a region of interest (ROI) basis and fit this model by relying on interspersing b0 images in the scan acquisitions. With those data, they were able to interpolate the mean signal of the defined ROI (the entire brain or phantom) to the temporal locations of the diffusion weighted scans and apply a correction to the unobserved reference signal that would have been temporally collocated. Moreover, Vos et al. show that the effect of temporal instability is present on scanners from multiple vendors [142].

To empirically illustrate this effect, Figures 1 and 2 show the resulting normalized signal and variance in the normalized signal in three selected ROIs in an ice-water phantom with 13 vials of varying PVP concentrations and a spherical isotropic sphere consisting of a single concentration of polyvinylpyrrolidone (PVP) respectively. The plots show the normalized signal across each volume corresponding to the 96 gradient directions acquired in 10 sequential scans for each phantom. The acquisition parameters are described in the next section. We observe that the variance in ADC is spatially dependent, which indicates that the signal drift occurs at different rates and even with different signs between ROIs. Moreover, the magnitude of the signal drift seems to decrease as time progressed in the session. This pattern in the signal drift requires a more complex correction that includes the spatial characteristics of the drift.

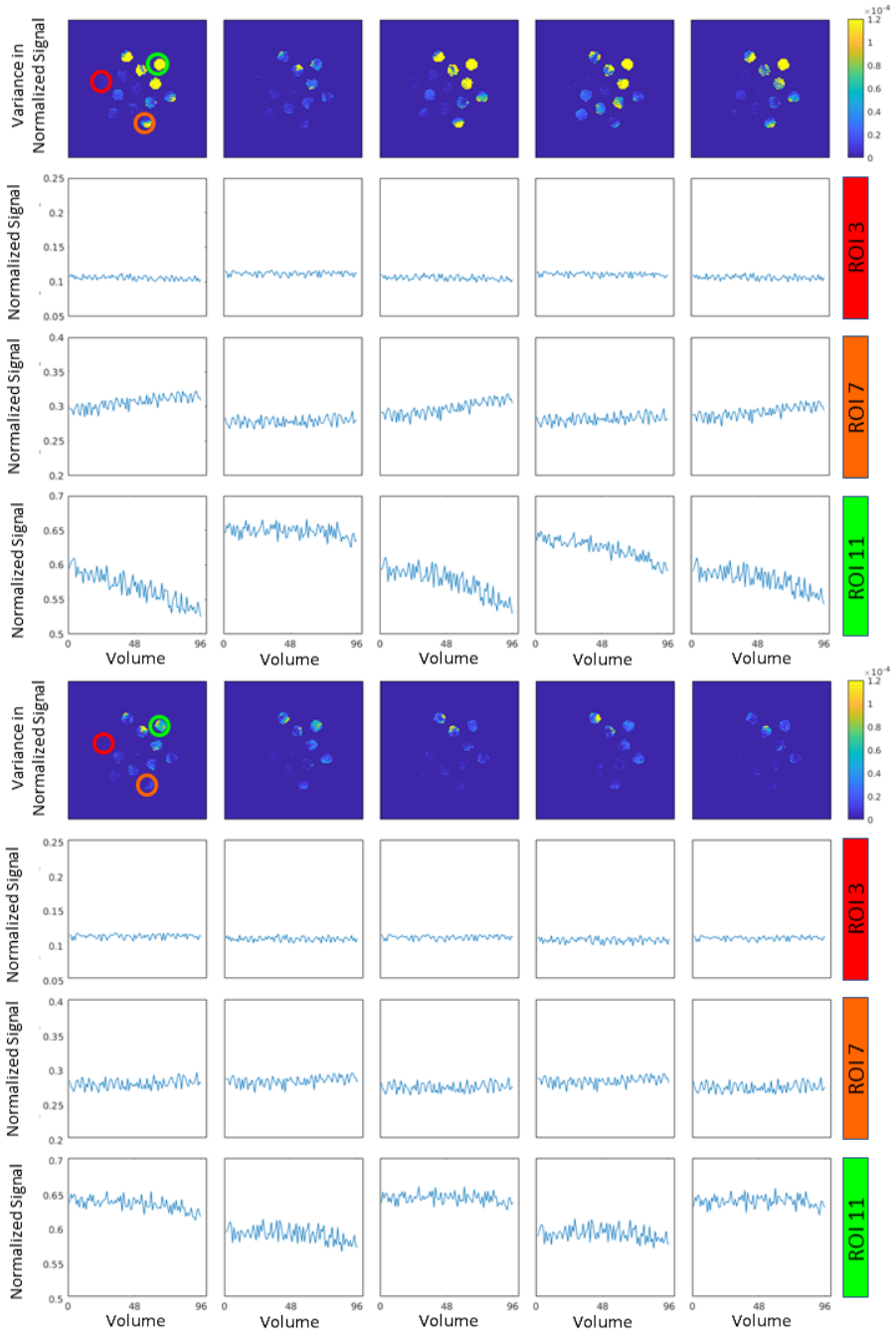


**Figure III-1. Empirical characterization of drift in the ice water phantom. This plot presents the variance in the ice water phantom in 10 different scans over the ¬course of the session (1st and 4th row) and normalized signal intensity within three ROIs (indicated in the top left plot). The top four rows represent the first five scans in the series with the bottom four representing the last five scans. Note that areas of high variance in the top plot row correspond to ROI's that have greater drift and require more substantial calibration. Of concern, observe that different ROIs within same scan are drifting with opposite signs, hence spatial correction of the signal drift is required. Furthermore, over the course of the session, the pattern of drift changes with time and with region. This complex drift pattern is motivation for the proposed spatial-temporal drift correction model, TS.**

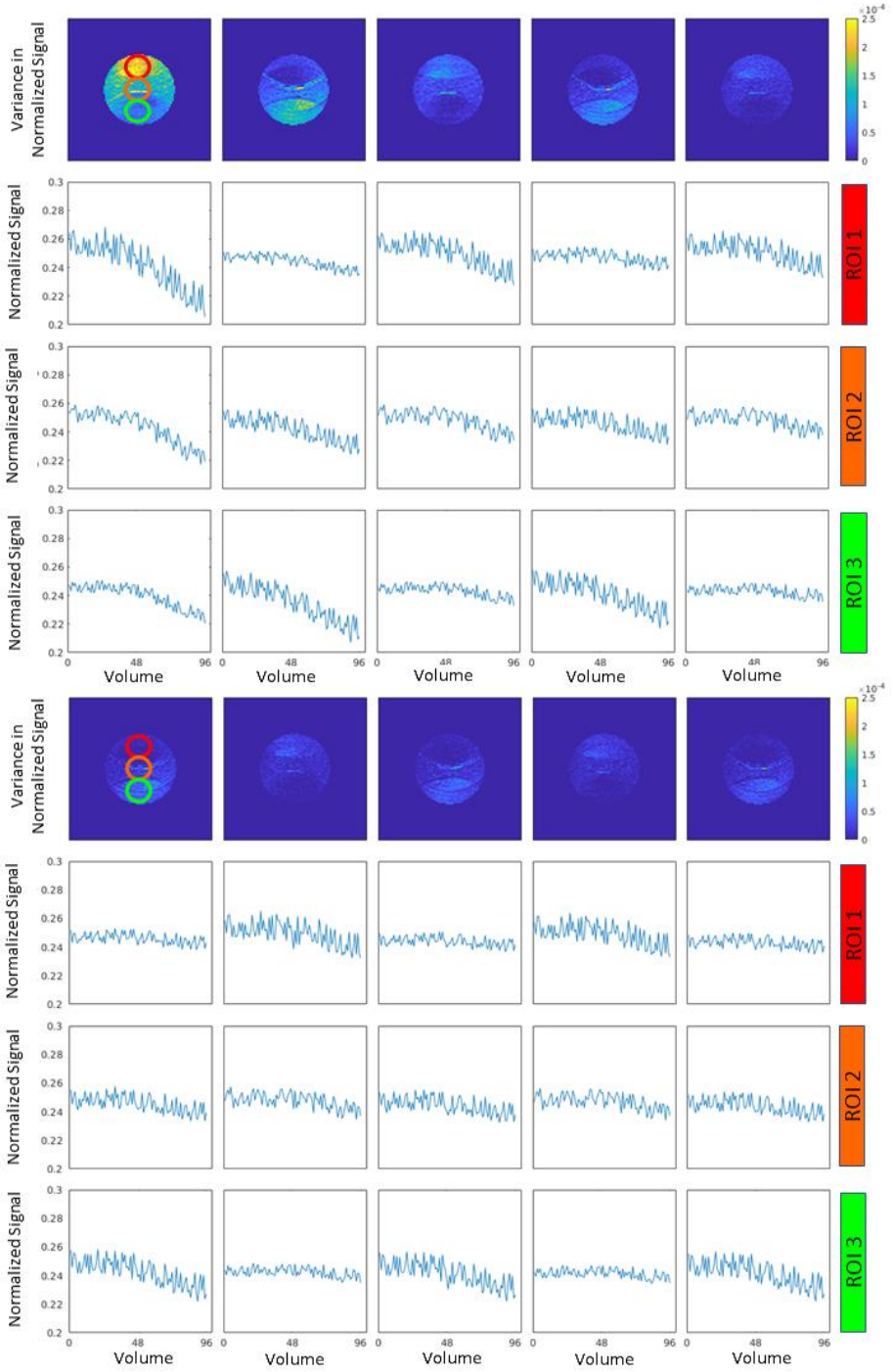


**Figure III-2. Empirical characterization of drift in the PVP phantom. This plot presents the variance in the PVP phantom in 10 different scans over the ¬course of one session (1st and 4th row) and the normalized signal intensity within three ROI's (indicated in the top left plot). The top four rows represent the first five scans in the series with the bottom four representing the last five scans. Observe the structural patterns in variance map that appear to correspond to parallel imaging artifacts. As with the ice water phantom (Figure 1), different ROIs within same scan are drifting at different rates, and spatial correction is required. Similarly, the pattern of drift changes with time and with region.**

To account for these effects, we proposed two key modifications. First, we generalize the ROI-temporal model to operate on a voxel-wise basis to provide for a higher degree of spatial specificity in the correction while alleviating the need for external context information (i.e., the specification of ROI's). Second, we propose a data-efficient model to capture the interaction effects between spatial and temporal nonlinearities. Herein, we combine these ideas to present a novel temporal-spatial model (TS) that accounts for the temporal instability of the scanner and spatial variation in the signal drift. We compare the new approach to uncorrected data, the temporal model (T) as proposed by Vos et al. [142], and a custom generalization of the Vos et al. that models nonlinearities on a voxel-wise basis. The novel method yields greater improvement in error than the alternative approaches. This work highlights the need to capture interleaved b0 data in DWMRI and provides an effective model to capture patterns of signal drift within a scan while yielding more accurate estimation of directional ADC.

## 2. Data

### 2.1. Acquisition

Figure 3 outlines the process for acquisition and processing. Images are acquired on a Philips 3T MRI system. For both the ice-water phantom and the PVP phantom, 10 scans were acquired in a single session with 96 gradient directions at a b-value of 2000 s/mm$^2$ and with a variable number of minimally weighted volumes of b=0.1 s/mm$^2$ interspersed throughout the scans. The acquisition parameters for all scans are as follows: b-value of 2000 s/mm$^2$, interspersed b-value of 0.1 s/mm$^2$, TR of 8394 ms, TE of 70 ms, SENSE acceleration of 2.5, slice thickness of 2.5 mm, and in-plane pixel dimensions of 2.5 mm. The number of these minimally weighted volumes was decreased every two scans giving a pair of scans of opposite phase encoded directions (left phase encoding and right phase encoding) with 13, 7, 4, 3, and 2 minimally weighted volumes at the beginning, end, and interspersed among the 96 directions in that chronological order.

### 2.2. Preprocessing

Each pair of scans with opposite phase encodings and the same b-value was concatenated for preprocessing which consisted of topup for susceptibility distortion correction [134] and eddy current correction [135]. Within each scan, relative signal intensity images were computed by normalizing the diffusion weighted volumes by the first minimally weighted volume ("b0") of the scan. Note that subsequent analyses were performed with additional b0-correction as described below. Next, each scan was registered to the T1 structural MRI of the phantom. The structural

MRIs were manually labeled using custom scripts in MATLAB (Mathworks, Natick, MA). For the ice-water phantom, the 13 ROIs correspond to the 13 vials of varying PVP concentrations. The spherical PVP phantom is simply filled with PVP at one concentration, and three spherical ROIs were defined within it to visually correspond to different regions as shown in Figure 2.

### 3. Methods

Each of the following correction methods was independently applied to each of the acquisitions, which resulted in five different output 4-D volumes for each scan: uncorrected images, images corrected using the Vos et al temporal model (T), images corrected using the Vos et al temporal model generalized to voxels ($T_x$), and images corrected using the temporal/spatial model (TS).

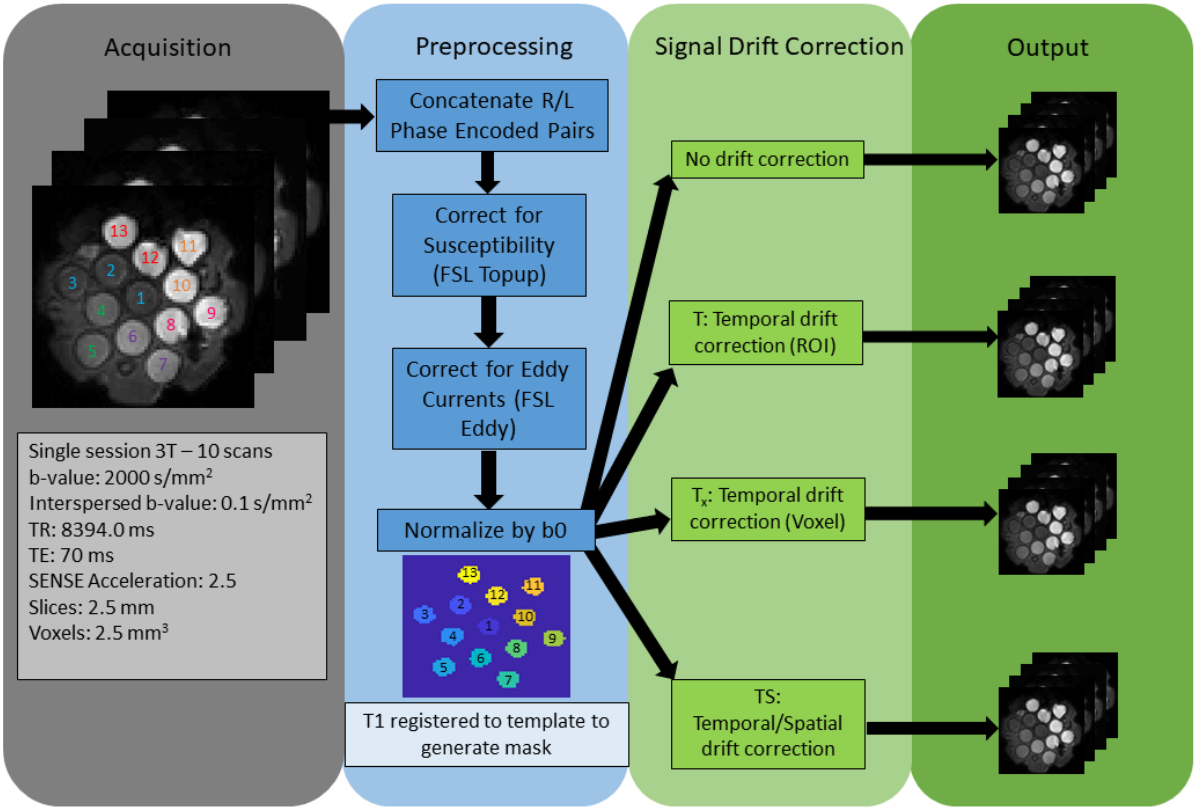


**Figure III-3. This shows our process from acquisition to correction. From the left, the acquisition parameters are shown, then the preprocessing pipeline, and finally the three different outputs resulting from the three different methods.**

### 3.1. Uncorrected

No additional was processing was performed.

### 3.2. Vos et al. Temporal Model (T)

The model defined by Vos et al. estimates a global signal decrease with a linear or quadratic fit through the mean signal intensities of the interspersed b0 images within a region of interest. Figure 2 illustrates that the observed drift is non-linear, and so the linear model would not suffice. However, we apply a linear model when the number of b0 images is less than 3. The linear and quadratic models are defined for the ROI using the b0 volumes among the scanned images by:

$$S(n) = dn + s_0 \tag{1}$$

and

$$S(n) = d_2 n^2 + d_1 n + s_0 \tag{2}$$

respectively where $n$ is the index of the volume; $S(n)$ is the mean signal within the b0 image at time index $n$ for the ROI; $d$, $d_1$, and $d_2$ are the modeled signal drift coefficients; and $s_0$ is the signal offset at the b0 image where $n = 0$. After the linear fitting of $s_0$, $d_1$, and $d_2$, a rescaling factor is used to correct each ROI:

$$\hat{S}\,(n) = S(n)\frac{100}{dn+s_0} \tag{3}$$

and

$$\hat{S}\,(n) = S(n)\frac{100}{d_2 n^2+d_1 n+s_0} \tag{4}$$

where $\widehat{S_j}\,(n)$ is the corrected signal intensity in the image normalized to 100 for quadratic corrections [142]. The linear model (Eq. 3) was applied when three or fewer b0s were acquired, while quadratic model was applied when more b0's were available.

### 3.3. Vos et al. Generalized to Voxels ($T_x$)

The Vos model was modified to fit the signal time course to each voxel allowing free form spatial correction. In this case a single mask is provided which denotes voxels that are in any of the ROIs. The linear and quadratic models then become:

$$V(X,n) = d_{\mathrm{X}} n + v_{0,\mathrm{X}} \tag{5}$$

and

$$V(X,n) = d_{1,\mathrm{X}}n^2 + d_{2,\mathrm{X}}n + v_{0,\mathrm{X}} \quad (6)$$

where $V(X,n)$ is the signal intensity of the voxel at the xyz-coordinate in the image; $X$ is the vector that specifies the xyz-coordinate; $d_{\mathrm{X}}$, $d_{1,\mathrm{X}}$, and $d_{2,\mathrm{X}}$ are the modeled signal drift coefficients; and $v_{0,\mathrm{X}}$ is the signal offset for the voxel in the b0 image where $n$ =0. The rescaling becomes:

$$\hat{V}(X,n) = V(X,n)\frac{100}{d_{\mathrm{X}}n+v_{0,\mathrm{X}}} \quad (7)$$

and

$$\hat{V}(X,n) = V(X,n)\frac{100}{d_{1,\mathrm{X}}n^2+d_{2,\mathrm{X}}n+v_{0,\mathrm{X}}} \quad (8)$$

where $V(X,n)$ is the uncorrected intensity for the voxel at the xyz-coordinate in image $n$, and $\hat{V}(X,n)$ corrected signal intensity for that voxel.

### 3.4. Temporal/Spatial Model (TS)

s We propose a new temporal/spatial (TS) model to take into account the temporal effects in the Vos model (e.g., Eq. 5-8), the spatial effects (e.g., as captured by the $v_{0,\mathrm{X}}$ terms), and their interactions. The degrees of freedom in a model that allows for voxel-wise variations in the interaction between temporal and spatial effects would quickly become untenable. Rather, we propose to use a second order Chebyshev polynomial decomposition of the spatial effects, while interacting with a polynomial expansion of temporal effects (as in Eq. 5 and 6). In the linear TS model, the basis functions are:

$$B_T(X,n) = d_1 n \quad (9)$$

$$B_{TSv}(X,n) = v_{0,\mathrm{X}}(d_2 x + d_3 y + d_4 z + d_5 xy + d_6 xz + d_7 yz + d_8 xyz) \quad (10)$$

$$B_{TSn}(X,n) = \mathrm{n}(d_9 x + d_{10} y + d_{11} z + d_{12} xy + d_{13} xz + d_{14} yz + d_{15} xyz) \quad (11)$$

and the combined linear TS model is:

$$V(X,n) = B_T(X,n) + B_{TSv}(X,n) + B_{TSn}(X,n) + v_{0,\mathrm{X}} \quad (12)$$

For the quadratic model, the basis functions are:

$$B_T(X,n) = d_1 n^2 + d_2 n \quad (13)$$

$$B_{TSv}(X,n) = v_{0,X}(d_3x + d_4y + d_5z + d_6xy + d_7xz + d_8yz + d_9xyz + d_{10}x^2 + d_{11}y^2 + d_{12}z^2 + d_{13}xy^2 + d_{14}xz^2 + d_{15}xy^2z + d_{16}xyz^2 + d_{17}xy^2z^2 + d_{18}x^2y + d_{19}x^2z + d_{20}x^2yz + d_{21}x^2y^2 + d_{22}x^2z^2 + d_{23}x^2y^2z + d_{24}x^2yz^2 + d_{25}x^2y^2z^2 + d_{26}y^2z + d_{27}yz^2 + d_{28}y^2z^2) \quad (14)$$

$$B_{TSn}(X,n) = n(d_{29}x + d_{30}y + d_{31}z + d_{32}xy + d_{33}xz + d_{34}yz + d_{35}xyz + d_{36}x^2 + d_{37}y^2 + d_{38}z^2 + d_{39}xy^2 + d_{40}xz^2 + d_{41}xy^2z + d_{42}xyz^2 + d_{43}xy^2z^2 + d_{44}x^2y + d_{45}x^2z + d_{46}x^2yz + d_{47}x^2y^2 + d_{48}x^2z^2 + d_{49}x^2y^2z + d_{50}x^2yz^2 + d_{51}x^2y^2z^2 + d_{52}y^2z + d_{53}yz^2 + d_{54}y^2z^2) \quad (15)$$

$$B_{TSn^2}(X,n) = n^2(d_{55}x + d_{56}y + d_{57}z + d_{58}xy + d_{59}xz + d_{60}yz + d_{61}xyz + d_{62}x^2 + d_{63}y^2 + d_{64}z^2 + d_{65}xy^2 + d_{66}xz^2 + d_{67}xy^2z + d_{68}xyz^2 + d_{69}xy^2z^2 + d_{70}x^2y + d_{71}x^2z + d_{72}x^2yz + d_{73}x^2y^2 + d_{74}x^2z^2 + d_{75}x^2y^2z + d_{76}x^2yz^2 + d_{77}x^2y^2z^2 + d_{78}y^2z + d_{79}yz^2 + d_{80}y^2z^2) \quad (16)$$

and the combined quadratic TS model is:

$$V(X,n) = B_T(X,n) + B_{TSv}(X,n) + B_{TSn}(X,n) + B_{TSn^2}(X,n) + v_{0,X} \quad (17)$$

where $B_T(X,n)$ has the same temporal components as T, and $B_{TSv}(X,n)$, $B_{TSn}(X,n)$, and $B_{TSn^2}(X,n)$ are the spatial-temporal components of the model that come from the cross product of three second order Chebyshev polynomials, each dealing with either the x, y, or z coordinate of the voxel. The rescaling in this method is defined by:

$$\hat{V}(X,n) = V(X,n)\frac{100}{B_T(X,n)+B_{TSv}(X,n)+B_{TSn}(X,n)+v_{0,X}} \quad (18)$$

and

$$\hat{V}(X,n) = V(X,n)\frac{100}{B_T(X,n)+B_{TSv}(X,n)+B_{TSn}(X,n)+B_{TSn^2}(X,n)+v_{0,X}} \quad (19)$$

To better account for outliers, all coefficients were estimated using robust bi-square regression [154].

### 3.5. Analysis

For the analysis of the methods, values were calculated in terms of signal intensity and ADC. For the ice-water phantom, ROIs corresponding to three vials of different PVP concentrations and spatial locations were chosen

and are denoted by the ROI numbers 3, 7, and 11 as shown in Figure 1. As for the spherical PVP phantom, the three ROIs selected correspond to the spherical ROIs as shown in Figure 2. For each diffusion volume in the 10 scans and for each method, the mean signal intensity and mean ADC was calculated for each ROI. The measurement of error within each ROI for every scan was calculated as the standard deviation of a third-degree polynomial fit to the mean ADC over the course of the scan. This metric quantifies the amount of residual drift across time. It also captures the variation from the isotropic properties of each ROI while ignoring the signal-to-noise ratio (SNR) variation between ROIs especially those of different PVP concentrations in the ice-water phantom. For the analysis of the number of b0 images needed, the same measurement of error is used, but only TS corrected images are used. The scans corrected using fewer than 13 interspersed b0 images for this analysis were artificially created by removing b0 images from both of the first two scans (those with 13 b0 images). To analyze the statistical difference between all methods, the ADC for all voxels in all scans in a valid ROI were considered for each method in a Wilcoxon rank sum test. In this way a p-value was generated for each pairing of methods.

**4. Results**

For both phantoms, the signal drift in ROIs of higher variance as seen in Figures 1 and 2 reached and sometimes surpassed 10% over the course of a scan without any correction. Correcting for the global signal drift as in T does result in improvement over the uncorrected method in most ROIs. However, without accounting for the different and rates of the drift, the method does not reduce the error in all ROIs. $T_x$ and TS show similar performances with small differences.

The mean normalized signal intensity and the corresponding standard deviation across the entire session (all ten scans) is reported in Figure 4 and Figure 5 for the ice-water and PVP phantoms respectively. In both phantoms, $T_x$ and TS have similar signal intensities. In the ice-water phantom the uncorrected method and T show a small difference. T performs more closely to $T_x$ and TS for the PVP phantom.

In Figure 6, the standard deviation of a third-degree polynomial fit to the mean ADC across diffusion volumes is reported for each method just as Figure 7 reports the same for the PVP phantom. The results for the ice-water phantom show the obvious shortcoming of correcting for the global signal drift. When a scan's error is particularly large with no correction, $T_x$ and TS outperform T, and in some cases when the ROI does not follow the global signal drift, T does worse than the uncorrected method whereas $T_x$ and TS reduce the error. Figure 6 also shows the effects of leaving out b0 images from the first two scans (those with 13 interspersed b0s) and applying signal drift correction

using TS. At the point where fewer than four b0 images are used, the linear model is used and is shown to be less stable in the two ROIs shown to have a higher variance in Figure 1.

Figure 7 shows similar results for the PVP phantom. The signal drift in the ROIs in this phantom agree with the direction of the global signal drift, but the rates at which the signal drift occurs within the ROIs varies. As a result, the errors in T and $T_x$ and TS are not as significantly different—see for example ROI 7 (middle row) in Figure 6. The central ROI (ROI 2) agrees closely with the global signal drift, but the outer two ROIs show improvement when using $T_x$ or TS as the spatial information becomes more necessary. In the right column in Figure 7, the effects of leaving out b0 images from the two scans that initially had 13 b0 images shows different results for the scan with right phase encoding. The scan with a left phase encoding direction shows an upward trend in error as fewer b0 images are used in all ROIs, but the scan with right phase encoding does not show a significant change (see Table 2 for all significance levels).

It can be seen from Figures 6 and 7 that even the scans with errors corresponding to a 10% difference in ADC due to drift can be corrected to error levels corresponding to a difference of less than 5%. Though methods Tx and TS perform very similarly in terms of error, it should be noted that the parameter space needed for fitting the Tx model is far greater than that of TS. Tx requires two parameters per voxel for the quadratic model while TS only requires a total of 81 parameters.

**5. Discussion**

Though Vos et al. most recently explored the signal drift caused by temporal instability of scanner systems in DWMRI, signal drift has been an issue in imaging systems that many have attempted to address. In functional magnetic resonance imaging, the effects of signal drift have been observed resulting in a few different methods for correction [155]. Gram-Schmidt orthogonalization [156] and high-pass filtering [157] have been used to eliminate signal drift, but other methods are very similar to $T_x$ in that they model the temporal drift in a voxel-wise manner using either a polynomial [158], a spline [159], or a wavelet [160]. Another method assumes that all voxels follow the trend of the global signal much like T and removes the global signal drift from each voxel [161]. However, we have seen that the assumption of a global trend does not always hold. As mentioned by Vos et al., most of these methods are fMRI specific as the drift is included as a confounding factor [142], but the premises can be adapted to DWMRI. Correction on an ROI basis is restricted by the defined ROIs, and therefore is not a commonly used method. TS is unique in that it models the spatial variation in the temporal drift where $T_x$ allows the temporal models to be spatially

independent and freely formed within a voxel.

It can be noted in Figure 6 that the error drops in the rlr phase encoded scan in the first ROI as two interspersed b0s are used instead of three, but the error is still higher than when using four or more b0s. Also, in Figure 7, the rlr phase encoded scans do not seem to have a particular trend as fewer b0s are used for correction. However, compared to the rll phase encoded scans denoted by 13L in the left column of Figure 7, the correction methods did not have as significant of an effect in the upper and lower ROIs for the rlr phase encoded scans denoted by 13R. We would expect if the resulting error is significantly reduced using all b0s available to the correction method then the error would have an upward trend as fewer b0s are used.

We find that we can correct even rather severe signal drift produced in an imaging situation intended to highlight the effects of the drift (e.g., isotropic phantoms with a large number of diffusion weighted volumes). Yet, to perform this correction a b0 is acquired at least every 32 volumes (or a minimum of 4 b0 volumes to support robust fitting of quadratic temporal models). For pragmatic technical reasons on the Philips systems, we use low (<5 s/mm$^2$), but non-zero, diffusion sensitized volumes to enable repeated b0's interleaved with volumes of higher diffusion weighting. Given the variable degree of signal drift observed within an imaging session and across similar phantoms, we strongly advocate for self-correction of individual datasets through interspersed b0's. Fortunately, these data can be acquired without time penalty as standard practice is already to acquire b0's in an approximate ration of 8:1.

**Table III-1. For each method and for each phantom the median of the errors from all ROIs indicated in Figures 1 and 2 is reported here along with the inter quartile range (IQR) of the errors. Additionally symbols have been placed next to the median values to indicate the rejection of the hypothesis that the method is equivalent to Uncorrected (*), T (†), or Tx ( ‡). The hypotheses were evaluated at a significance level of 5% using the Wilcoxon rank sum statistical test. Here we see that TS has a lower median error, but Tx shows the lowest IQR.**

| Method | Ice-water Med. Error | IQR | PVP Med. Error | IQR |
|---|---|---|---|---|
| | x $10^{-5}$ | x $10^{-5}$ | x $10^{-5}$ | x $10^{-5}$ |
| Uncorrected | 0.35 | 0.35 | 0.85 | 0.76 |
| T | 0.29 * | 0.37 | 0.51 * | 0.23 |
| $T_x$ | 0.16 * | 0.05 | 0.35 *† | 0.23 |
| TS | 0.16 *†‡ | 0.10 | 0.28 *†‡ | 0.29 |

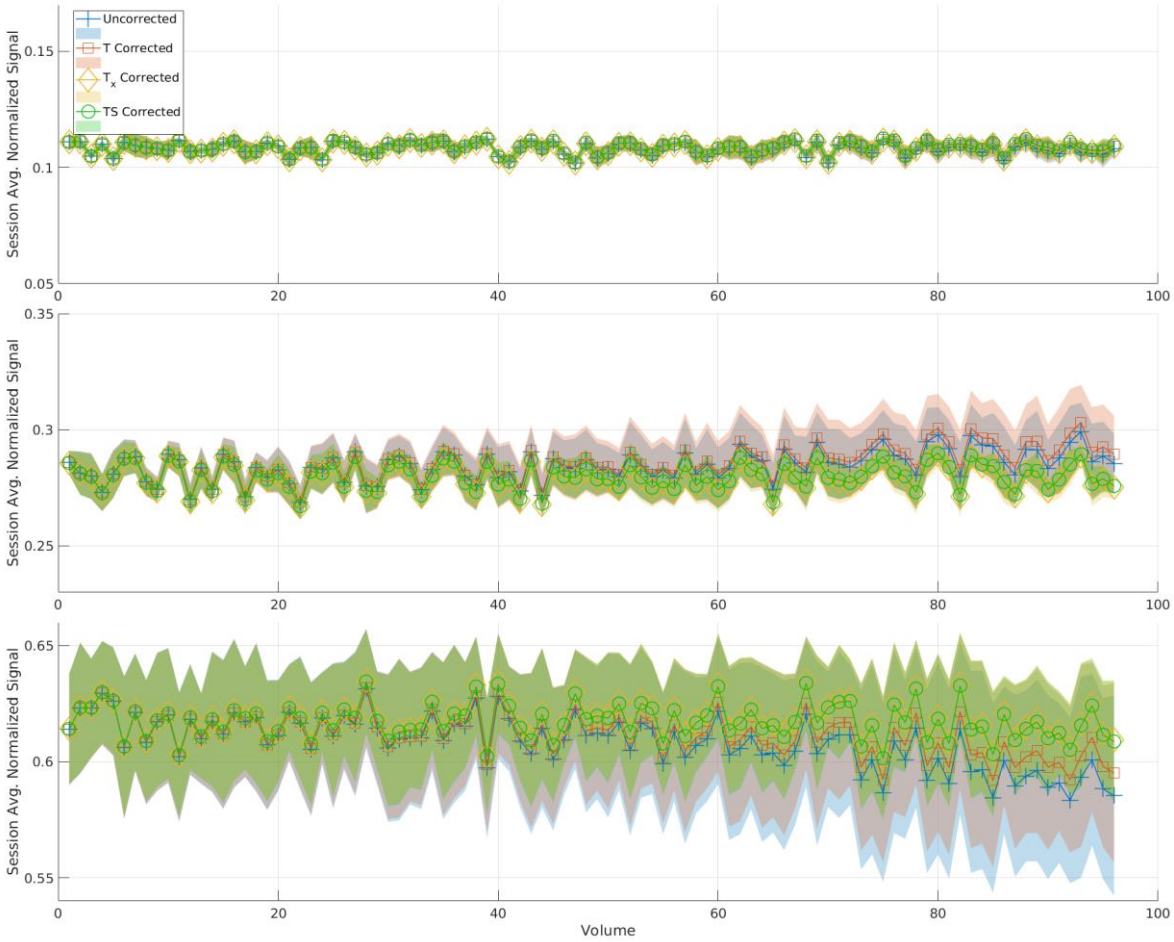


**Figure III-4. This plot represents the mean signal across the 10 scans in the session for the ice-water phantom for each method. Each line also has a shaded area representing the standard deviation among those scans at each volume number. Each plot represents one of the three selected ROIs from Figure 1. From top to bottom those ROIs are 3, 7, and 11. In ROI 3 there is no difference between the methods, but in the ROIs that show higher variance in Figure 1, the uncorrected method and T deviate from Tx and TS.**

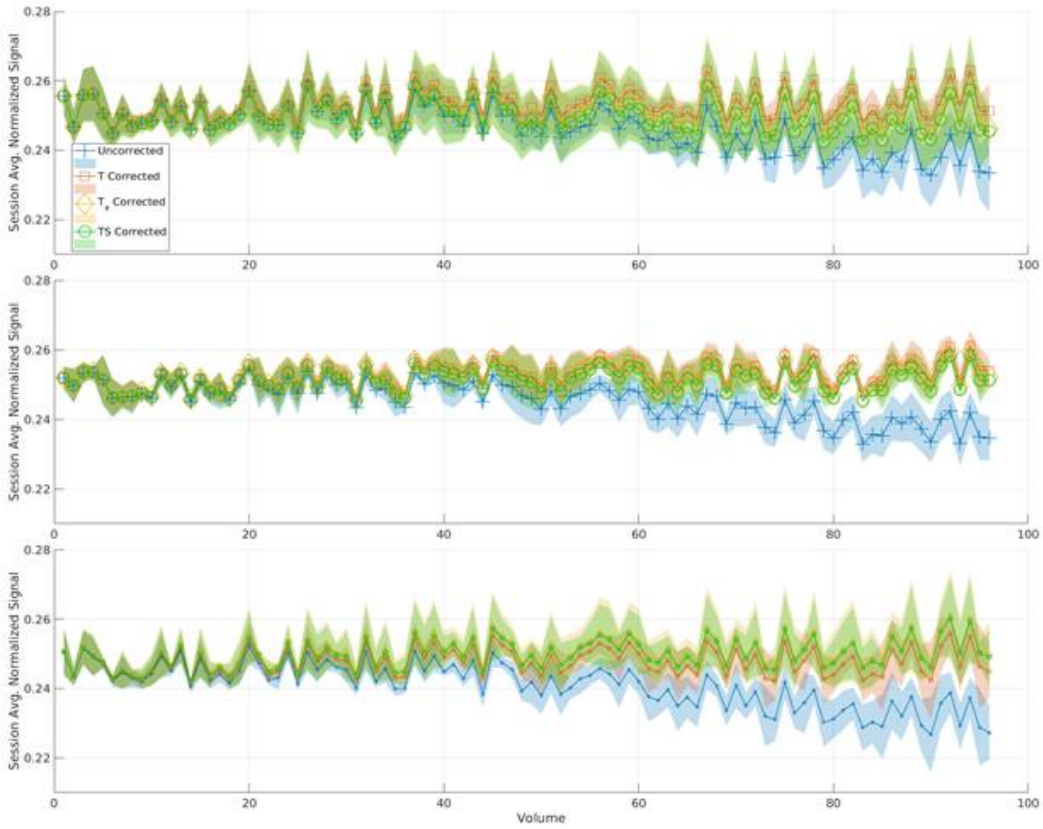


**Figure III-5. This plot represents the mean signal across the 10 scans in the session for the PVP phantom for each method. Each line also has a shaded area representing the standard deviation among those scans at each volume number. Each plot represents one of the three selected ROIs from Figure 1. From top to bottom those ROIs are 1, 2, and 3. In all 3 ROIs there is a noticeable difference between the uncorrected method and the correction methods. In ROI 1 and 3 there is also a slight difference between T and methods Tx and TS.**

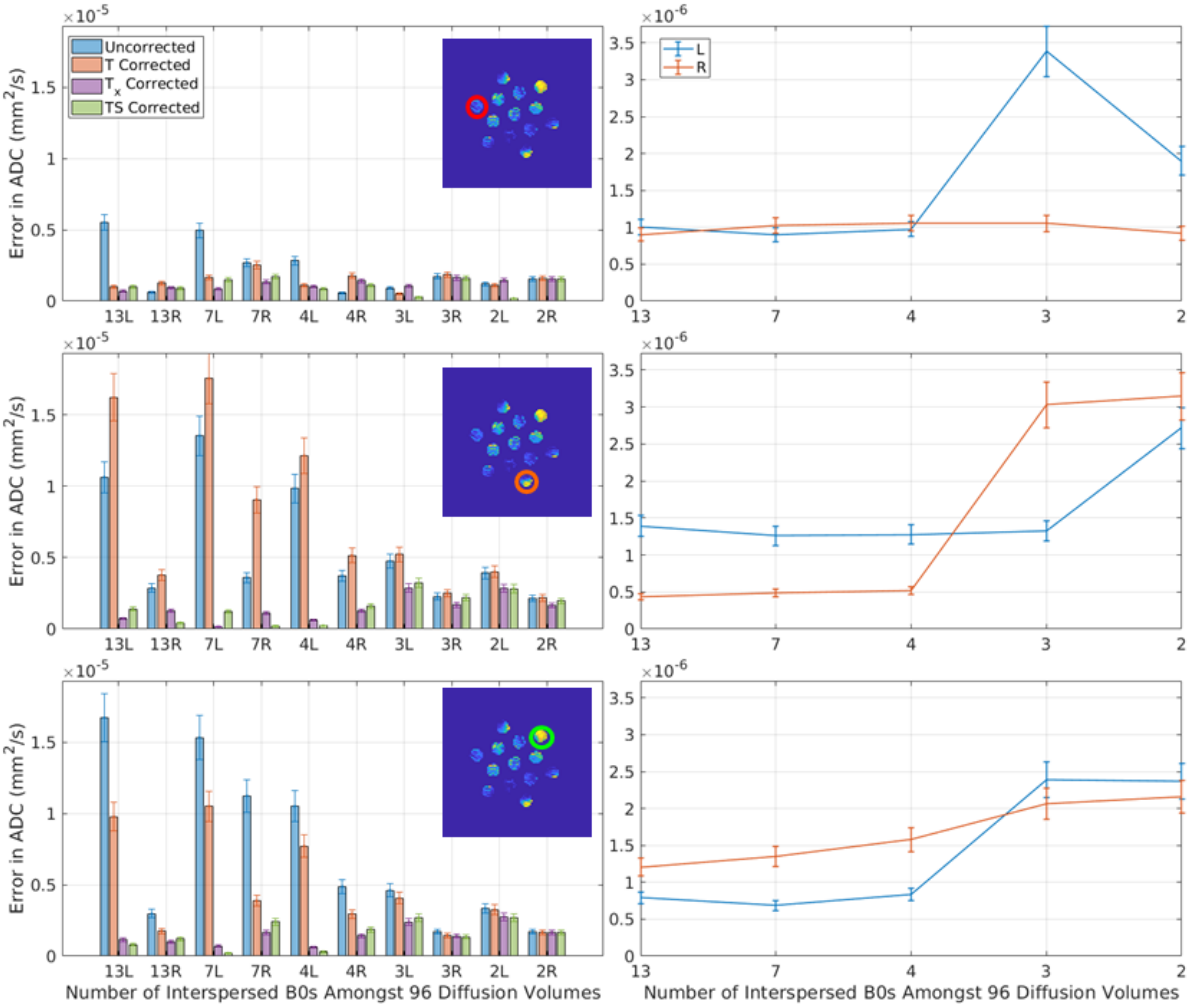


**Figure III-6. This plot presents the average error in ADC (standard deviation in a 3rd degree polynomial fit to the mean ADC of an ROI over the course of a scan) after correction for each of the five methods for 10 consecutive sessions using the ice water phantom with varying numbers of interspersed minimally weighted ("b0") volumes (labeled in the x-axis). The appended letter of the x-axis label indicates the phase encoding direction (L = rll R= rlr). The three rows correspond to the three ROIs in Figure 1, as indicated. The left column presents a comparison of the five methods. In the low variance ROI (first row), overall errors are small and little difference is observed between methods. In the two ROIs of higher variance, Tx, and TS outperform T for all scans. Note that in some scans, the uncorrected method outperforms the T corrected scans. The right column studies simulated rate of b0 volumes by dropping out the b0s from the first two scans. Observe that with at least 4 b0s, the model errors are stable and low, which is intuitive as a second-degree model is fit for #b0s>3 and a first-degree model is fit for #b0<=3.**

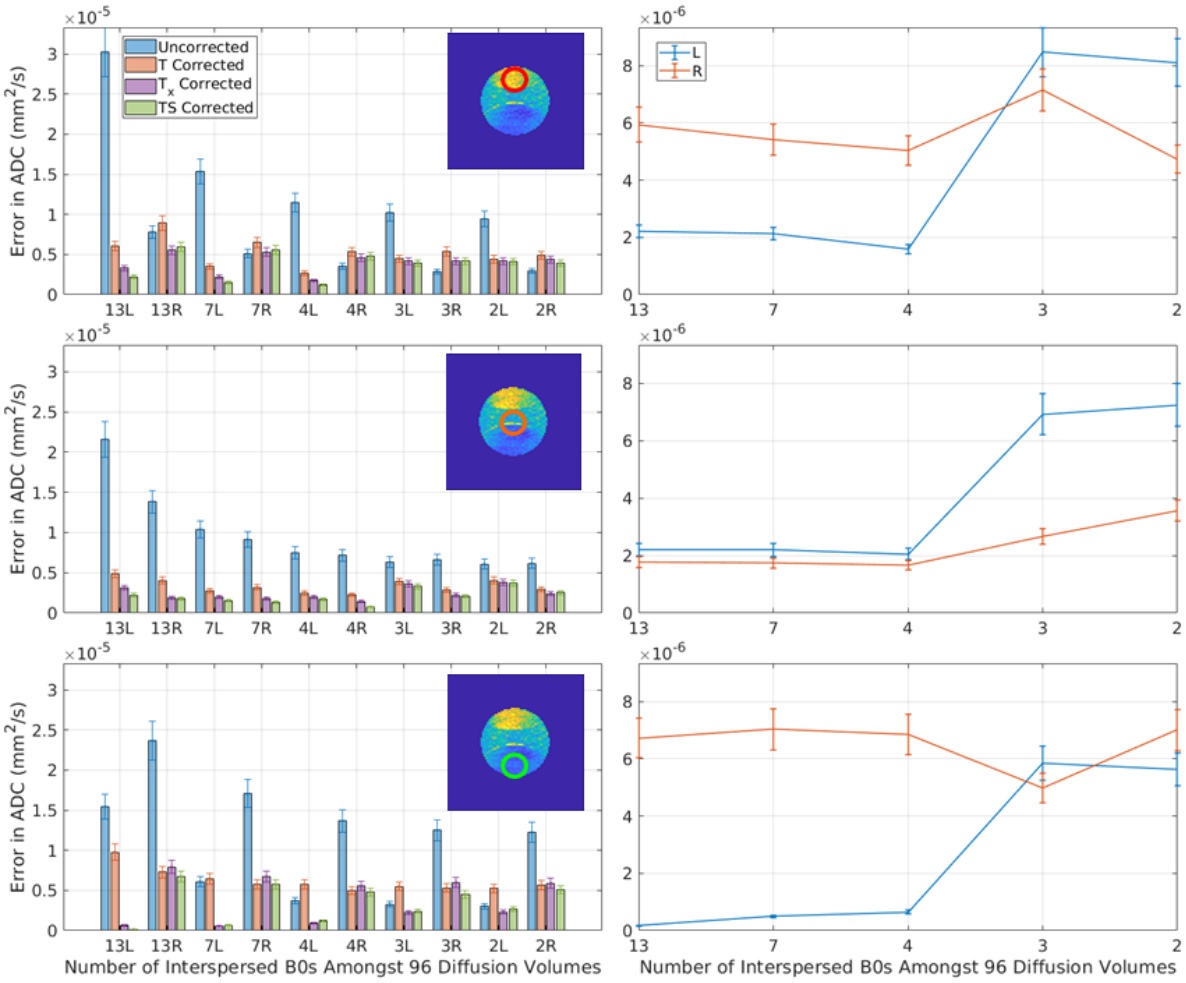


**Figure III-7. This plot presents the error in ADC (standard deviation in a 3rd degree polynomial fit to the mean ADC of an ROI over the course of a scan) after correction for each of the five methods for 10 consecutive sessions using the PVP phantom with varying numbers of interspersed minimally weighted ("b0") volumes (labeled in the x-axis). The appended letter of the x-axis label indicates the phase shift direction (L = rll R= rlr). The three rows correspond to the three ROIs in Figure 1, as indicated. The left column presents a comparison for the five methods. In the middle ROI (second row) T, Tx, and TS show very similar performance as the drift in the center ROI is similar to the average drift across all ROIs. In the outer two ROIs (especially row 3) Tx and TS shows significant improvements over T. The right column studies simulated rate of b0 volumes by dropping out the b0s from the first two scans. Observe that with at least 4 b0s, the model errors are stable and low, which is intuitive as a second-degree model is fit for #b0s>3 and a first-degree model is fit for #b0<=3.**

## Chapter IV. Empirical field mapping for gradient nonlinearity correction of multi-site diffusion weighted MRI

### 1. Introduction

Physics underlying magnetic resonance imaging (MRI) gradient coil designs result in nonuniform magnetic field gradients during acquisition. This leads to spatial image warping [162-165] in magnetic resonance images and gradient distortion in diffusion weighted magnetic resonance imaging (DW-MRI) [166-170]. The introduced spatial variation can impact estimated diffusion tensor information [171] or high-angular resolution diffusion measurements [172]. Bammer et al. show in extreme cases the gradient nonuniformity can lead to an overestimation in the diffusion coefficient up to 30% and an underestimation up to 15% [30]. The severity of the effect increases with distance from the magnet's isocenter [30] and with higher gradient amplitudes [30, 173]. The artifact becomes especially troubling for multi-site studies that have varying scanner models and manufacturers [174] and for studies utilizing very large gradient amplitudes such as in the human connectome project (HCP) which utilized amplitudes up to 300 mT/m [135, 173, 175]. Recent work has shown the effect of gradient nonlinearities in the HCP cohort results in considerable bias in tractography results and potentially incorrect interpretations in group-wise studies [176].

Various estimates of the coil magnetic field nonlinearities have been applied to improve accuracy within and across sites [32, 33, 177, 178]. An adaptive correction of diffusion information proposed by Bammer et al. relies on calculating the spatially varying gradient coil $L$. This approach is achieved by relating the actual gradients with the desired gradients [30], and has become standard practice [179, 180]. However, this approach assumes that the gradient calibration specified by the manufacturer is readily available. Spherical harmonics (SH) based techniques are already implemented by manufacturers in the scanning systems to account for the spatial image warping effects of gradient nonlinearities [31, 162, 181-183]. Yet, the spherical harmonic coefficients are not usually provided to regular users and may be subject to non-disclosure criteria. Additionally gradient nonlinearity correction has been approached using noncartesian MR image reconstruction [184].

To remove the need for the manufacturer supplied specifications, we demonstrate an empirical field-mapping procedure which can be universally applied across platforms as defined by Rogers et al. [29, 141]. At two scanners (scanner A and scanner B), a large oil-filled phantom is used to measure the magnetic field produced by each gradient coil. To estimate the achieved diffusion gradient directions and b-values on a voxel-wise basis, solid harmonic basis

functions are fit to the measured magnetic field. The measured diffusivity (MD) and fractional anisotropy (FA) are compared without nonlinearity correction, with nonlinearity correction using estimated fields, and with nonlinearity

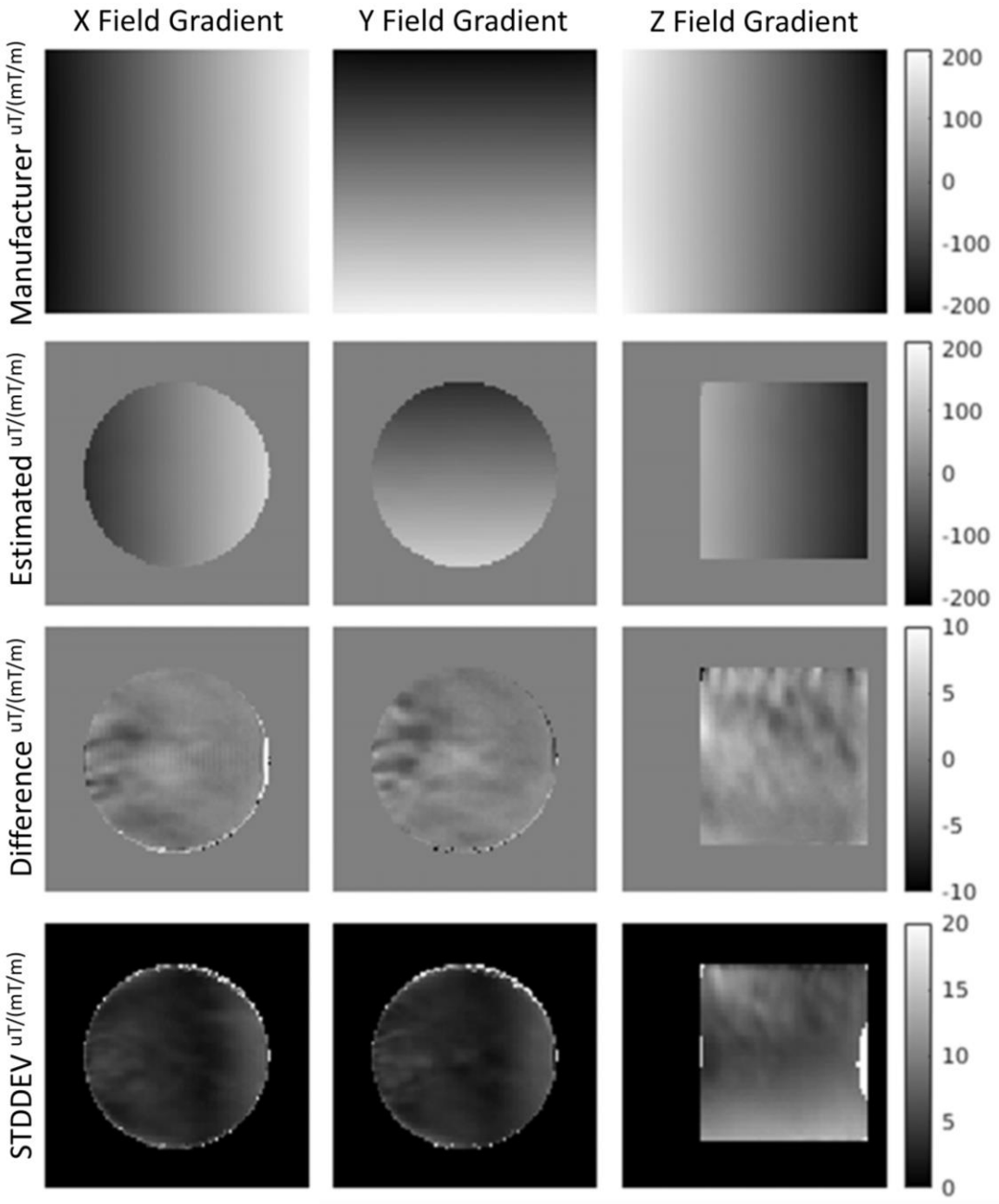


**Figure IV-1. Here we show the manufacturer specified fields (top), the averaged empirically estimated (directly measured) fields (middle-top), the difference between these (middle-bottom), and the standard deviation in the empirically estimated fields across time (bottom) in units of uT (per mT/m of applied gradient). The field of view is 384mm by 384mm, and a mask is applied to the fields according to the usable regions within the oil phantom (135mm radius from isocenter). The x and y magnetic field gradients are shown as an axial slice at isocenter (192mm), and the z magnetic field gradient is shown as a sagittal slice at isocenter (192mm).**

correction using fields specified by the manufacturer for an ice-water diffusion phantom. The reproducibility is compared between without nonlinearity correction and with nonlinearity correction with the estimated fields for a subject scanned at two positions within the scanner at scanner A. We show that our method removes the need for manufacturer specified spherical harmonic coefficients and that the method reduces MD reproducibility error in-vivo when the effect of gradient nonlinearities is present.

## 2. Methods

### 2.1. Measurement of gradient coil-generated magnetic fields

Data were acquired across two 3T scanners: Scanner A and scanner B. Both of these are 94 cm bore Philips Intera Achieva MR whole-body systems and have a gradient strength of 80 mT/m, a 200 T/m/s slew-rate. A phantom is used to estimate the gradient coil fields. The phantom is 24 liters of a synthetic white oil (SpectraSyn 4 polyalphaolefin, ExxonMobil) in a polypropylene carboy with an approximate diameter of 290mm and height of 500mm [29]. This oil is used by the manufacturer for some of their calibration phantoms which made it a reasonable choice. The phantom was placed approximately at scanner isocenter and imaged with a dual echo GRE-based field mapping sequence. Images are acquired at two echo times 1ms apart, and the fieldmap is computed from the phase difference of the two images. This follows the manufacturer's field mapping and provides a field map with minimal phase wrapping or distortion. Four field maps were acquired, one with shim field set to 0.05 mT/m on each axis X, Y, Z plus a final image with gradient coil shim fields set to zero. Each used a 384 mm field of view with 4 mm isotropic voxel size. Total scan time was approximately 5 minutes. Gradient coil fields were estimated by subtracting the zero-shim field map from each coil's respective 0.05 mT/m field map. It should be noted that the proposed method requires that the field maps are made using the same coils used to produce the diffusion gradients, and systems that utilize gradient coil inserts may not be able to directly utilize the technique. Field maps were acquired on 40 dates over the course of a year at scanner B while scanner A only one session was acquired with the fieldmapping phantom.

For each coil, we modeled the magnetic field spatial variation as a sum of solid harmonics [30, 185, 186] to 7th order, excluding even order terms due to the coils' physical symmetry. These basis functions were fit to the field measurements with robust least squares, using all voxels within a 270 mm diameter sphere at isocenter. For comparison, the general shape of the human head is an ellipsoid with an average height of 180 to 200mm [187]. The result was an analytically differentiable estimate of the true magnetic field produced by each gradient coil (Figure 1).

This fitting procedure was performed on an average field map derived from a series of scans to ensure stability. On Scanner B, the fitting procedure is also performed on the scanner manufacturer's estimate of the coil fields as measured during manufacturing and installation. These are provided as a set of solid harmonic functions and corresponding coefficients. The series of scans which are averaged are defined for each subject session according to the closest 10 field map sessions in terms of date for scanner B whereas 10 acquisitions were acquired within a single session at scanner A which are averaged.

### 2.2. Estimating achieved b-values and gradient directions

A spatially varying tensor $L$ relates the achieved magnetic field gradient to the intended one [30]:

$$L = \begin{bmatrix} \frac{\partial B_z^{(x)}}{\partial x} & \frac{\partial B_z^{(y)}}{\partial x} & \frac{\partial B_z^{(z)}}{\partial x} \\ \frac{\partial B_z^{(x)}}{\partial y} & \frac{\partial B_z^{(y)}}{\partial y} & \frac{\partial B_z^{(z)}}{\partial y} \\ \frac{\partial B_z^{(x)}}{\partial z} & \frac{\partial B_z^{(y)}}{\partial z} & \frac{\partial B_z^{(z)}}{\partial z} \end{bmatrix} \quad (1)$$

where $B_z^{(x)}$ is the z component of the magnetic field produced by unit amplitude of a nominal x-gradient coil current, and similarly for (y) and (z). This tensor may be computed analytically from the solid harmonic approximation to the measured field, then evaluated at spatial locations of interest. We can use $L$ to relate the assumed gradient vector to the achieved gradient field and as well as the assumed b-value to the achieved one. If we assume $|g| = 1$ then the adjusted gradient vector and b-value become:

$$g' = Lg \quad (2)$$

$$g'' = \frac{g'}{|g'|} \quad (3)$$

$$b' = b|g'|^2 \quad (4)$$

where $b'$ is the adjusted b-value and $g''$ is the adjusted and normalized gradient vector. In the common situation where the scanner reports the intended gradient direction and amplitude but the full b-matrix [188-190] is not known, an approximate correction to adjust the signal $S_i$ for the $i^{\text{th}}$ diffusion acquisition relative to the reference signal $S_0$ is [177]:

$$\ln\left(\frac{S_i}{S_0}\right) = -bg_i'^T D g_i' = -bg_i^T L^T D L g_i \quad (5)$$

where $b$ is the scalar b-value, $g$ is the intended gradient vector, $g'$ is the actual gradient vector, and $D$ is the diffusion tensor. If we substitute with $b'$ and $g''$ equation 5 can be re-written as:

$$\ln\left(\frac{S_i}{S_0}\right) = -b' g_i^{"T} D g_i^{"} = -b|g_i'|^2 \frac{g_i'^T}{|g_i'|} D \frac{g_i'}{|g_i'|} = -b g_i^T L^T D L g_i \quad (6)$$

Importantly, this is spatially varying and processing occurs voxel-wise, but this may be used in any desired way for further processing of the diffusion images. Figure 2 shows $L$ for each voxel estimated using our empirical fieldmapping acquired on scanner B.

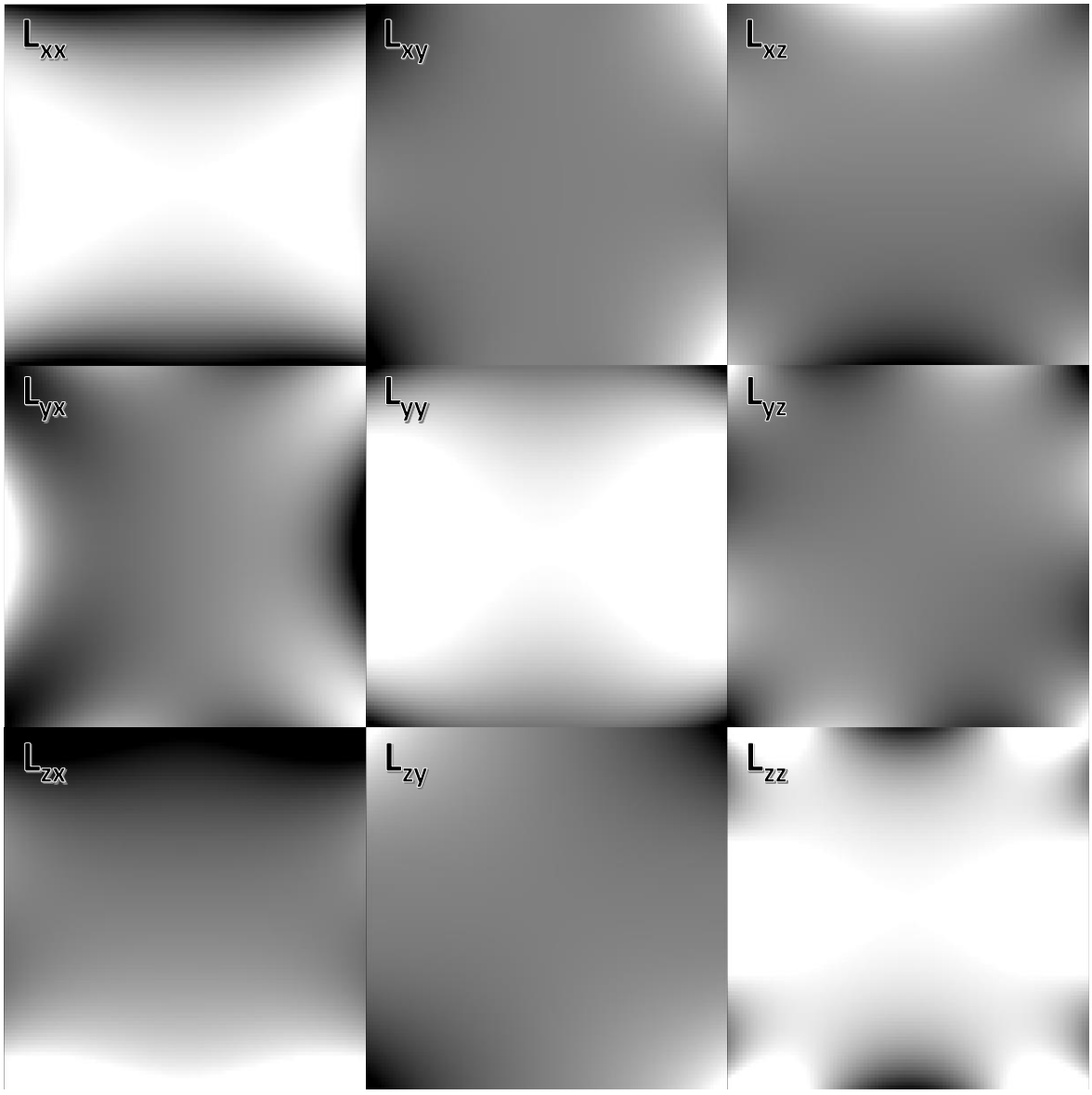


**Figure IV-2. Gradient coil tensor L(r) (sagittal view) for each voxel position using 7th order spherical harmonic expansion using only odd order terms. This was generated using the coefficients estimated from the empirical field mapping procedure.**

## 3. Experiments

This section describes the set of analyses which aim to show the accuracy of the estimated fields as well as their impact on resulting DW-MRI metrics in phantom and human data. All DW-MRI are corrected for susceptibility distortion [134] and eddy current distortion [135] using FSL.

### 3.1. Empirically Estimated Fieldmaps

Gradient nonlinearity correction is only viable if we can depend on the estimation to match the true fields. To investigate if the magnitude estimated fieldmaps closely approximate the true fields, we compare them to the fieldmaps specified by the manufacturer on scanner B. This was not done for scanner A as the manufacturer specifications for scanner A were not provided. For comparison, we take the average fieldmap from the latest 10 oil phantom scans on scanner B and calculate the voxel-wise difference between this and the manufacturer specified fields. To evaluate the stability of the empirical estimations, we report the variance across fields estimated from 40 individual oil phantom scans acquired over time on scanner B. These additional acquisitions are unnecessary for practical use and are strictly for evaluation purposes. Only a single acquisition would be needed for this method to be deployed on a scanner to be applied to all previous and future acquisitions. All evaluations on the empirical fields use a spherical mask with a radius of 135mm from isocenter.

### 3.2. Polyvinylpyrrolidone (PVP) phantom

To evaluate the intra-scanner performance of the gradient field nonlinearity correction with the empirical fieldmaps in a controlled environment, we use a 43% Polyvinylpyrrolidone (PVP) aqueous solution in a sealed spherical container that is 160mm in diameter (PVP phantom) [191]. The PVP phantom is a large homogeneous material, and estimated metrics are expected to be the same across the entire volume. Additionally, toxicology has shown PVP to be safe for use, and PVP is stable and uniform. At scanner B, the phantom was scanned at three positions along the magnet axis: superior (4cm above isocenter), isocenter, and inferior (8cm below isocenter). At each position DWI data was acquired with diffusion weighting applied in twelve directions at a b-value of 1000 s/mm$^2$ and twelve more were acquired at 2000 s/mm$^2$ with a TR of 7775, a voxel resolution of 2.5mm by 2.5mm by 2.5mm, and a FOV of 240mm by 240mm by 170mm. Susceptibility distortion correction and eddy current distortion correction are applied without movement correction. Signal to noise ratio (SNR) was calculated by fitting the signal to a tensor in the phantom and taking the residuals after the fit. Using all diffusion volumes at each position, MD is calculated without

and with gradient nonlinearity correction using the empirically derived fields and using the manufacturer specified fields. When calculating MD with the correction, the estimated achieved b-values and gradient directions for each voxel are used. We report error in terms of absolute percent error (APE) between each scan out of isocenter and the scan at isocenter. All non-diffusion volumes to a structural T1 image using a rigid body transform restricted to only use translations, and this registration is applied to the calculated MD before analysis.

### 3.3. Human subject

To evaluate the intra-scanner and inter-scanner performance of the gradient field nonlinearity correction with the empirical fieldmaps in-vivo, we scanned a single subject at scanner A and scanner B. At scanner B, two sessions were acquired of the subject with one session acquired with the bridge of the subject's nose positioned at isocenter within the magnet and one session acquired with the subject positioned 6cm superior from isocenter. At scanner A, only one session is acquired at isocenter. Each session consisted of twelve gradient directions at a b-value of 1000 s/mm$^2$, twelve at a b-value of 2000 s/mm$^2$, a TR of 3700ms, a voxel resolution of 2.5mm by 2.5mm by 2.5mm, and a FOV of 240mm by 240mm by 170mm. Susceptibility distortion correction and eddy current distortion correction are applied with movement correction for each session. Using all diffusion volumes from each session, MD is calculated without and with gradient nonlinearity correction using the empirically derived fields. At scanner B, MD is also calculated after correction with the manufacturer specifications. For analysis the scans are registered to a T1 acquired at isocenter using FSL Flirt [192]. We report MD error as the absolute percent error between the two scans acquired at scanner B and between the scan acquired at scanner A and the out of isocenter scan acquired at scanner B.

We also evaluate the performance of the empirical correction with higher quality acquisitions on scanner A. Again, two sessions are acquired of the subject: one with the bridge of the subject's nose positioned at isocenter and one where the subject is shifted 4cm inferior from isocenter. Each session consisted of 384 gradient directions at a b-value of 1000 s/mm$^2$, a voxel resolution of 2.5mm by 2.5mm by 2.5mm, and a FOV of 240mm by 240mm by 170mm. Susceptibility distortion correction and eddy current distortion correction are applied with movement correction for each session. Using all diffusion volumes from each session, MD is calculated without and with gradient nonlinearity correction using the empirically derived fields. For analysis the scans are registered to a T1 acquired at isocenter using FSL Flirt [192]. We report MD error as the absolute percent error between the two scans.

## 4. Results

### 4.1. Empirically Estimated Fieldmaps

There are small differences between the manufacturer and the measured field produced by the gradient coil. These are shown in Figure 1 in units of uT scaled by the intensity (mT/m) of the applied gradient (uT/(mT/m), or mm). On average the difference at a given voxel is approximately 1 uT/(mT/m) in the x and y magnetic field gradients and 2 uT/(mT/m) in the z gradient field within 135mm of isocenter. The difference maps indicate the presence of some structural artifacts. The average standard deviation at a given voxel after 40 acquisitions acquired throughout a year is approximately 4 uT/(mT/m) in the x and y fields and 6 uT/(mT/m) in the z field within 135mm of isocenter.

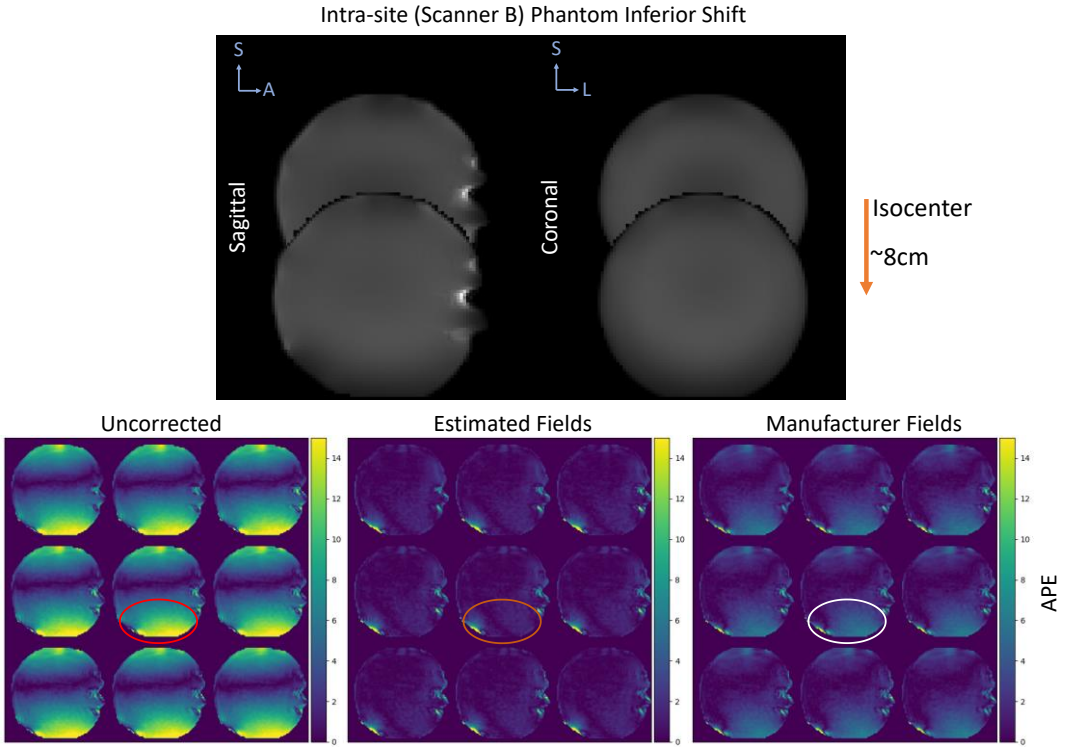


**Figure IV-3. The absolute percent error (APE) in MD is shown for the PVP phantom with one session acquired at isocenter and another acquired 8cm inferior from isocenter. The top plot shows the sagittal and coronal view of the b0 from each session to demonstrate the shift within the scanner. The bottom plots show the APE for nine sagittal slices before correction, after correction using the estimated fields, and after correction using the manufacturer specifications. The error before correction is most prominent in the inferior regions of the phantom as those were the furthest from isocenter during the second acquisition.**

### 4.2. PVP phantom

The mean absolute percent error within the phantom between the inferior scan and the isocenter scan is approximately 5% before correction. After correction using the manufacturer fields, this falls to approximately 1.6%. Correcting with the empirically derived fields leads to 0.9% mean error. Figure 3 shows most of the error before correction in the inferior regions of the phantom which were furthest from isocenter in the inferior scan.

When uncorrected, the mean absolute percent error within the phantom between the superior scan and the isocenter scan is approximately 4.9%. After correction using the manufacturer fields, this falls to approximately 2%. Correcting with the empirically derived fields leads to 1.3% mean error. Figure 4 shows most of the error before correction in the superior regions of the phantom which were furthest from isocenter in the superior scan.

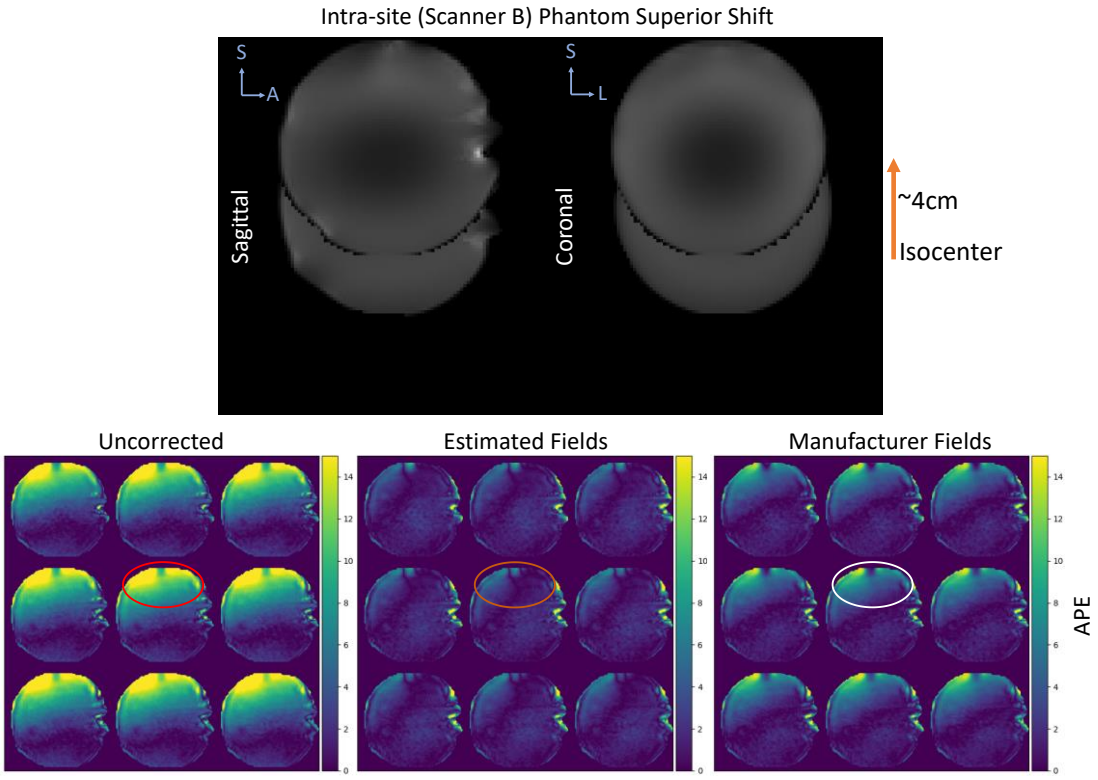


**Figure IV-4. The absolute percent error (APE) in MD is shown for the PVP phantom with one session acquired at isocenter and another acquired 4cm superior from isocenter. The top plot shows the sagittal and coronal view of the b0 from each session to demonstrate the shift within the scanner. The bottom plots show the APE for nine sagittal slices before correction, after correction using the estimated fields, and after correction using the manufacturer specifications. The error before correction is most prominent in the superior regions of the phantom as those were the furthest from isocenter during the second acquisition.**

Additionally, the PVP phantom was corrected using fieldmaps estimated with various orders of solid harmonics. Regardless of the order, both FA and MD reproducibility errors decrease when compared to the

uncorrected error. However, we find that a 3rd order basis results in the lowest FA error but a higher MD error. Between the higher order basis, the 7th order solid harmonics achieves lower MD error (Figure 5).

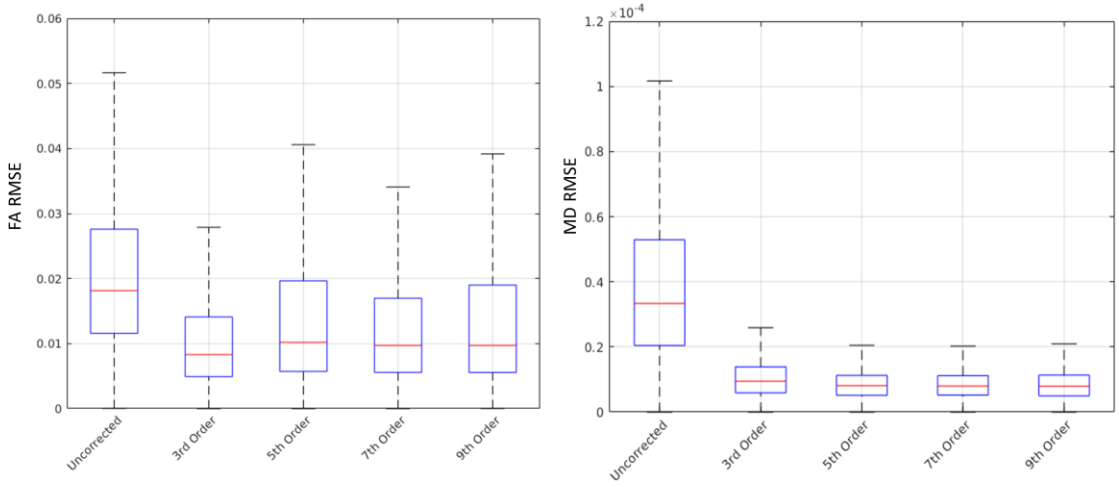


**Figure IV-5. The reproducibility error in FA and MD for the PVP phantom are calculated using the estimated fieldmap utilizing different orders of solid harmonics. Orders higher than 3rd achieve lower MD RMSE but tend to have higher FA RMSE.**

### 4.3. Human repositioned

The intra-scanner sessions on scanner B result in a mean absolute percent error of 5.9% before correction within the brain volume excluding CSF regions. After correcting the scans using the empirically estimated fields, the mean error is reduced to 5.6% and further to 5.4% if the manufacturer specifications are used during correction. Just as in with the phantom, the error attributable to the gradient nonlinearities before correction appears in the superior regions of the brain which were furthest from isocenter during one of the sessions (Figure 6).

For the inter-scanner experiment, the mean absolute percent error before correction is 7.2% and is reduced 6.9% after correction using the estimated fields. Clearly the error that is accounted for in the correction is the superior regions of the brain which were furthest from isocenter during the session acquired on scanner B (Figure 7).

The intra-scanner sessions acquired on scanner A using a significantly higher number of gradient directions results in a mean absolute percent error of 4.6% when no correction is applied. After correction using the empirically estimated fields, the mean error is reduced to 4.2%. The difference can be seen in the inferior regions of the brain, specifically the cerebellum which was furthest from isocenter during one of the sessions (Figure 8). Figure 9 shows the mean absolute percent error across all voxels within the phantom and within the brain volume excluding cerebrospinal fluid (CSF) regions for each method.

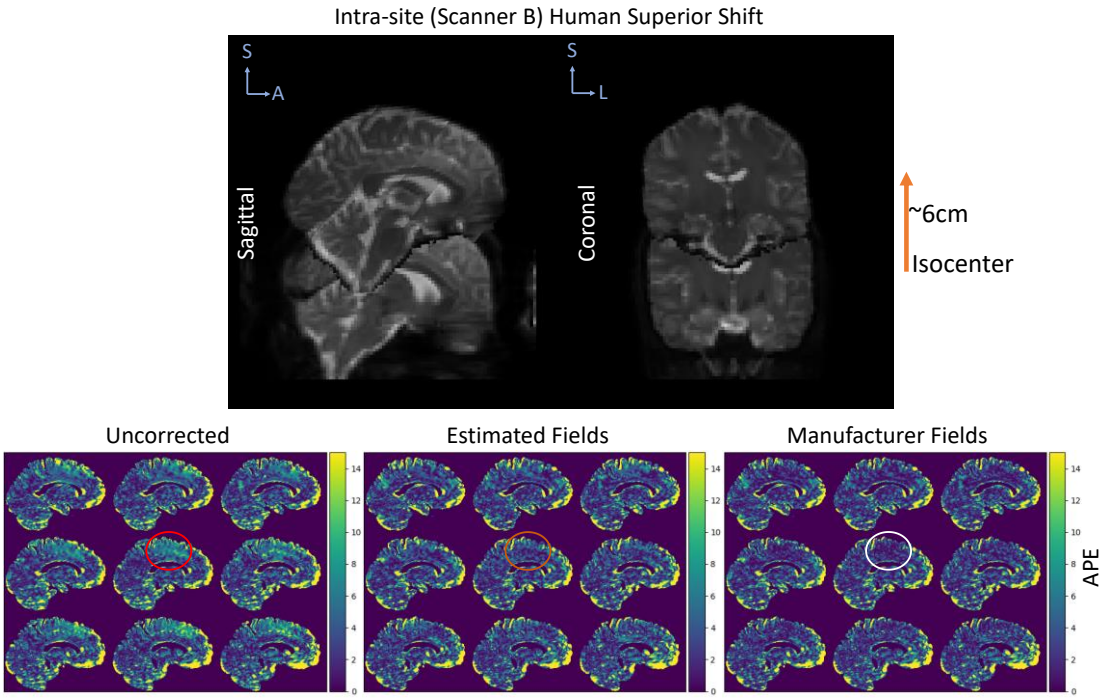


**Figure IV-6. The absolute percent error (APE) in MD is shown for the human subject with one session acquired at isocenter and another acquired 6cm superior from isocenter on scanner B. The top plot shows the sagittal and coronal view of the b0 from each session to demonstrate the shift within the scanner. The bottom plots show the APE for nine sagittal slices before correction, after correction using the estimated fields, and after correction using the manufacturer specifications. The error before correction is most prominent in the superior regions of the phantom as those were the furthest from isocenter during the second acquisition.**

## 5. Discussion

In comparing the empirically estimated fields to the fields specified by the manufacturer, we find that our approximations are very similar. The largest differences are in the z gradient field which corresponds to the largest variations in all the estimated fields across 40 oil phantom acquisitions. In this study we use an average of fieldmaps across 10 acquisitions each acquired a week apart, but this should not be necessary as the field produced by the gradient coil depends only on the coil geometry and the current flowing in the coils. Unaltered system need only acquire the fields once for this method, but further study on the stability of the empirical mapping may be necessary. Additionally, further study on the stability of the fit of the spherical harmonics and the need for higher order basis may be necessary.

The experiments with the PVP phantom show in a large isotropic volume the impact of the gradient nonlinearities within the magnet and the effectiveness of the correction. The small superior shift of 4cm results in over 15% error in the superior voxels. In the case of a large inferior shift and a smaller superior shift, the mean error is increased by a factor of two to five if these effects are not accounted for. If we consider the experiments involving the human subject, we can see the impact of this correction is reduced. This could in part due to imperfect registration which seems to have contributed to error in the anterior regions of the brain. Results may vary depending on registration strategy. We have tried multiple techniques with similar results. Though the absolute percent error only changed by 0.3% to 0.4%, some small regions see a similar magnitude of improvement, and it is qualitatively clear that the correction is impacting regions we expect. The differences between resulting absolute percent error using the empirical fields and the manufacturer fields is varies between the phantom and the human subject. The results for the

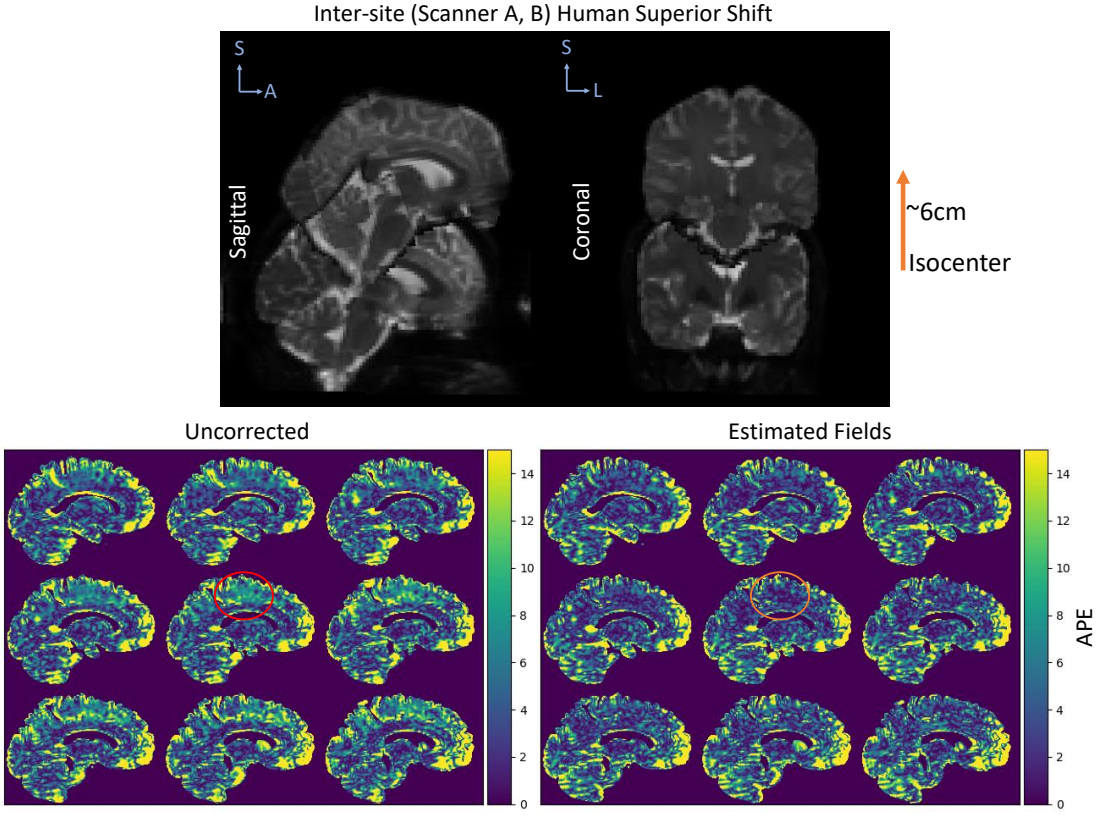


**Figure IV-7. The absolute percent error (APE) in MD is shown for the human subject with one session acquired at isocenter on scanner A and another acquired 6cm superior from isocenter on scanner B. The top plot shows the sagittal and coronal view of the b0 from each session to demonstrate the shift within the scanner. The bottom plots show the APE for nine sagittal slices before correction, after correction using the estimated fields, and after correction using the manufacturer specifications. The error before correction is most prominent in the superior regions of the phantom as those were the furthest from isocenter during the second acquisition.**

phantom indicate that the estimated fields improve performance of the method, but the human subject results show a small advantage for using the manufacturer field directly.

Though all intra-scanner results on scanner B are compared against using the manufacturer field directly, future work should investigate the sensitivity of our proposed method and compare with other field mapping methods such as proposed by Janke et. al [181] even though these methods require that the manufacturer provide the solid harmonic coefficients. In recent work, another approach is proposed for correcting voxel-wise b-value errors. Instead of correcting for gradient nonlinearities in the coil, this method directly estimates a voxel-wise b-value map that is used to correct resulting diffusion metrics [193]. While this method could account for errors that stem from other sources of deviation than just gradient nonlinearities, the model requires an estimation of more parameters and likely it would be best practice to acquire a calibration scan along with every subject acquisition. In comparison to apply the

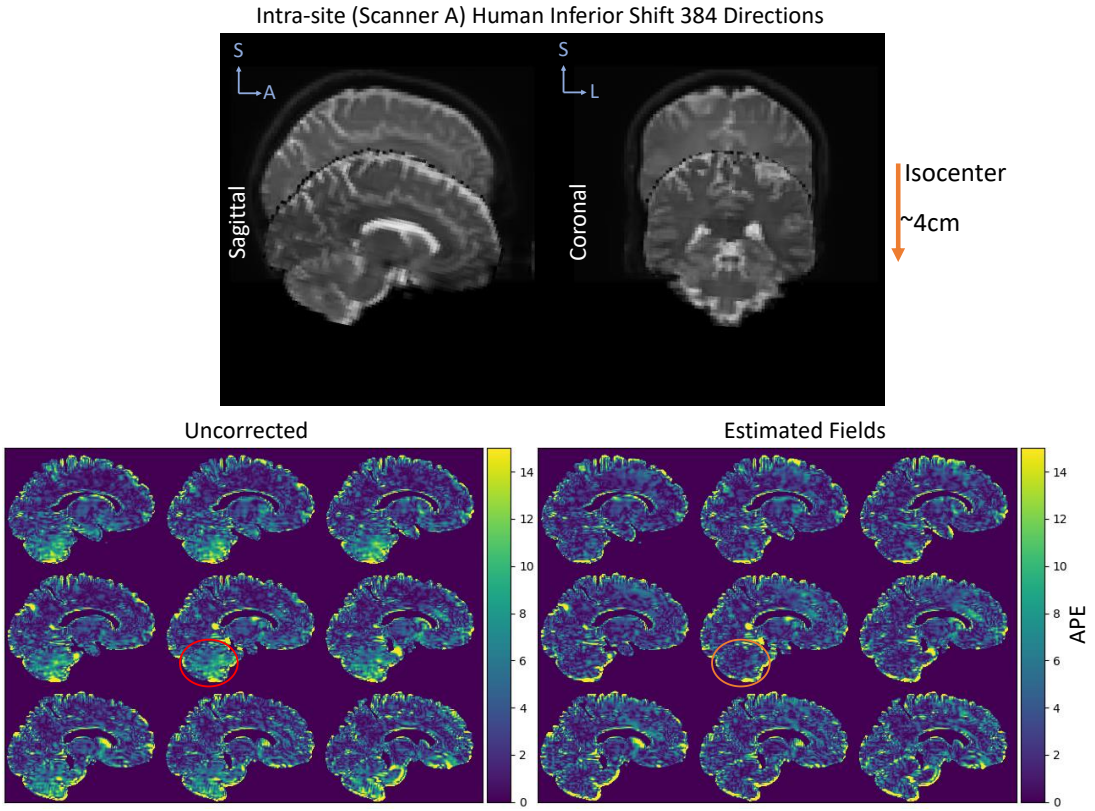


**Figure IV-8. The absolute percent error (APE) in MD is shown for the human subject with one session acquired at isocenter and another acquired 4cm inferior from isocenter on scanner A. These acquisitions were acquired with 384 directions. The top plot shows the sagittal and coronal view of the b0 from each session to demonstrate the shift within the scanner. The bottom plots show the APE for nine sagittal slices before correction, after correction using the estimated fields, and after correction using the manufacturer specifications. The error before correction is most prominent in the inferior regions of the phantom as those were the furthest from isocenter during the second acquisition.**

approach proposed in this work, only a single calibration scan is necessary for each system.

While this method is successful in circumventing the need for manufacturer specifications which are not always readily available, it should be noted that vendor-provided on-scanner gradient nonlinearity correction is preferred for translation in a clinical environment. Additionally, when working with any DICOM data coordinating world coordinate frame and patient frames can be incredibly nuanced and should be considered carefully when applying any corrections post acquisition. However, our approach remains as a solution to correct retroactively to enable the use of acquired datasets which should be corrected for gradient nonlinearity effects for use in clinic and in research.


**Figure IV-9. The mean APE within the phantom and brain excluding CSF regions are shown for each experiment without correction, after correction with the estimated fieldmaps, and after correction with the manufacturer specifications when available.**

## 6. Conclusion

This work shows that the errors caused by gradient nonlinearities is apparent in metrics derived from DW-MRI but can be reduced using the correction outlined by Bammer et al. Using empirically derived fields, we can achieve similar results without needing manufacturer specification of the hardware. In both phantom and in-vivo data, error in MD can be significantly reduced by applying this correction. We advocate for the use of gradient nonlinearity correction in standard diffusion preprocessing pipelines and provide a simple method for empirically measuring the fields necessary to account for the achieved b-values and b-vectors.

## Chapter V. The Value of Nullspace Tuning Using Partial Label Information

### 1. Introduction

Semi-supervised learning methods attempt to improve a model learned from labeled examples by using information extracted from unlabeled examples [194-196]. One effective approach is to use some form of data augmentation, in which a new example is created by transforming an unlabeled example [197, 198], and then encouraging the model to predict the same label for both.

In some learning problems, we already know that some unlabeled examples have the same label, even though that label is missing. For example, we may know that multiple photos are of the same object because of the way the photos were acquired, despite not having a label for that object. In the medical domain, repeated imaging of the same patient is common [199-201], and if the learning task is to predict something that does not change over time (or at least not over the short time between images), then we may know that the repeated images have the same label depending on the domain (Figure 1). We call this knowledge partial label information and distinguish it from the standard semi-supervised assumption that there is no label information at all for the unlabeled examples.

In prior work, Nath [8] used partial label information to predict fiber orientation distributions in magnetic resonance imaging, and Huo [202] used it for coronary artery calcium detection in non-contrast computed tomography scans. But there has been no careful investigation of how much the partial label information can add to a model's performance. In this paper, we evaluate the performance benefit of partial labels using the rigorous, standardized approach described by Oliver [203], which is designed to realistically assess the relative performance of semi-supervised learning approaches.

In this work, we use the term equivalence class to indicate a subset of unlabeled examples for which the label is known to be the same. An equivalence class need not contain all of the examples with the same label. Formally, an equivalence class $Q$ of examples $x$ in a data subset $D$ under a true but unknown labeling function $f$ is defined as:

$$Q = \{x \in D | f(x) = c\} \tag{1}$$

where $c$ is a constant. If we know that a particular pair $x_1, x_2 \in Q$, we express this as $x_1 \sim x_2$. If the labeling function $f$ is a linear operator, then the difference between any two examples in $Q$ lies in the nullspace of $f$:

$$f(x_1) = f(x_2) \Leftrightarrow f(x_1 - x_2) = 0 \tag{2}$$

Because the term nullspace is so evocative, we abuse it here to conceptually refer to comparisons between elements of the sets defined by (1), even though for a nonlinear function $f$ the relationship (2) does not hold.

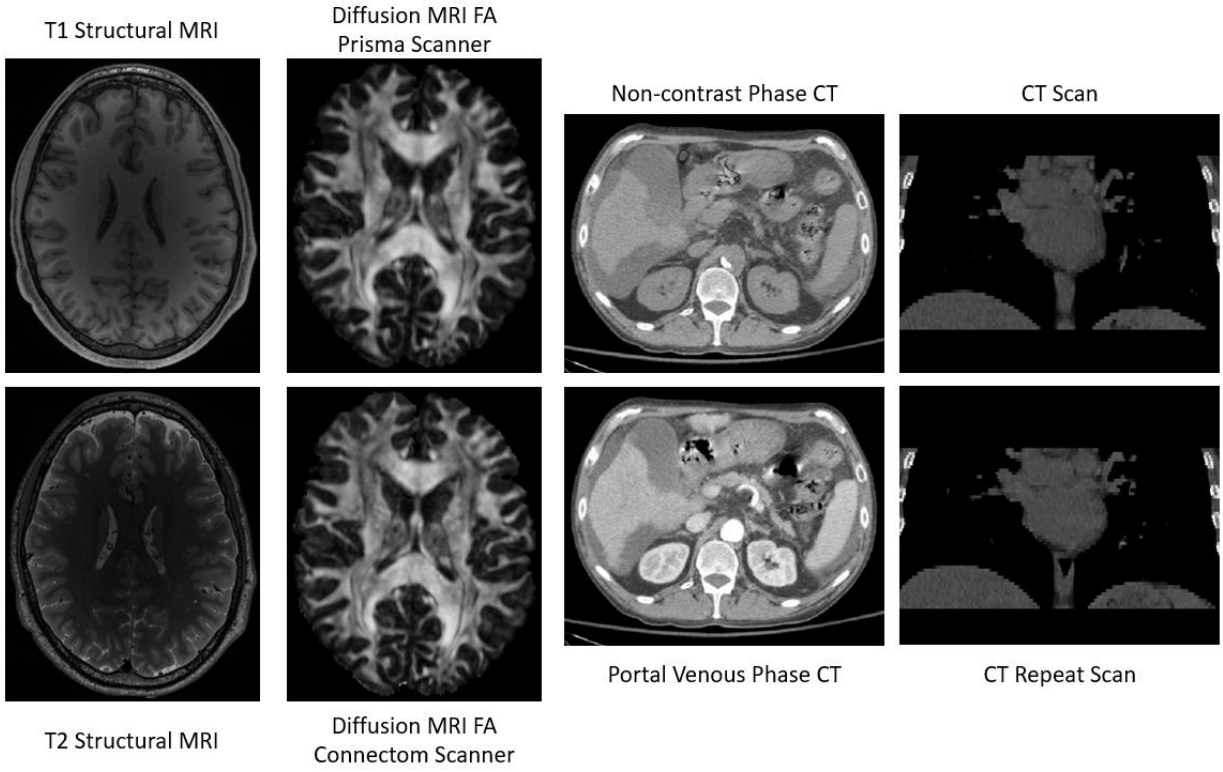


**Figure V-1. In medical imaging repeat partial label information commonly comes in the form of repeat acquisitions of a subject. Assuming these acquisitions are acquired within a reasonable amount of time such that aging does not affect the anatomy, models can leverage the differences between acquisitions that may arise from differences in acquisition parameters. This may be differences in contrast such as the difference between T1 weighted MRI (top left) and T2 MRI (bottom left) or between non-contrast phase CT (top center-right) and portal venous phase CT (bottom center-right). The manufacturer of the imaging equipment may be a factor as well as is shown in the diffusion MRI fractional anisotropy (FA) estimated from a Prisma scanner (top center-left) and the FA estimated from a Connectom scanner (bottom center-left). Using repeat acquisitions with the same parameters and hardware can also provide useful information such as in repeat heart CT (right).**

We can use our knowledge of the natural equivalence classes in a dataset to help tune a model using a procedure that we call Nullspace Tuning, in which the model is encouraged to label examples $x_1$ and $x_2$ the same when $x_1 \sim x_2$. Nullspace tuning is easily implemented by adding a term to the loss function to penalize the difference in label probabilities $h(x_1) - h(x_2)$ assigned by the current model $h$ when $x_1 \sim x_2$ (Figure 2).

Nullspace Tuning is related to the idea of Data Augmentation and is a form of contrastive learning. Data Augmentation is used when we know specific transformations that should not affect the label, and we create new

training examples by transforming existing examples and keeping the label constant, whether that label is known or unknown. This implicitly places those transformations into the architecture's nullspace. Contrastive Learning constrains a network using two samples from the same or different classes. Some semi-supervised contrastive methods create two instances from the same sample using data augmentation [204]. In contrast, Nullspace Tuning explicitly places naturally occurring but unknown transformations into the nullspace when nature provides examples of them by way of partial labels. The contribution of this work is the empirical analysis of the use of partial label information to improve model generalizability through contrastive learning. In specific cases, unlabeled data may contain partial label information in the form of equivalence classes which Nullspace Tuning proposes to use to identify non-class altering differences between two samples.

## 2. Related Work

Data augmentation, semi-supervised learning, and contrastive learning are existing learning approaches that are closely related to Nullspace Tuning, and there is a large literature for each. Thorough reviews are available elsewhere [88, 89, 205, 206]. In this section we discuss specific work that is most closely related, and work that we use as experimental baselines.

### 2.1. Data Augmentation

Data Augmentation artificially expands a training dataset by modifying examples using transformations that are believed not to affect the label. Image deformations and additive noise are common examples of such transformations [197, 207, 208]. The most effective data augmentations may be specific to the learning task or dataset and driven by domain knowledge. Elastic distortions, scale, translation, and rotation are used in the majority of top performing MNIST models [208-211]. Random cropping, mirroring, and color shifting are often used to augment natural images [74]. Recent work automatically selects effective data augmentation policies from a search space of image processing functions [197]. Though data augmentation can provide useful variation which enable more generalizability, they only approximate the natural differences within a class which can impact a model. When partial label information is available where the labels themselves are not, the contrast between equivalent samples is more informative than artificial augmentations.

### 2.2. Semi-supervised Learning

Recent approaches to Semi-supervised Learning add to the loss function a term computed over unlabeled

data that encourages the model to generalize more effectively. While there are many examples of this approach [203], we describe those here that we use for comparison in our experiments.

Π -Model encourages consistency between multiple predictions of the same example under the perturbations of data augmentation or dropout. The loss term penalizes the distance between the model's prediction of two perturbations of the same sample [90, 91].

Mean Teacher [92] builds on Π-Model by stabilizing the target for unlabeled samples. The target for unlabeled samples is generated from a teacher model using the exponential moving average of the student model's weights. This allows information to be aggregated after every step rather than after every epoch.

Virtual Adversarial Training (VAT) approximates a tiny perturbation which, if added to unlabeled sample $x$, would most significantly change the resulting prediction without altering the underlying class [94]. VAT can be used in place of or in addition to data augmentation.

Pseudo-labeling uses the prediction function to repeatedly update the class probabilities for an unlabeled sample during training [93]. Probabilities that are higher than a selected threshold are treated as targets in the loss function, but typically the unlabeled portion of the loss is regulated by another hyperparameter [93].

MixUp creates augmented data by forming linear interpolations between examples. If the two source examples have different labels, the new label is an interpolation of the two [212]. MixMatch was developed by taking key aspects of dominant semi-supervised methods and incorporating them into a single algorithm. The key steps are augmenting all examples, guessing low-entropy labels for unlabeled data, and then applying MixUp to provide more interpolated examples between labeled, unlabeled, and augmented data (using the guessed labels for unlabeled data) [198].

Berthelot et al. [198] compares these semi-supervised methods to their proposed MixMatch method. They evaluate these on the CIFAR-10 dataset [213] and on the Street View House Number (SVHN) dataset [214] as they simulate labeled and unlabeled data. They split the training set such that the models are trained at 250, 500, 1,000, 2,000, and 4,000 labeled data with the remaining treated as unlabeled data each time. In SVHN dataset, they find that MixUp generally has the worst performance reaching a 40\% test error with 250 labeled data while MixMatch has the best performance staying below 4\% test error at 250 labeled data. MixMatch also shows superior performance in the CIFAR-10 dataset achieving 11\% test error at 250 labeled data where the next best performing method VAT achieves 36\% test error. While these semi-supervised approaches can incorporate unlabeled data, they rely on artificial data

augmentation or on the model's prediction function. While these can be important sources of information, they would ignore partial label information present in unlabeled data.

### 2.3. Equivalence Classes in Labeled Data

An idea similar to Nullspace Tuning was used by Bromley [215] in fully supervised learning, where $x_1 \sim x_2$ is known because their labels are observed. They used this fact to improve a signature verification model by minimizing distance between different signatures from the same person, essentially tuning the nullspace of the network with labeled equivalence classes. Contrastive loss extends this concept to learn from the contrast of two samples whether they are from the same or different classes [216, 217]. This idea inspired triplet networks [218] that learn from tuples $(x, x^+, x^-)$, where $x \sim x^+$ and $x \nsim x^-$, and the predicted probabilities are encouraged in the loss function to be respectively near or far. There are multiple works that indicate usage of Siamese networks for person re-identification [219-221]. Nullspace Tuning extends these ideas to the case where the labels are missing but still known to be the same.

Semi-supervised contrastive learning uses data augmentation with a contrastive objective. Here $x^+$ can be generated from $x$ using some augmentation function [204]. Similarly, semi-supervised null space learning uses a positive and negative sample, but this technique uses two samples of the same object. For person re-identification, Zhang et al. relies on learning the null Foley-Sammon transform (NFST) [222] from a labeled set and then using the model's current prediction function alongside a nearest neighbors clustering to estimate groups of images which each consist of a single individual to create positive and negative samples [223]. The goal of NFST and contrastive learning is to learn an embedding that satisfies zero within-class scatter and positive between-class scatter. The main difference between semi-supervised contrastive learning and semi-supervised null space learning is the use of two real samples in the same class rather than augmentations of a sample. This work extends the idea of null space learning to deep learning classification models while focusing on the specific case of contrastive learning with partial label information.

## 3. Methods

This section describes the method of Nullspace Tuning using partial labels. We describe it first as a standalone approach, and then to illustrate how it can be combined with existing methods, we illustrate it in combination with MixMatch.

### 3.1. Nullspace Tuning

Given a set of labeled data $\{x_i, y_i\} \epsilon D$ and unlabeled data $\{x_i^*\} \epsilon D^*$ for which some equivalence classes are known, we perform Nullspace Tuning by adding to a standard loss function $\mathcal{L}_s$ a penalty on the difference in the predicted probabilities for pairs of elements of $D^*$. The new loss function $\mathcal{L}$ becomes

$$\mathcal{L} = \mathcal{L}_s(h(x_i), y_i) + \lambda \parallel h\left(x_j^*\right) - h(x_k^*) \parallel_2^2 \tag{3}$$

where $h$ is the vector-valued prediction function of the model, $\lambda$ is a hyperparameter weighting the contribution of the nullspace loss term, and $x_j^*$ and $x_k^*$ are two unlabeled samples such that $x_j^* \sim x_k^*$ is known. Note that $x_j^*$ and $x_k^*$ have no required relationship to the labeled $x_i$. By minimizing the distance between equivalent unlabeled data $x_j^*$ and $x_k^*$, we encourage the model towards zero within-class scatter similar to null space learning [223], but rely on the supervised term to ensure positive between class scatter. In the first experiment, we use cross entropy as the standard loss function component$\mathcal{L}_s$.

### 3.2. MixMatchNST

In our second experiment, we modify the MixMatch loss function with a Nullspace Tuning term and denote this model MixMatchNST. In brief, MixMatch assigns a guessed label $\bar{q}$ to each unlabeled example $x^*$ by averaging the model's predicted class distributions across $K$ augmentations of $x^*$:

$$\bar{q}_j = \frac{1}{K} \sum_{k=1}^{K} h\left(x_{j,k}^*\right) \tag{4}$$

Temperature sharpening is then applied to the probability distribution of guessed labels to lower the entropy of those predictions for each example:

$$q = \frac{\bar{q}_i^{\frac{1}{T}}}{\sum_{j=1}^{L} \bar{q}_i^{\frac{1}{T}}} \tag{5}$$

Where $T$ is a hyperparameter which is chosen to be $T = 0.05$ as per Goodfellow et al. [224]. Mixup [212] is then applied to the labeled data $\{x_i, y_i\}$ and unlabeled data $\left\{x_j^*, q_j\right\}$ to produce interpolated data $\{\tilde{x}_i, \tilde{y}_i\}$ and $\left\{\tilde{x}_j^*, \tilde{q}_j\right\}$. For a pair of two examples with their corresponding label probabilities $(\mathrm{x}_1, \mathrm{p}_1), (\mathrm{x}_2, \mathrm{p}_2)$, MixUp computes $(\tilde{x}, \tilde{p})$ as:

$$\lambda \sim \mathrm{Beta}(\alpha, \alpha) \tag{6}$$

$$\lambda' = \max(\lambda, 1 - \lambda) \tag{7}$$

$$\tilde{\mathrm{x}} = \lambda' \mathrm{x}_1 + (1 - \lambda') \mathrm{x}_2 \tag{8}$$

$$\tilde{\mathrm{p}} = \lambda' \mathrm{p}_1 + (1 - \lambda') \mathrm{p}_2 \tag{9}$$

Where $\alpha$ is a hyperparameter which is chosen to be $\alpha = 0.75$ as per Goodfellow et al. [224]. For each labeled sample and each unlabeled sample being $x_1$, another sample, labeled or unlabeled within the batch, is randomly selected as $x_2$ for MixUp which will result in $\{\tilde{x}_i, \tilde{y}_i\}$ and $\{\tilde{x}_j^*, \tilde{q}_j\}$. Weight decay is used during training to prevent overfitting [225, 226].

With the addition of Nullspace Tuning, the loss function for MixMatchNST becomes a combination of terms: the loss term $\mathcal{L}_X$ for labeled data, which in this case is the cross-entropy loss $\mathcal{L}_s$, the MixMatch loss term $\mathcal{L}_U$ for unlabeled data and guessed labels, and the Nullspace Tuning loss term $\mathcal{L}_E$:

$$\mathcal{L}_X = \mathcal{L}_s(h(\tilde{x}_i), \tilde{y}_i) \tag{10}$$

$$\mathcal{L}_U = \| h(\tilde{x}_j^*) - \tilde{q}_j \|_2^2 \tag{11}$$

$$\mathcal{L}_E = \| q_j - q_k \|_2^2 \tag{12}$$

$$\mathcal{L} = \mathcal{L}_X + \lambda_U \mathcal{L}_U + \lambda_E \mathcal{L}_E \tag{13}$$

where $\lambda_U$ and $\lambda_E$ are hyperparameters controlling the balance of terms, and $x_k^*$ is chosen so that $x_j^* \sim x_k^*$. The added Nullspace Tuning term (12) is calculated between the guessed labels $q_j$ and $q_k$ before the MixUp step, whereas the MixMatch terms (10) and (11) are calculated using interpolated, post-MixUp examples, as usual.

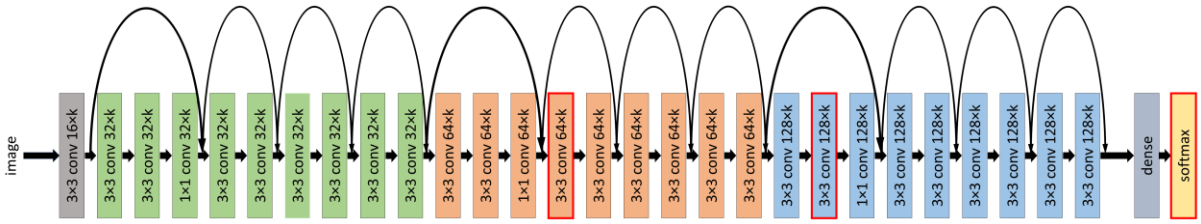


**Figure V-3. To investigate the impact of Nullspace Tuning on the feature space, we visualize extracted feature maps for both MixMatch and MixMatchNST models. The convolutional operations in the Wide ResNet-28 model are bordered in red where we choose to extract the feature maps for all test images in CIFAR-10 for each chosen model.**

### 4. Experiments

We evaluate the benefit of Nullspace Tuning over partial label information using standard benchmark datasets. We follow the precedent of simulating randomly unlabeled data in these datasets, and we likewise simulate partial labels and their equivalence classes.


**Figure V-4. Samples from CIFAR-10 (left) and SVHN (right) are shown here. CIFAR-10 contains natural images of animals and vehicles and SVHN contains natural images of house numbers where the centered number is the one of interest.**

### 4.1. Implementation Details

All experiments use a ``Wide ResNet-28" model [227], with modifications made to the loss function as needed to instantiate the various comparison methods. The training procedure and error reporting follows Oliver [203] for our experiment comparing standalone semi-supervised methods, and Berthelot [198] for our experiment comparing combined methods.

### 4.2. Standalone Methods

In this experiment the simple use of Nullspace Tuning over partial labels was compared against four good semi-supervised learning methods, using benchmark datasets CIFAR-10 [213] and SVHN [214], and the comparison framework designed by Oliver [203] re-implemented in PyTorch [228]. The comparison methods were Π-Model [90, 91], Mean Teacher [92], VAT [94], and Pseudo-Label [93], all using the Oliver framework [203].

To simulate semi-supervised data, labels were removed from the majority of training data, leaving a small portion of labeled data, the size of which was systematically varied as part of the experiment. To simulate partial label information, equivalence classes were computed on the set chosen to be unlabeled, but before the labels were removed, one equivalence class per unique label value. Performance in these experiments represents an upper bound on the benefit we can expect to achieve using Nullspace Tuning over similar data, because natural partial labels are not

always known so completely.

Test error and standard deviation was computed for labeled dataset sizes between 250 and 8000, for five randomly seeded splits each. CIFAR-10 has a total of 50000 examples of which 5000 are set aside for validation, and SVHN a total of 73,257 of which 7325 are set aside for validation. The standard test set for each dataset are used to evaluate models. Hyperparameters for this experiment were set to those used by Oliver [203].

### 4.3. Combined Methods

These experiments evaluate the benefit of adding Nullspace Tuning to an existing powerful semi-supervised learning approach. MixMatch is a good example for this demonstration, because aside from being state of the art, it uses several techniques in combination already, and therefore has a fairly complex loss function.

Unlabeled and partially labeled examples were computed as in Experiment 1. For this experiment, we evaluate only on CIFAR-10, on which MixMatch has previously achieved the largest error reduction compared to other methods [198]. We used the TensorFlow [229] MixMatch implementation, written by the original authors [198], augmenting it to produce our MixMatchNST algorithm. Test error and standard deviation was computed for labeled dataset sizes between 250 and 4000, with five random splits each.

MixMatch hyperparameters were set at the optimal CIFAR-10 settings established by Berthelot. For MixMatchNST we set the Nullspace Tuning weight $\lambda_E$, which generally works well for most experiments. However, to investigate whether the addition of the Nullspace Tuning term altered the loss landscape, we also performed univariate grid search over the MixUp hyperparameter $\alpha$ and the loss component weights $\lambda_E$ and $\lambda_U$ for MixMatchNST. We did not apply a linear rampup [92] to $\lambda_E$ as is done for $\lambda_U$.

Further characterization of MixMatchNST is accomplished through modifying the unlabeled data. With 500 labeled data in the CIFAR-10 training set, the model is trained with a varied amount of unlabeled data starting with all 44,500 and ending with 5,000 for one experiment. Another experiment uses all 44,500 unlabeled data but increases the chance of a nonequivalent pair provided to the Nullspace Tuning term.

Additionally, we evaluate MixMatchNST on the CIFAR-100 dataset with 10,000 labeled data using a wider model. In following the parameters used by Berthelot et al., we set $\lambda_U = 150$. To find the optimal $\lambda_E$, we incrementally increase the value and retrain the model.

To investigate how the network was responding to Nullspace Tuning, we visualized three layers in the Wide

ResNet-28 model (Figure 3) for both MixMatch and MixMatchNST. The feature maps for the CIFAR-10 test set were extracted after training with 500 and 2000 labeled examples and then were reshaped into a vector for each sample. These flattened feature maps were then embedded in a 2D manifold fit with UMAP [230] resulting in a single coordinate for each sample.

## 5. Results

### 5.1. Standalone Methods

The use of partial labels generally provided a performance improvement at least as large as the difference between the best and the worst semi-supervised methods, except at the smallest labeled set sizes (Figure 5). Surprisingly, the benefit of partial labels essentially maxes out at the relatively small number of 2000 labeled examples (vs. 43000 unlabeled examples) in CIFAR-10, and at less than 250 examples (vs. 65682 unlabeled examples) in SVHN, while the semi-supervised methods continue to improve with more labeled data.

We attribute the generally weaker performance of all methods on CIFAR-10 vs. SVHN (Figure 5), including the large number of labeled examples needed to approach asymptotic accuracy, to the higher complexity of the images and the greater difficulty of the task (Figure 4)

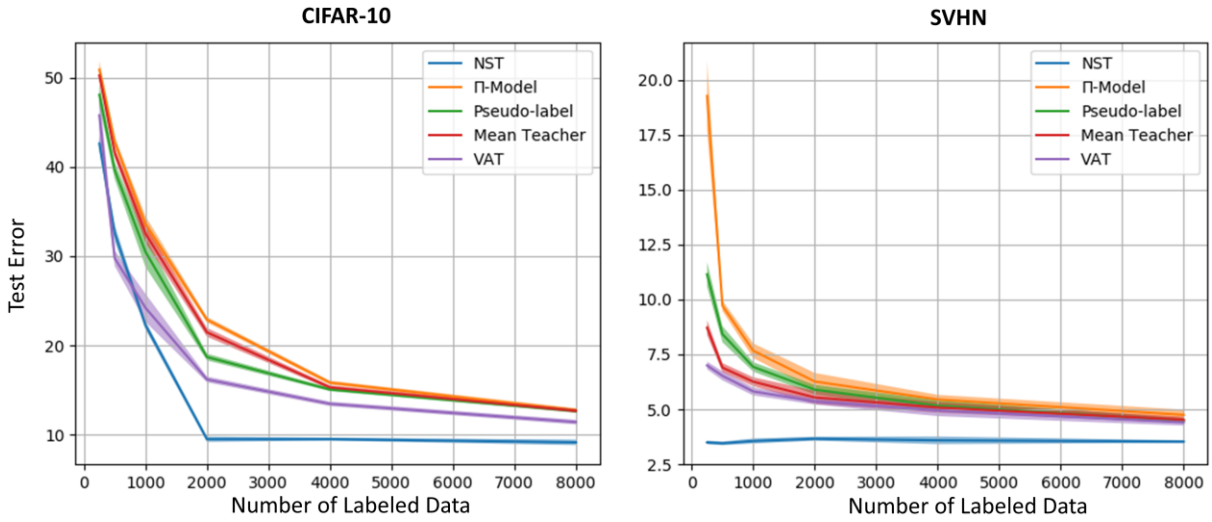


**Figure V-5. The added partial label information results in a substantial improvement in performance over baseline methods. This is shown in a percent test error and standard deviation (shaded region) comparison of Nullspace Tuning to baseline methods on CIFAR-10 (left) and SVHN (right) for a varied number of labeled data between 250 and 8000. The most significant improvement between Nullspace Tuning and the next best performing method (VAT) occurs in CIFAR-10 at 2000 labeled data.**

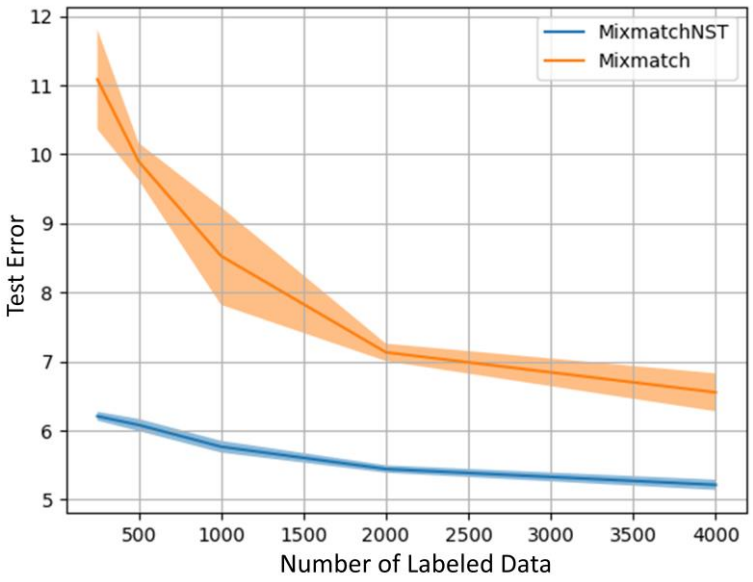


**Figure V-6. The additive performance of Nullspace Tuning on top of the state-of-the-art MixMatch algorithm is considerable. This is especially evident at 250 labeled data in CIFAR-10 where error is reduced by a factor of 1.8. This is shown in a percent test error and standard deviation (shaded regions) comparison of MixMatchNST to MixMatch on CIFAR-10 for a varying number of labels.**

**Table V-1. CIFAR-10 and SVHN percent classification error is reported here for all methods at 250 and 2000 labeled data. Bolded valuesindicate the best performing method for the number of labeled data in the dataset.**

| Method | CIFAR-10 Error 250 Labeled Data | CIFAR-10 Error 2000 Labeled Data | SVHN Error 250 Labeled Data | SVHN Error 2000 Labeled Data |
|---|---|---|---|---|
| Π-Model | 50.88 ±0.94 | 22.88 ±0.30 | 19.28 ±1.58 | 6.27 ±0.39 |
| Mean Teacher | 50.22 ±0.39 | 21.46 ±0.49 | 8.72 ±0.33 | 5.54 ±0.30 |
| Pseudo-Label | 48.06 ±1.24 | 18.70 ±0.38 | 11.15 ±0.55 | 5.89 ±0.22 |
| VAT | 45.76 ±2.81 | 16.19 ±0.32 | 7.00 ±0.17 | 5.36 ±0.14 |
| NST | 42.60 ±0.82 | 9.50 ±0.30 | **3.49 ±0.08** | **3.66 ±0.10** |
| MixMatch | 11.08 ±0.72 | 7.13 ±0.13 | NA | NA |
| MixMatchNST | **6.21 ±0.06** | **5.44 ±0.05** | NA | NA |

### 5.2. Combined Methods

The performance of MixMatch on CIFAR-10 was better than any algorithm alone, including Nullspace Tuning, in the first experiment (Table 1). Despite this impressive gain, performance was improved further by including

Nullspace Tuning together with the MixMatch innovations. Doing so reduced test error by an additional factor of 1.8 on the smallest labeled set size, and about 1.3 at the largest set size (Figure 6).

Adding Nullspace Tuning to MixMatch with even a small number of labeled examples dramatically improved the performance on CIFAR-10 (Figure 6), suggesting complementary and synergistic use of information in the two methods; either method on its own needed over 2000 labeled examples to approach its asymptotic accuracy.

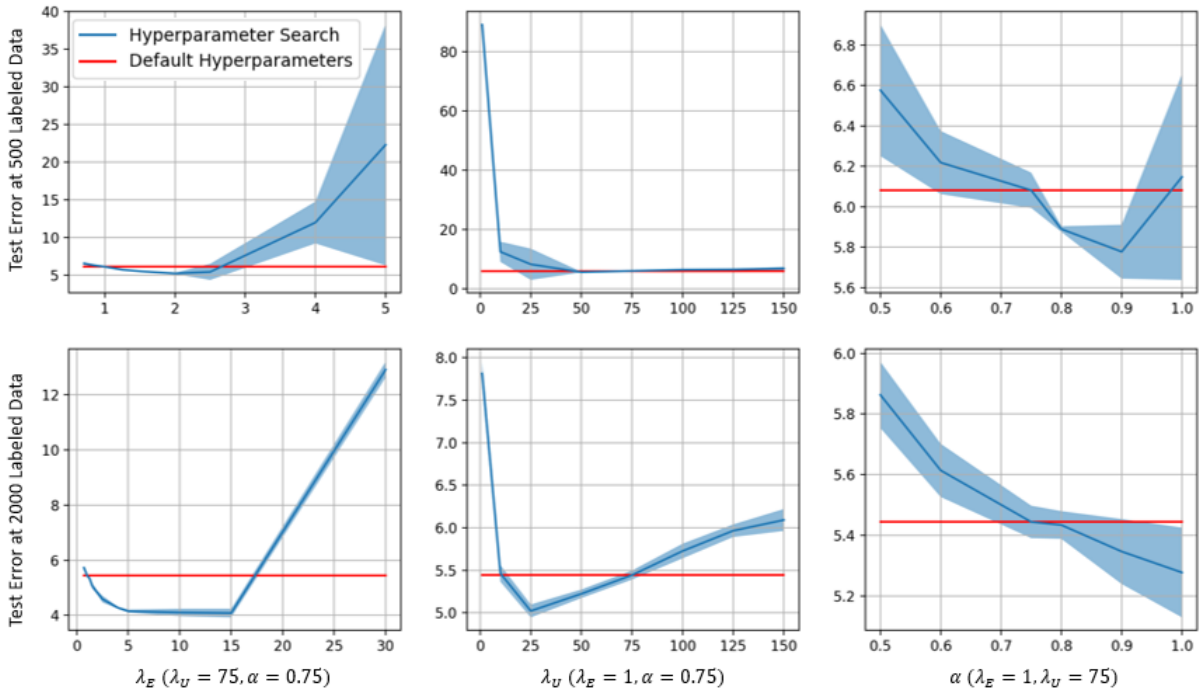


**Figure V-7. MixMatchNST models can benefit from hyperparameter tuning at each number of labeled data. The hyperparameter $\lambda_U$ shows the greatest need for this as the optimal value at 2000 labeled data is at least double of that at 500 labeled data and would reduce the test error by approximately 1.4%. Test errors and standard deviations are reported (shaded regions) at 500 labeled data (top) and 2000 labeled data (bottom) as the hyperparameter space is searched for $\lambda_E$ (left), $\lambda_U$ (center), and $\alpha$ (right). Red lines indicate the performance before fine tuning. As one hyperparameter is tuned, the other two are set to the previously used values. The error achieved with the hyperparameters in Figure 4 is indicated by the red line.**

The hyperparameter search shows that the MixMatch loss landscape was modestly altered with respect to the MixMatch hyperparameters $\alpha$ and $\lambda_U$, and small gains could be had by tuning them further (Figure 7). Tuning our nullspace weight made a larger relative difference, providing a further improvement of about 20% at 500 labeled datapoints, and about 30% at 2,000 labeled datapoints, over what is shown in Figure 6.

The robustness of MixMatchNST is evaluated as we altered the amount of unlabeled data and retrained the model. In reducing the amount of unlabeled data, we found that MixMatchNST can outperform the MixMatch model

trained with all 44,500 unlabeled data when only 20,000 unlabeled data are available. We also found that MixMatchNST can outperform MixMatch when there is a 50% chance that the partial label information which provides an equivalence class pair is incorrect (Figure 8).

MixMatchNST also sees a large increase in performance over MixMatch when evaluated on a more difficult task and using a larger model. MixMatchNST reduces test error in the CIFAR-100 dataset by more than 4% with $\lambda_E = 50$ (Figure 9).

Image features from MixMatch models and MixMatchNST models show that comparable learning happens with fewer examples with the addition of Nullspace Tuning (Figure 10), and that this learning occurs deep inside the model, rather than superficially at a later layer. The clusters in convolutional layers under Nullspace Tuning with 500 labeled examples look comparable to those for 2,000 labeled examples without it, and the clustering appears slightly clearer with Nullspace Tuning given the same number of labeled examples. Differences in the softmax layer are subtler, but their presence is evident by the overall model performance.

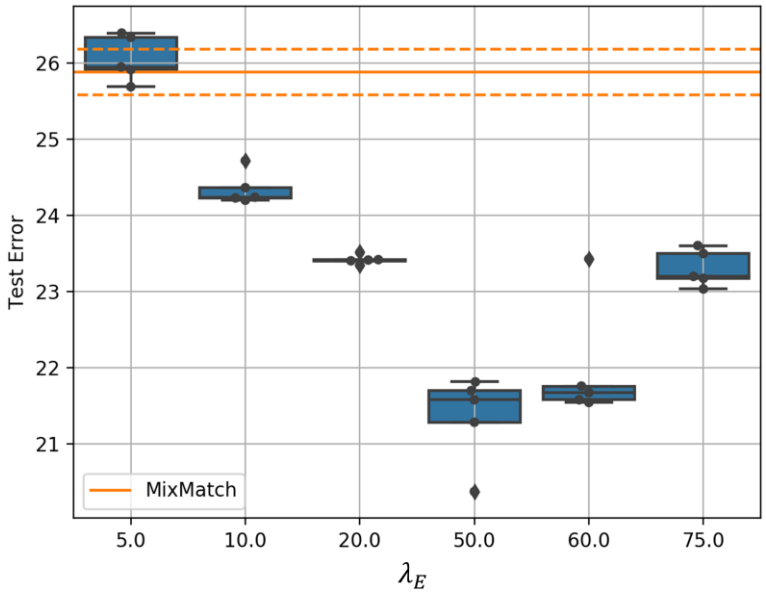


**Figure V-8. Using a wider network, we evaluate MixMatchNST on the CIFAR-100 dataset as we set $\lambda_U = 150$ and $\alpha = 0.75$ as we increase $\lambda_E$. A much larger $\lambda_E$ is needed as compared to the smaller model in the CIFAR-10 dataset.**

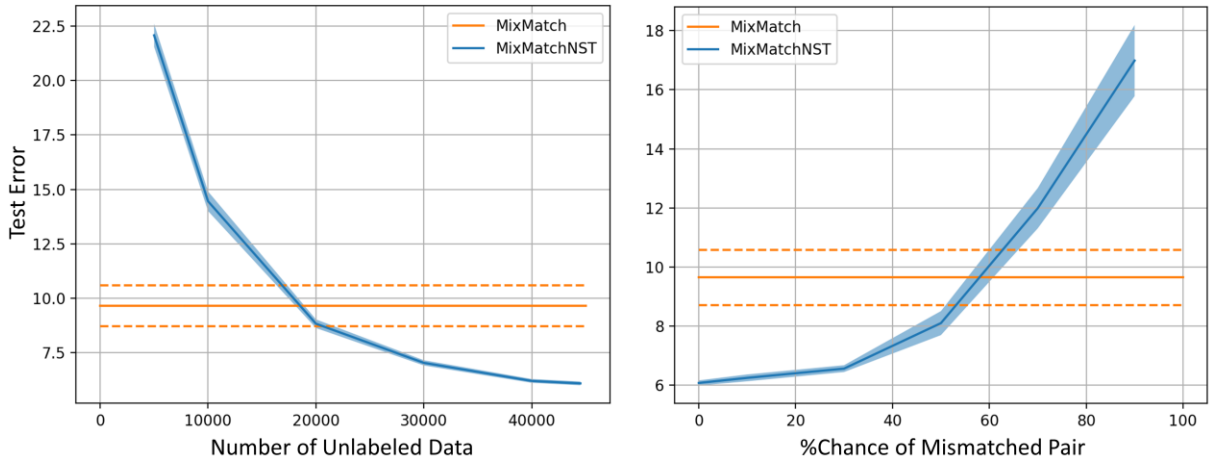


**Figure V-9. MixMatchNST as we alter the unlabeled data is compared to the baseline MixMatch as reported by Berthelot et al. with 500 labeled data all unlabeled data. We choose to alter the amount of unlabeled data (left) and to simulate error in the chosen equivalency classes (right). If we take away approximately half of the unlabeled data or if we have a 50% chance of incorrectly choosing an equivalent pair, MixMatchNST still outperforms MixMatch with all unlabeled data.**

## 6. Discussion

The main contribution of this work is the systematic demonstration that tuning the nullspace of a model using the partial label information that may reside in unlabeled data can provide a substantial performance boost compared to treating them as purely unlabeled data. It is not surprising that adding new information to a model provides such an improvement; our goal with this work was to quantify just how much improvement one could expect if equivalence classes were known within the unlabeled data. This idea is important because identifying or obtaining equivalence classes within unlabeled data may be cheaper than obtaining more labels, if standard semi-supervised methods provide insufficient performance.

The gain from using partial label information is fairly constant over the range of labeled dataset size tested, as long as a minimum threshold of labeled data is met. This makes sense from the perspective of tuning the null space of the model, because most of that tuning can be done with equivalence classes, but a small amount of labeled data is needed to anchor what is learned to the correct labels.

Increasing the number of labeled examples beyond the threshold is essentially trading partial label information for full label information. The relative value of that information for a given learning problem is suggested by the slope of the error curve. For the standalone methods comparison, the nearly horizontal slope suggests that

partial information is nearly as good as full label information. The performance of the architecture on a fully-labeled dataset was 2.59% error, which reinforces this idea. The steeper (but still mild) slope found in the combined methods comparison suggests a stronger tradeoff, although performance of MixMatch on the full 40,000 examples is 4.2% [198], which is fairly close to the 5.5% that we get using more than 90% partial labels, or even the 6.0% that we get with 99% partial labels. We conclude that at least in some cases, partial labels can get us most of the way there.

We can infer something about what the models are learning from the ordering of model performance: MixMatchNST > MixMatch > Nullspace Tuning > single data-augmentation models. Explicitly learning the shape of the nullspace from partial labels was much more effective than implicitly placing data transformations into that space by the standalone algorithms, although combining those transformations into MixMatch was more effective still. But the fact that MixMatchNST performed better than either MixMatch or Nullspace Tuning alone demonstrates that MixMatch is learning somewhat different aspects of the nullspace than that provided by the partial labels.

Decreasing the amount of unlabeled data has a nonlinear effect on the performance of the MixMatchNST method, but when the amount unlabeled data is decreased by 55\%, the partial label information is able to compensate achieving better performance than a model without partial label information with all the unlabeled data. The nullspace tuning term in the MixMatchNST method is directly shown to be resilient to noise in the equivalency classes showing improvement even while 50\% of the pairs provided are not equivalent. This would suggest that even less than perfect partial labelling methods may still adequately tune the model.

One strength of this method is its simplicity – it can be added to nearly any other semi-supervised learning algorithm, as long as we have access to the loss function, and we can provide appropriate example pairs from an equivalence class.

This experiment used the largest possible equivalence classes – one class for each label value. Naturally occurring equivalence classes are likely not to be so large, especially if they are obtained by repeated observations of the same object. Our experimental design investigated the most we could gain from using the partial information in equivalence classes, but if the classes are smaller and more numerous, then we might expect that gain to be smaller. But because the partial labels are given to the algorithm as example pairs, with no required relationship between those pairs and the labeled pairs, Nullspace Tuning can still be used even with equivalence classes as small as two examples. And if those equivalence classes are well distributed over the data space, their diminished size may not actually impact the benefit by much. One could imagine that even a relatively small number of relatively small equivalence classes

could be rather effective at tuning the null space. The large number of trained models needed to characterize how the benefit changes with respect to the size and number of equivalence classes placed that question out of scope for this paper, but it will be an interesting direction for future work.

And of course, not all learning problems have natural equivalence classes embedded in them at all. Benchmark public datasets tend not to, except in simulations like ours, partly because information about how they were collected has been lost. But it may be cheap to instrument data collection pipelines to record information that does provide this information. In addition to the medical use cases described above, where the patient identity is tracked through repeated observations, unlabeled objects may be tracked through sequential video frames, fixed but unlabeled regions may be identified for multiple passes of a satellite, or the unlabeled sentiment of all sentences in a paragraph might be considered to form an equivalence class. We expect that there are many creative ways to find partial labels in naturally occurring datasets, and when we find them, Nullspace Tuning is a promising method to exploit them. Nullspace tuning is a flexible approach that is amenable to real world learning scenarios and promises to enable use of partial label information that is not accessible with current standard neural network approaches.

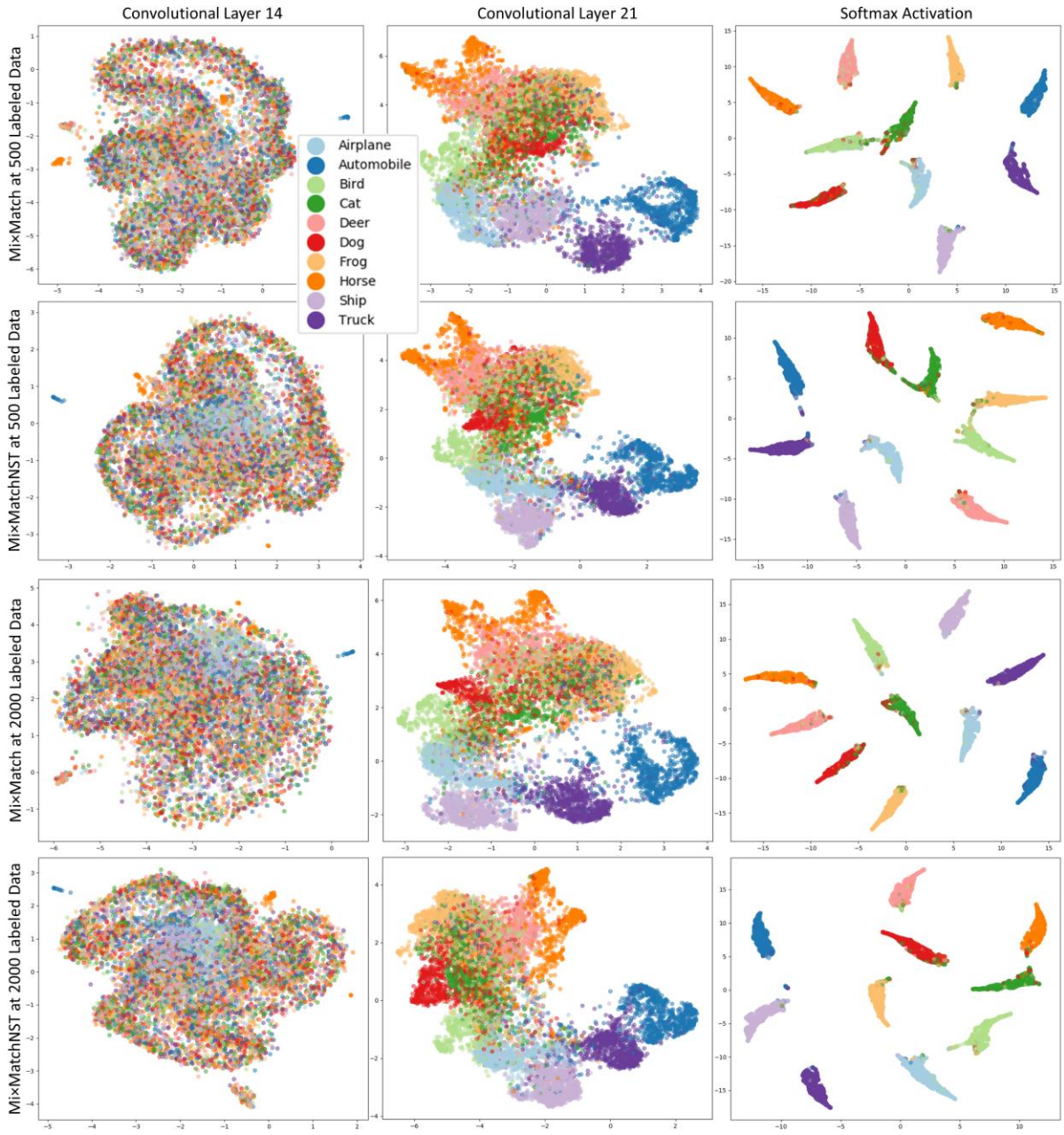


**Figure V-10. Nullspace Tuning provides better learning with fewer labeled examples, as evidenced by discernably clearer clusters in a 2Dmanifold space learned from the feature maps. In general, MixMatchNST does about as well in the CIFAR-10 test set with 500 labeled raining points (second row) as MixMatch does with 2000 (third row). Each point represents a sample's feature maps flattened to a single vector from two convolutional layers (first and second column) and the final softmax layer (third column), embedded in a 2D space learned with UMAP (McInnes et al., 2018). These are shown for a single fold for MixMatch and MixMatchNST for datasets with 500(top two rows) and 2000 (bottom two rows) labeled examples. With 500 labeled examples, a cluster is forming at layer 14 for the class "Airplane" in MixMatchNST, with no clear counterpart in MixMatch. At layer 21, several clusters are slightly clearer, with separation between Cat and Dog further along. With 2000 labeled examples, both methods are starting to form clusters for Airplane at layer 14, but MixMatchNST now also has a cluster formed for "Ship". At layer 21, several clusters are again slightly clearer for MixMatchNST, with separation especially evident for "Frog".**

## Chapter VI. Semi-supervised Machine Learning with MixMatch and Equivalence Classes

### 1. Introduction

Semi-supervised learning methods seek to leverage performance in models using information extracted from both labeled and unlabeled data [231]. Many forms of semi-supervised learning and regularization rely on data augmentation as well as the stochasticity of deep learning models. In data augmentation, a sample is transformed to introduce new example variations to which a model should be robust without altering the label of the sample. An effective semi-supervised approach is to encourage models to make the same prediction for two different variants of the same sample [197, 198]. Recent success in the CIFAR-10 classification task with limited labeled training data has been achieved through applying Mixup [212] to both labeled and unlabeled data in an algorithm called MixMatch [198]. However, the variations introduced by data augmentation are typically dataset specific. This is especially true for medical imaging tasks in which data augmentation must not alter the image outside of what is possible, considering the anatomy involved and the type of acquisition.

In some tasks, pairs or groups of unlabeled examples can be identified as having the same label even if the label itself is unknown. This is an advantage in medical imaging as many studies typically have repeat acquisitions of the same subject. Assuming the time between acquisitions is not large enough that the anatomy or diagnosis should change, then we know these same subject acquisitions have the same label. We call this knowledge *partial label information*. In prior work, partial label information has been used to predict fiber orientation distributions in diffusion weighted magnetic resonance imaging [8] and to detect coronary calcium in non-contrast computer tomography (CT) [202].

We use the term *equivalence class* to indicate a subset of unlabeled examples for which the label is known to be the same. Formally, an equivalence class $Q$ of examples $x$ in a data subset $D$ under a true but unknown labeling function $f$ is defined as:

$$Q = \{x \in D | f(x) = c\} \tag{2}$$

where c is a constant. We use the expression $x_1 \sim x_2$ to indicate a pair of samples such that $x_1, x_2 \in Q$. If the labeling function $f$ is a linear function, the difference between a pair of examples $x_1 \sim x_2$ from $Q$ lies in the nullspace of $f$:

$$f(x_1) = f(x_2) \Leftrightarrow f(x_1 - x_2) = 0 \tag{2}$$

We abuse the term *nullspace* by using it to conceptually refer to comparisons between elements in an equivalence class, even though (2) does not hold for nonlinear functions. Using the equivalence classes that can be found naturally in medical imaging, we can help tune a model by encouraging it to make the same predictions for $x_1$ and $x_2$ when $x_1 \sim x_2$ in a process we call *Nullspace Tuning*.

The purpose of this work is to show the effectiveness of recent methods MixMatch and Nullspace Tuning in medical imaging tasks and characterize their performance with diminishing labeled data. Additionally, we explore how these methods can be used in tandem to leverage aspects from both methods in training models. We do this for natural images in the task of skin lesion diagnosis using the HAM10000 skin lesion dataset [232] and for CT in the task of lung cancer diagnosis using data from the National Lung Screening Trial (NLST) with follow up confirmed diagnoses [233].

## 2. Related Work

### 2.1. Data Augmentation

Data augmentation artificially expands a training dataset by modifying examples using transformations that are believed not to affect the label. Image deformation and additive noise are common examples of such transformations [197, 207, 208]. Natural images can be effectively augmented using random cropping, mirroring, and color shifting [74]. In CT, data augmentation can consist of spatial deformations, translations, rotations, and non-rigid deformations [234]. Effective data augmentation policies can be automatically selected from a search space of image transformations [197]. Generative adversarial networks are also being used to generate anatomically informed data augmentations as well as completely new data to supplement training [235, 236].

### 2.2. Equivalence Classes in Labeled Data

Some tasks exist in which the equivalence classes describe the label completely. Signature verification and facial recognition are two examples. The verification model tunes the nullspace through minimizing the distance between different signatures or images of the same person [215, 237]. Contrastive loss extends this concept to learn from the contrast of two samples whether they are from the same or different classes [216, 217]. Triplet networks [218] use a similar concept to learn from tuples $(x, x^+, x^-)$, where $x \sim x^+$ and $x \nsim x^-$, and the predicted class probability pairs are encouraged to be near or far, respectively.

### 2.3. Semi-supervised Learning

Recent semi-supervised learning methods constrain the model through an additional term in the loss function that is computed over unlabeled data. The goal of these methods is to extract useful features from unlabeled data that will allow the model to generalize more effectively to unseen data. This can be done by penalizing the distance in predictions for two perturbations of the same sample [90, 91], by stabilizing the target for unlabeled data through obtaining predictions from a moving average of model weights during training [92], or by using the prediction function to update a guessed label for the unlabeled data periodically during training [93]. Virtual Adversarial Training approximates a small perturbation which, if added to $x$, would most significantly change the resulting prediction without altering the underlying class [94]. Of particular interest is the method called MixMatch which was developed by taking key aspects of dominant semi-supervised methods and incorporating them in to a single algorithm [198]. The key steps are augmenting all examples, guessing low-entropy labels for unlabeled data, and then applying MixUp to provide more interpolated examples between labeled, unlabeled, and augmented data [212].

## 3. Methods

Nullspace Tuning is a form of contrastive learning, but unlike some semi-supervised contrastive methods [238], Nullspace Tuning does not rely on data augmentation. Rather it relies on the natural augmentations that exist between samples that can be identified as being equivalent in class. This section describes the use of partial labels in Nullspace Tuning. First, it is described as a standalone method. Second, we illustrate how to combine Nullspace Tuning with MixMatch.

### 3.1. Nullspace Tuning

To perform Nullspace Tuning, we add a penalty on the distance between predicted probabilities for known equivalence class pairs to a standard loss function $\mathcal{L}_s$. If we have labeled data $\{x_i, y_i\} \epsilon\, D$ and unlabeled data $\{x_i^*\} \epsilon\, D^*$, the new loss function can be defined using the model's vector-valued prediction function $h$ and a known equivalence class paring $x_j^* \sim x_k^*$ as:

$$\mathcal{L} = \mathcal{L}_s(h(x_i), y_i) + \lambda \parallel h\left(x_j^*\right) - h(x_k^*) \parallel_2^2 \tag{3}$$

where $\lambda$ is a hyperparameter weighting the contribution of the nullspace loss term. It is not necessary to make any assumptions about the relationship between the labeled data $x_i$ and the unlabeled data $x_j^*$ and $x_k^*$. Additionally, in cases where the equivalency class has more than two elements, the randomization of chosen pairs within the

equivalency class can provide further data augmentation. We choose cross entropy as the standard loss function $\mathcal{L}_s$ in all experiments contained in this work.

### 3.2. MixMatchNST

The original MixMatch algorithm uses two forms of data augmentation. The first is a set of dataset-specific transformations. By averaging the predicted class distribution function across $K$ augmentations, a guessed label distribution $q$ is assigned to each unlabeled sample $x^*$. To reduce entropy, temperature sharpening is applied to $q$ [224]. The second form of data augmentation applies MixUp [212] to the labeled data $\{x_i, y_i\}$ and the unlabeled data $\{x_j^*, q_j\}$ to produce interpolated data $\{\tilde{x}_i, \tilde{y}_i\}$ and $\{\tilde{x}_j^*, \tilde{q}_j\}$. A hyperparameter $\alpha$ controls how much the examples are altered during MixUp. To prevent overfitting, weight decay is applied using an exponential moving average during training [225, 226].

MixMatchNST modifies the MixMatch loss function with the addition of a Nullspace-Tuning term. The loss function then becomes a combination of the standard loss $\mathcal{L}_s$ calculated using labeled data, the unlabeled loss term weighted by hyperparameter $\lambda_U$, and the Nullspace Tuning term weighted by hyperparameter $\lambda_E$:

$$\mathcal{L} = \mathcal{L}_s(h(\tilde{x}_i), \tilde{y}_i) \; + \; \lambda_U \parallel h\left(\tilde{x}_j^*\right) - \tilde{q}_j \parallel_2^2 \; + \; \lambda_E \parallel q_j - q_k \parallel_2^2 \tag{4}$$

where $x_j^*$ is chosen such that $x_j^* \sim x_k^*$. The Nullspace Tuning term is calculated using the guessed labels $q_j$ and $q_k$ before the MixUp step, whereas the labeled and unlabeled MixMatch terms are calculated using MixUp interpolated examples.

## 4. Experiments

We evaluate the benefit of Nullspace Tuning over partial label information as well as the benefit of MixMatch over unlabeled data in two medical imaging examples. The first is skin lesion diagnosis in natural images, and the second is lung cancer diagnosis in CT. We follow the precedent of simulating randomly unlabeled data in these datasets to characterize these methods as the amount of labeled data diminishes while the amount of unlabeled data increases [203].

**4.1.** Implementation details

All experiments were implemented in PyTorch 1.0.0 [228] and trained on Nvidia 2080Ti GPUs. In both datasets there is a class imbalance which must be considered in both the labeled and in unlabeled data. For the supervised loss, we sample evenly from each class in the labeled data. For the semi-supervised loss, the average prediction for each equivalence class is used as a guessed label, and the unlabeled or paired data are sampled evenly across the guessed labels. Additionally, for each fold, a balanced validation set is created to evaluate the model during training. The class imbalance is kept in proportion when splitting the data in to test sets for each fold, so we report balanced multi-class accuracy and AUC in our evaluation for diminishing amount of labeled. For each method, we perform a hyperparameter search on the $\lambda$ loss hyperparameters.

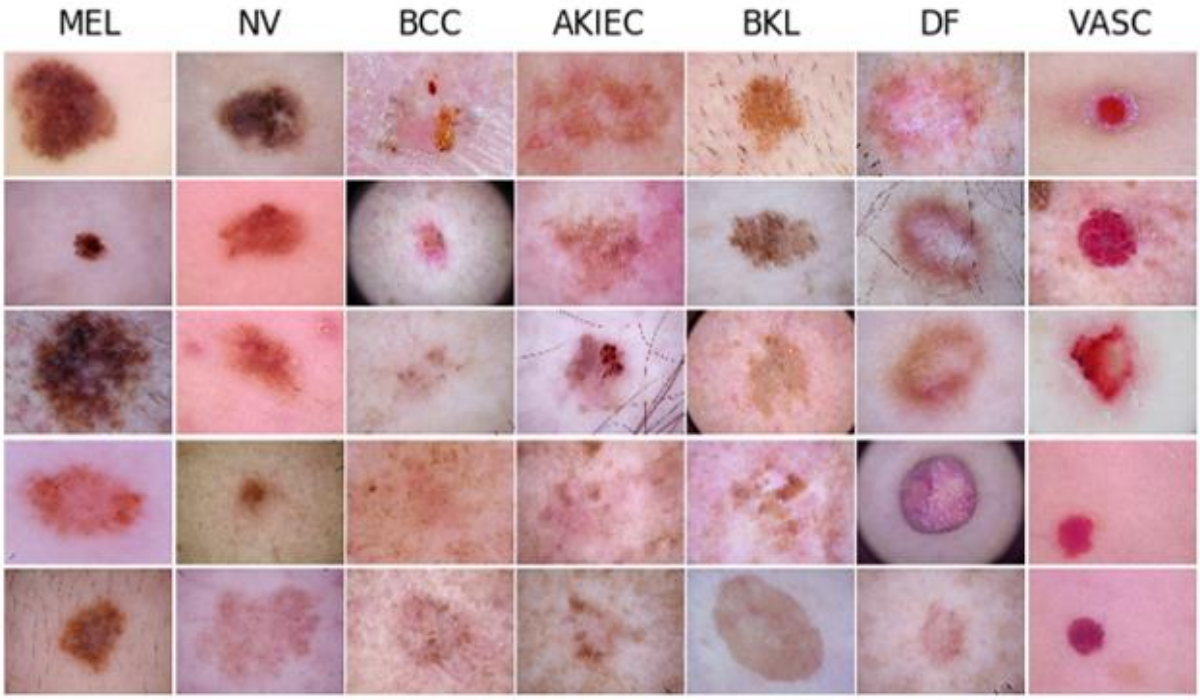


**Figure VI-1. The difficulty in the skin lesion diagnosis task is the similarity between classes and the variation within classes. This can be seen as especially true for melanoma.**

**Experiment 1:** Using the HAM10000 skin lesion dataset, we train supervised, Nullspace Tuning, MixMatch, and MixMatchNST models, using varying numbers of labeled examples. The supervised model ignores unlabeled examples. The dataset consists of 10,015 color photographs (RGB format, 600×450 pixels) of skin lesions categorized as: melanoma (MEL) (1113 images); melanocytic nevus (NV) (6705 images); basal cell carcinoma (BCC) (514

images); actinic keratosis and intraepithelial carcinoma (AKIEC) (327 images); benign keratosis, solar lentigo, and lichen-planus (BKL) (1099 images); dermatofibroma (DF) (115 images); or vascular lesions (VASC) (142 images) [232]. Fig. 1 shows examples of each class. For the network architecture, we use a DenseNet [239] which was the top performing single model in the ISIC 2018 challenge which did not use external data [240]. The method is defined by Li et al. and serves as our baseline. The weights of this model are initialized from a model pretrained on Imagenet [74]. Unlike the lung cancer data, HAM10000 does not have natural equivalence classes. We simulate these by randomly pairing the unlabeled data once at the beginning of training such that there are many unchanging equivalence classes of size two, where each example in the pair has the same known but withheld label. Random data affine transforms, mirroring, and color shifting is applied as data augmentation strategies. Validation is performed using k-fold cross validation (k=5).

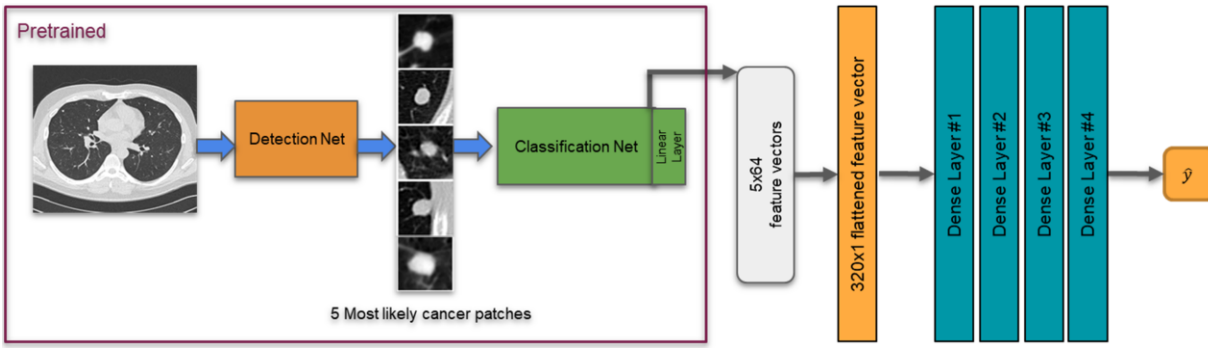


**Figure VI-2. The feature vectors extracted from the top five most likely cancer patches from the Liao pretrained model are used to train a four-layer FCNN with approximately 300,000 total parameters.**

**Experiment 2:** Here we use the NLST as well as a pretrained model from the top performing method in the 2017 Data Science Bowl lung cancer diagnosis challenge [241]. The pretrained model is defined by Liao et al. and was pretrained on a dataset provided by the National Cancer Institute which included some of the NLST data. From the NLST, data used consists of 5710 subjects and a total of 16,053 CT scans with follow-up confirmed diagnoses that successfully passed the preprocessing of Liao et al. There are 1055 subjects with a positive final diagnosis and 4655 with a negative final diagnosis. Most subjects have multiple longitudinal scans which are used as natural equivalence classes for Nullspace Tuning when a subject is simulated as unlabeled data. When splitting the data into training, validation, and testing sets as well as into labeled and unlabeled data, we keep subject data together to avoid bias in the model. We obtain the feature vectors of the five most likely nodule patches just before the final fully

connected layer in the pretrained model and train a fully connected neural network on the NLST data as described in Fig. 2. This is similar to the method used by Gao et al. [242]. For data augmentation, a small amount of random Gaussian noise is applied to the feature vectors obtained from the pretrained model. Validation is done by repeating 100 rounds of training and testing under 80/20 random splits. The training data is further split into sets of labeled, unlabeled, and validation data.

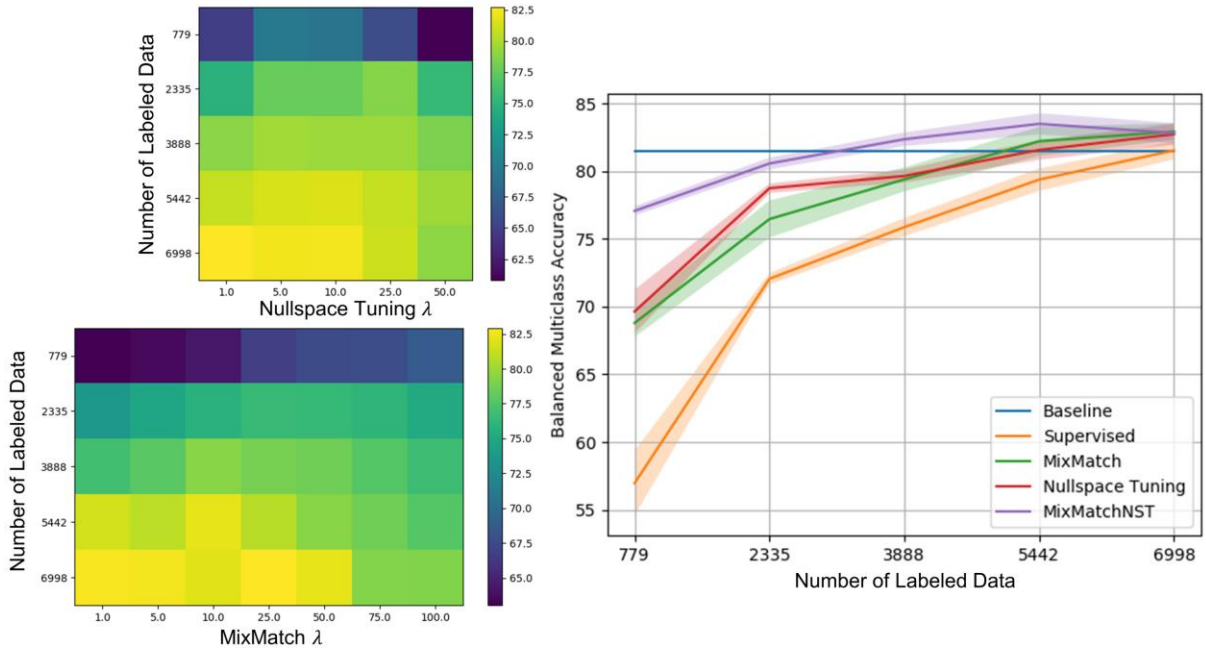


**Figure VI-3. For experiment 1 using HAM10000, the mean balanced multiclass accuracy across five folds is shown for the hyperparameter search for MixMatch (bottom left) and Nullspace Tuning (top left). The highest performing hyperparameter is used in reporting the final performance (right) where the baseline is the balanced multiclass accuracy reported by the ISIC 2018 challenge for the Li method. The shaded region represents the standard error of the mean.**

### 4.2. Results

**Experiment 1:** For the skin lesion data, balanced multiclass accuracy is reported for models trained using from 779 to 6998 labeled examples (Fig. 3). Both MixMatch and Nullspace tuning show large performance gains over the standard supervised model. When only 779 samples are labeled in the training set, both methods achieve an increase in balance multiclass accuracy of over 20%. At the same point, MixMatch achieves an increase of approximately 7% over the next best method and achieves the best performance at all amounts of labeled data (Fig. 3). At 3888 labeled data or approximately 40% of the original challenge training set, MixMatchNST achieves comparable performance to the that achieved by the Li et al. in the withheld challenge test set, and comparable

performance to using all 6998 labeled examples in a supervised model. For both methods, a larger $\boldsymbol{\lambda}$ which controls the contribution of the loss term generally achieves better performance when less data is available (Fig. 3).

**Experiment 2:** The AUC is reported after training each model with between 40 and 400 labeled subjects (Fig. 4). Here, the baseline represents the AUC from applying the Liao et al. pretrained model. All other methods train a small fully connected network using pretrained feature vectors as described by Fig. 2. Other than the baseline, Nullspace Tuning and MixMatchNST achieve the highest AUC at 200 and 400 labeled data whereas MixMatch achieves nearly the same AUC as the standard supervised approach. In general, a $\boldsymbol{\lambda}$ of 5 achieves the best Nullspace Tuning performance and a $\lambda$ of 0.1 achieves the best MixMatch performance.

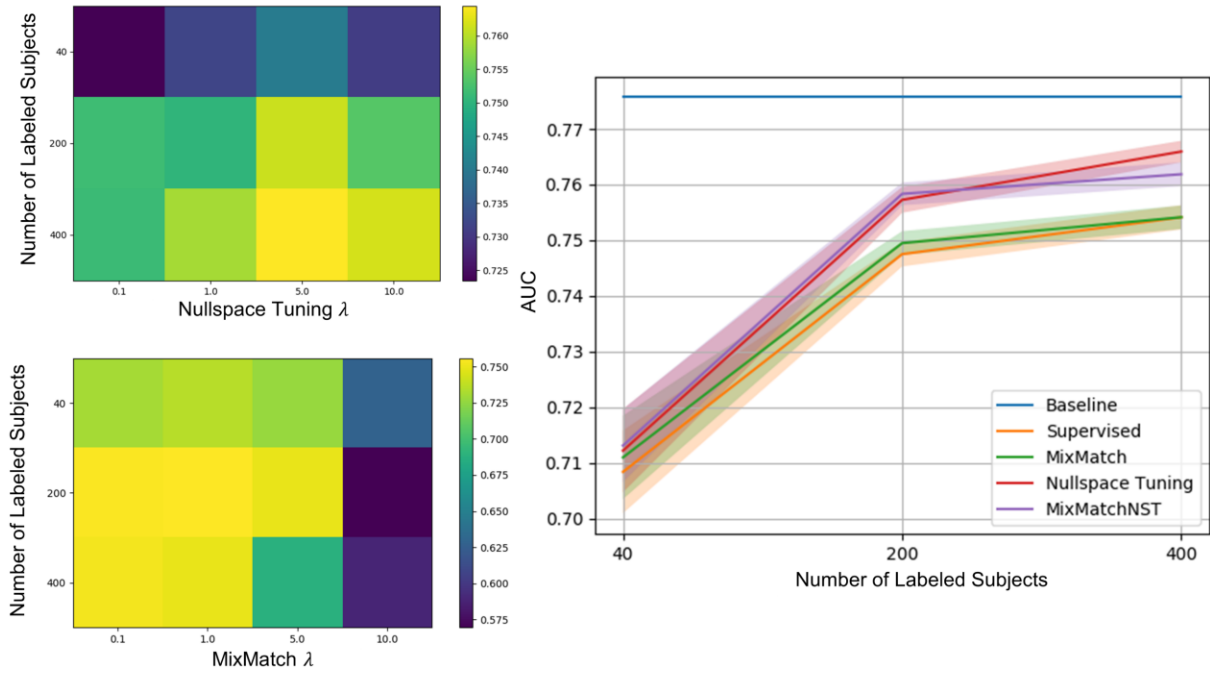


**Figure VI-4. For experiment 2 using the NLST, the mean AUC across five folds is shown for the hyperparameter search for MixMatch (bottom left) and Nullspace Tuning (top left). The highest performing hyperparameter is used in reporting the final performance (right) where the baseline is the AUC reported from directly applying the Liao model to the NLST dataset. The shaded region represents the standard error of the mean.**

## 5. Discussion

Experiment 1 depicts the full extent of the semi-supervised methods' ability to regularize the model. Even though MixMatch and Nullspace Tuning appear to have similar performance, the high performance of MixMatchNST suggests that features extracted or constrained by each method is additive to the generalizability of the model. In experiment 2, we see that even in fine tuning a pretrained model, the scarcity in labeled data has a large impact on the performance of the model. Here, the semi-supervised learning methods have a small but distinct advantage when labeled data is limited. It is possible the MixMatch algorithm is at a disadvantage when data augmentation is limited to the addition of noise rather than a full suite of randomized transforms. Additionally, the choice of using longitudinal scans as equivalence classes introduces noise due to only fine-tuning the diagnosis model without training the detection model at all. Two sets of patches each from different scans of the same subject then may not belong to the same class. While this work does not show this method is clinically applicable, it does show the added value of these semi-supervised methods in medical imaging tasks.

**Conclusion**: The use of semi-supervised learning methods such as MixMatch can greatly benefit tasks in which labeled data is scarce or annotations are expensive to obtain. We advocate for the adoption of these methods to medical image processing especially when domain specific data augmentations are available. Additionally, the ability to acquire partial label information such as equivalence classes should be considered when full labels are impractical. Incorporating partial label information and unlabeled information in semi-supervised learning paradigms can largely benefit models used in medical image processing domains.

## Chapter VII. 4-D White matter bundle population-based atlases derived from diffusion MRI fiber tractography

### 1. Introduction

**Note**: This chapter is the result of equal contribution by another author and myself.

The creation and application of medical image-based brain atlases is widespread in neuroanatomy and neuroscience research. Atlases have proven to be a valuable tool to enable studies on individual subjects and facilitate inferences and comparisons of different populations, leading to insights into development, cognition, and disease [3, 243-245]. Through the process of spatial normalization, images can be aligned with atlases to facilitate comparisons of brains across subjects, time, or experimental conditions. Additionally, atlases can be used for label propagation, where anatomical labels are propagated from the atlas to new data in order to identify a priori regions of interest. With these applications in mind, a number of human brain atlases have been created (Figure 1), with variations in the number of labels, the regions of the brain that are delineated, the methods used to generate labels, and the population or individuals used to create the atlas (for a review of the existing atlases and their standardization, see recent work by Lawrence et al. [3].

Despite the wide variety of human brain atlases available to the research community, there is a distinct lack of resources available to describe the white matter of the brain. For example, most atlases emphasize cortical or sub-cortical gray matter, and do not contain a label for white matter [99, 102, 103, 105-109, 246-257] or only label white matter as a single homogenous structure, or simply separate into the "cerebral white matter" of the left and right hemispheres [110, 111, 252, 258].

Some atlases do indeed include labels for white matter. However, in many cases these labels are for "regions" of the white matter rather than labels for specific white matter bundles [98, 112, 113, 259-261] For example, an atlas may contain a label for the "anterior limb of the internal capsule" or "corona radiata" which are descriptions of regions through which several white matter bundles are known to pass. While these regions are certainly scientifically useful, the white matter pathways themselves would be more informative for network neuroscience investigations or applications where white matter structure, connectivity, and location are paramount. Additionally, regional labels do not overlap, whereas the fiber bundles of the brain are known to be organized as a complex mixture of structures, overlapping to various degrees.

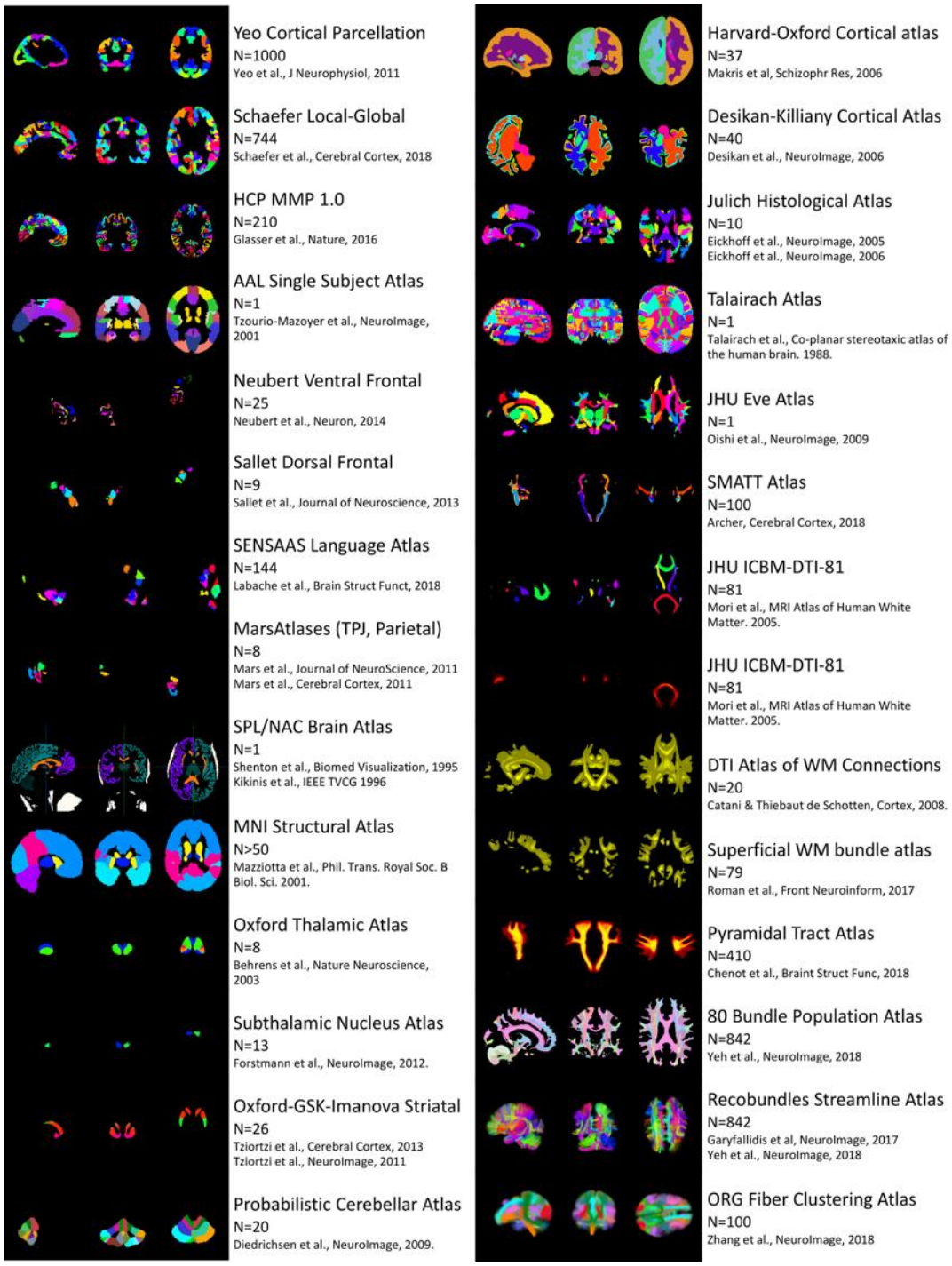


**Figure VII-1. Comparison of types of human brain atlases and regions present in each. Visualizations were made using FSLview tri-planar view for volumetric atlases and using MI-brain 3D-view for streamline atlases. Note that because atlases are in different spaces, visualized slices, anatomy, and orientation is not guaranteed to be the same across atlases. This figure is not exhaustive and is only representative of the types of atlases and the information they contain. In general, from top-to bottom, left-to-right, atlases focus on cortical and sub-cortical gray matter, to regional white matter labels, to tractography-derived white matter pathways, to streamline-based atlases. Figure inspired by work on standardizing gray matter parcellations (Figure 1 of Lawrence et al. [3]).**

To overcome these limitations, several atlases have been created using diffusion MRI fiber tractography, a technique which allows the investigator to perform a "virtual dissection" of various white matter bundles of the brain. Examples include population-based templates [262, 263] or atlases of association and projection pathways [66, 264-267], atlases of the superficial U-fibers connecting adjacent gyri [268, 269], and atlases created from tractography on diffusion data averaged over large population cohorts [265, 270, 271]. In particular, several atlases have been made with a focus on a single pathway or a set of pathways with functional relevance [272], for example the pyramidal tract [273], the sensorimotor tracts [274], or lobular-specific connections [66, 275, 276]. Existing tractography-based atlases, however, typically suffer from one or more limitations: (1) small population sample sizes, (2) restriction to very few white matter pathways, and (3) the use of out-dated modeling for tractography (specifically the use of diffusion tensor imaging which is associated with a number of biases and pitfalls). Further, it is not clear whether the same pathway defined using one atlas results in the same structure when compared to another atlas due to differences in the procedures utilized to define and dissect the bundle under investigation. A final type of atlas, streamline-based atlases [262, 269, 270, 277, 278] have become popular in recent years. These are composed of millions of streamlines and can be used as a resource to cluster sets of streamlines on new datasets, thus they nicely complement the use and application of volumetric atlases when diffusion MRI is available.

In this work, we introduce the Pandora* white matter bundle atlas. The Pandora atlas is actually a collection of 4-dimensional population-based atlases represented in both volumetric and surface coordinates in a standard space. Importantly, the atlases are based on a large number of subjects, and are created from multiple state-of-the-art tractography and dissection techniques, resulting in a sizable number of (possibly overlapping) white matter labels. In the following, we describe the creation of these atlases, the data records of the files and their formats, and validate the use of multiple subject populations and multiple tractography methodologies. The Pandora atlas is freely available (https://www.nitrc.org/projects/pandora_atlas; https://github.com/MASILab/Pandora-WhiteMatterAtlas) and will be a useful resource for parcellation and segmentation.

## 2. Methods

Figure 2 presents an overview of the pipeline and methodology used to create these atlases. Briefly, we retrieved and organized data from 3 large repositories (Figure 2, Data). For each subject, we performed six different automated methods of tractography and subsequent white matter dissection (Figure 2, Subject-level processing:

tractography), and registered all data to a standard volumetric space (Figure 2, Subject-level processing: registration). Next, a probabilistic map was created separately for each white matter bundle in standard space in order to create the volumetric atlases (Figure 2, Volumetric atlas creation). Finally, a surface mesh of the boundary between white and gray matter was created, and the volumetric maps were used to assign probabilities along this surface to create the surface-intersection atlases (Figure 2, Surface Atlas creation).

### 2.1. Data

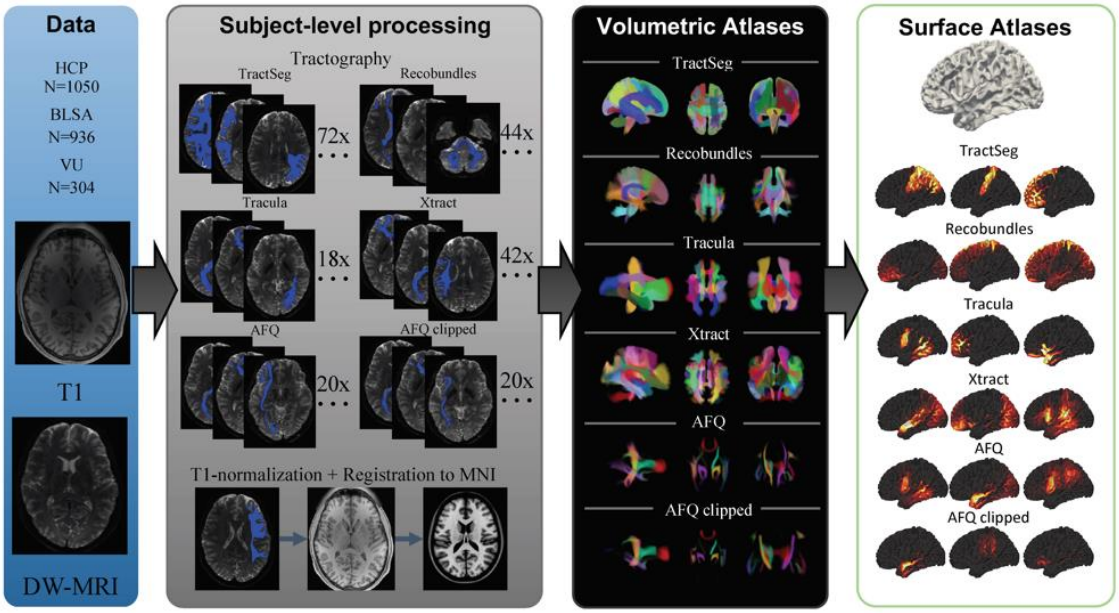


**Figure VII-2. Experimental workflow and generation of Pandora atlases. Data from three repositories (HCP, BLSA, and VU) were curated. Subject-level processing includes tractography and registration to MNI space. Volumetric atlases for each set of bundle definitions is created by population-averaging in standard space. Point clouds are displayed which allow qualitative visualization of probability densities of a number of fiber pathways. Finally, surface atlases are created by assigning indices to the vertices of the MNI template white matter/gray matter boundary.**

We used de-identified images from the Baltimore Longitudinal Study of Aging (BLSA), Human Connectome Project (HCP) S1200 release, and Vanderbilt University (Figure 2, Data). The BLSA is a long-running study of human aging in community-dwelling volunteers and is conducted by the Intramural Research Program of the National Institute on Aging, NIH. Cognitively normal BLSA participants with diffusion MRI data were included in the present study, using only one scan per participant, even if multiple follow-ups were available. HCP data are freely available and unrestricted for non-commercial research purposes, and are composed of healthy young adults. This study

accessed only de-identified participant information. All datasets from Vanderbilt University were acquired as part of a shared database for MRI data gathered from healthy volunteers. A summary of the data is given in Table 1, including number of subjects, age, sex, and handedness. All human datasets were acquired under research protocols approved by the local Institutional Review Boards.

All datasets included a T1-weighted image, as well as a set of diffusion-weighted images (DWIs). Briefly,

**Table VII-VII-1. Meta-data information. Note that several inputs are not provided due to confidentiality and data release agreements.**

| | HCP | BLSA | VU |
|---|---|---|---|
| **Subjects** | 1060 | 963 | 303 |
| **Age** | 28.8±3.5 | 66.2±14.82 | 29.7±11.5 |
| **Age Range** | [22 35] | [22.4 95.1] | [18 75] |
| **Handedness** | N/A | 86L; 843R; 35N/A | 30L; 270R; 3N/A |
| **Sex** | 488M; 572F | 431M; 532F | 134M; 169F |

the BLSA acquisition (Philips 3T Achieva) included T1-weighted images acquired using an MPRAGE sequence (TE = 3.1 ms, TR = 6.8 ms, slice thickness = 1.2 mm, number of Slices = 170, flip angle = 8 deg, FOV = 256x240mm, acquisition matrix = 256×240, reconstruction matrix = 256×256, reconstructed voxel size = 1x1mm). Diffusion-weighted images were acquired using a single-shot EPI sequence, and consisted of a single b-value (b = 700 s/mm2), with 33 volumes (1 b0 + 32 DWIs) acquired axially (TE = 75 ms, TR = 6801 ms, slice thickness = 2.2 mm, number of slices = 65, flip angle = 90 degrees, FOV = 212*212, acquisition matrix = 96*95, reconstruction matrix = 256*256, reconstructed voxel size = 0.83x0.83 mm). HCP acquisition (custom 3T Siemens Skyra) included T1-weighted images acquired using a 3D MPRAGE sequence (TE = 2.1 ms, TR = 2400 ms, slice thickness = 0.7 mm, flip angle = 8 deg, FOV = 224x224mm, acquisition, voxel size = 0.7x0.7mm). Diffusion images were acquired using a single-shot EPI sequence, and consisted of three b-values (b = 1000, 2000, and 3000 s/mm2), with 90 directions (and 6 b=0 s/mm2) per shell (TE = 89.5 ms, TR = 5520 ms, slice thickness = 1.25 mm, flip angle = 78 degrees, FOV = 210*180, voxel size = 1.25mm isotropic). The scans collected at Vanderbilt included healthy controls from several projects (Philips 3T Achieva). A typical acquisition is below, although some variations exist across projects. T1-weighted images acquired using an MPRAGE sequence (TE =2.9 ms, TR = 6.3 ms, slice thickness = 1 mm, flip angle = 8 deg, FOV = 256x240mm, acquisition matrix = 256×240, voxel size = 1x1x1mm). Diffusion images were acquired using a single-shot EPI sequence, and consisted of a single b-value (b = 1000 s/mm2), with 65 volumes (1 b0 + 64 DWIs per

shell) acquired axially (TE = 101 ms, TR = 5891 ms, slice thickness = 2.2 mm, flip angle = 90 degrees, FOV = 220*220, acquisition matrix = 144*144, voxel size = 2.2mm isotropic). Data pre-processing included correction for susceptibility distortions, subject motion, eddy current correction [134], and b-table correction [279].

**2.2. Subject-level processing: tractography**

Six methods for tractography and virtual bundle dissection were employed on all diffusion datasets in native space (Figure 2, Subject-level processing). These included (1) TractSeg [280] (2) Recobundles [277], (3) Tracula [264], (4) Xtract [281], (5) Automatic Fiber-tract Quantification (AFQ) [282], and (6) post-processing of AFQ where only the stem of the bundle was retained, which we call AFQ-clipped. Algorithms were chosen because they are fully automated, validated, and represent a selection of the state-of-the art methods in the field. In all cases, algorithms were run using default parameters or parameters recommended by original authors.

Briefly, TractSeg is based on convolutional neural networks and performs bundle-specific tractography based on a field of estimated fiber orientations [280, 283], and delineates 72 bundles. We implemented the dockerized version at which generates fiber orientations using constrained spherical deconvolution using MRtrix software [284]. Recobundles segments streamlines based on their shape-similarity to a dictionary of expertly delineated model bundles. Recobundles was run using DIPY [285] software after performing whole-brain tractography using spherical deconvolution and DIPY LocalTracking algorithm. The bundle-dictionary contains 80 bundles, but only 44 were selected to be included in the Pandora atlas after consulting with the algorithm developers based on internal quality assurance (for example removing cranial nerves which are often not used in brain imaging). Of note, Recobundles is a method to automatically extract and recognize bundles of streamlines using prior bundle models, and the implementation we chose uses the DIPY bundle dictionary for extraction, although others can be used. Tracula uses probabilistic tractography with anatomical priors based on an atlas and Freesurfer [286-288] cortical parcellations to constrain the tractography reconstructions. Tracula used the ball-and-stick model of diffusion from FSL's [289] bedpostx algorithm to reconstruct white matter pathways, and resulted in 18 bundles segmented per subject. Xtract is a recent automated method for probabilistic tractography based on carefully selected inclusion, exclusion, and seed regions, selected for 42 tracts in the human brain. Xtract also utilized the ball-and-stick model (bedpostx) of diffusion for local reconstruction. AFQ is a technique that identifies the core of the major fiber tracts with the aim of quantifying tissue properties within and along the tract, although we only extracted the bundle profile itself. The default in AFQ is to use tensor based deterministic tractography, followed by fiber segmentation utilizing methodology defined by

Wakana et al. [290], and removal of outlier streamlines, using AFQ_run MatLab script. In our case, we extracted the full profile of the bundle, as well as the core of the bundle which was performed in the AFQ software by a clipping operation (dtiClipFiberGroupToROIs). For this reason, we called these AFQ and AFQ-clipped, respectively. Both of these methods resulted in 20 bundles. In total, we present 216 bundles in the atlas. A list of the bundles from each pipeline is given in Appendix A.

Output from all algorithms were in the form of streamlines, tract-density maps, or probability maps. In all cases, pathways were binarized at the subject level, indicating the voxel-wise existence or non-existence of the bundle in that subject, for that pathway. These binary maps were used to create the population atlases after deformation to standard space.

Exhaustive manual quality assurance (QA) was performed on tractography results. QA included displaying overlays of binarized pathways over select slices for all subjects, inspecting and verifying appropriate shape and location of all bundles on all subjects. We note that not all methods were able to successfully reconstruct all pathways on all subjects, for this reason, some atlases contain information from fewer than all 2443 subjects.

### 2.3. Subject-level processing: registration

In order to create the atlases, all images were registered and transformed to a standard space (Figure 2, Subject-level processing). For this work, we chose the MNI standard space, a commonly used space in neuroimaging literature. To do this, the T1 image was intensity normalized using FreeSurfer's mri_nu_correct, mni, and mri_normalize which perform N3 bias field correction and intensity normalization, respectively on the input T1 image [291]. Next, the diffusion b0 image was coregistered to the T1 using FSL's epi_reg [292] (a rigid-body 6 degrees of freedom transformation). The T1 was then nonlinearly registered using ANTS antsRegistrationSyn to a 1.0 mm isotropic MNI ICBM 152 asymmetric template [140]. The FSL transform from epi_reg was converted to ANTS format using the c3d_affine_tool. Afterwards, all data could be transferred from subject native diffusion space to MNI space (and vice-versa) through antsApplyTransforms tools. Thus, all binarized pathways for all subjects were transformed to MNI space using both linear and nonlinear transforms. Transforms were also applied to the normalized T1 images to transform these structural images to standard space.

QA was performed to verify acceptable image registration. This again included generating and visualizing overlays of the b0 images, pathways, and T1 images in MNI spaces overlaid and/or adjacent to the MNI ICBM template image.

### 2.4. Volumetric atlas creation

Once all data were in MNI space, population-based atlases were created by following methods previously used to create tractography atlases [273, 293, 294] For each pathway, the binarized maps were summed and set to a probabilistic map between 0 and 100% population overlap (Figure 2, Volumetric Atlas). Thus, each pathway was represented as a 3D volume, and concatenation of all volumes results in the 4D volumetric atlas. Atlases were additionally separated based on the method used to create the atlas, as well as separated by dataset (BLSA, HCP, VU) if population-specific or method-specific analysis is required (see Technical Validation, below).

### 2.5. Surface-intersection atlas creation

To overlay each pathway onto the MNI template surfaces, a standard FreeSurfer pipeline [291] was used to reconstruct the white/gray matter cortical surfaces directly from the MNI ICBM template image. Each of the probability maps overlaid over the volumetric atlas was then transferred to the reconstructed surfaces to create the surface atlas. However, the reconstructed cortical surfaces do not necessarily guarantee unique voxel-to-vertex matching (normally, more than one vertex belongs to a single voxel) even if they perfectly trace the white- and gray-matter boundary. This potentially degenerates vertex-to-voxel mapping without a voxel-wise resampling scheme. Therefore, the probability to a given vertex was obtained by tri-linear resampling of the associated voxel for sub-voxel accuracy.

### 2.6. Data visualization and validation

Qualitative validation of the atlases included pathway visualization as an overlay of the population probability on the MNI ICBM template image, or visualization of population-probability on the white matter/gray matter surface. These displays were used in QA during atlas creation, ensuring acceptable probability values, as well as agreement with expected anatomy, shape, and location.

To quantify similarities and differences across pathways and methods, a pathway-correlation measure was used. The pathway-correlation was calculated between two pathways by taking the correlation coefficient of all voxels where either pathway has a probability > 0. This correlation coefficient ranged from -1 to 1, where a value of 1 indicates a perfect correlation of population densities. Thus, this metric measures the coherence between population maps obtained from the bundles and was used to assess if the distribution of population probabilities in space is similar. We used this measure to test similarities/differences between the pathways from different bundle dissection methods

(to justify the use of different tractography methods) as well as between pathways generated from the different datasets (to justify making available atlases separated by dataset, as well as understand differences in results based on populations).

Finally, a uniform manifold approximation and projection (UMAP) [230] was used for dimensionality reduction in order to further assess similarities and differences in pathways across methodologies. The UMAP is a general non-linear dimension reduction that is particularly well suited for visualizing high-dimensional datasets.

**3. Technical Validation**

We begin with a qualitative validation of the data, thoroughly inspecting and visualizing all volumes and surfaces from each atlas. An example visualization for 10 selected pathways from the TractSeg sets of atlases is shown in Figure 3. All pathways overlay in the correct location, with the correct shape and trajectory, as expected. Population agreement is generally high in the core of the bundle (values ~1) with larger variability along the periphery of pathways. Through this qualitative validation process, differences in the methodologies were noted including some possessing high sensitivity (larger volumes, greater agreement across subjects) and those with higher specificity (smaller, well-defined pathways with lower population agreement).

Next, to assess differences within and between tractography techniques, we show pathway-correlations against all other pathways as a large 216x216 matrix of correlations (Figure 4, a) and also plotting the UMAP projection of each pathway on a 2D plane (Figure 4, b). As expected, most pathways are quite different from others (for example we do not expect the optic radiations to share any overlap whatsoever with the uncinate fasciculus, regardless of methodology), however there are clearly clusters of pathways sharing some similarity, due to both spatial overlap of pathways with comparable anatomies (for example inferior longitudinal fasciculus and inferior frontal occipital fasciculus), as well as methods representing the same pathway. We identified a core group of 20 pathways that are commonly dissected in all methods, and clusters of these pathways are apparent in the UMAP projection (for example, the corticospinal tracts, forceps major and minor, optic radiations, and inferior longitudinal fasciculi are quite similar across algorithms). Thus, certain pathways are similar, but not exactly the same, across methodologies, justifying the use of all six state-of-the art methods for bundle dissection.

Finally, we quantify differences across datasets by showing boxplots of the pathway-correlations after separating by source of data (Figure 4, c). While all methods show quite high correlations, it is clear that BLSA and VU datasets and bundles are more similar to each other than to HCP datasets. This is expected as HCP data quality,

SNR, resolution, and acquisitions are quite different from the more clinically feasible BLSA and VU sets. Thus, bundles are also different based on dataset source. Because of this, in addition to combining results from all subjects, we also supply atlases separated by dataset.

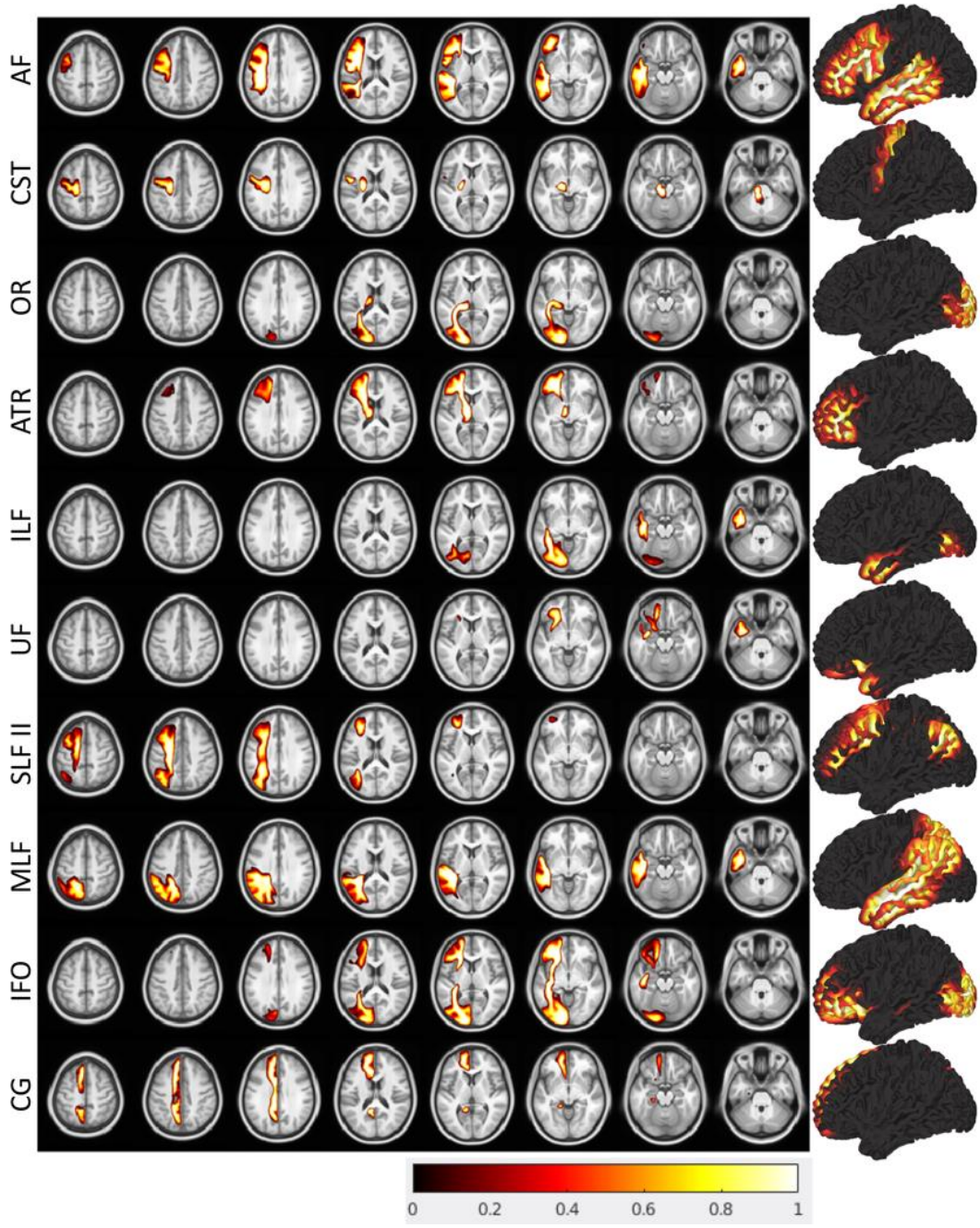


**Figure VII-3. Visualization of data contained in example volumetric and surface atlases. Example visualization for 10 pathways in the TractSeg nonlinear atlas are shown as both overlays and surfaces.**

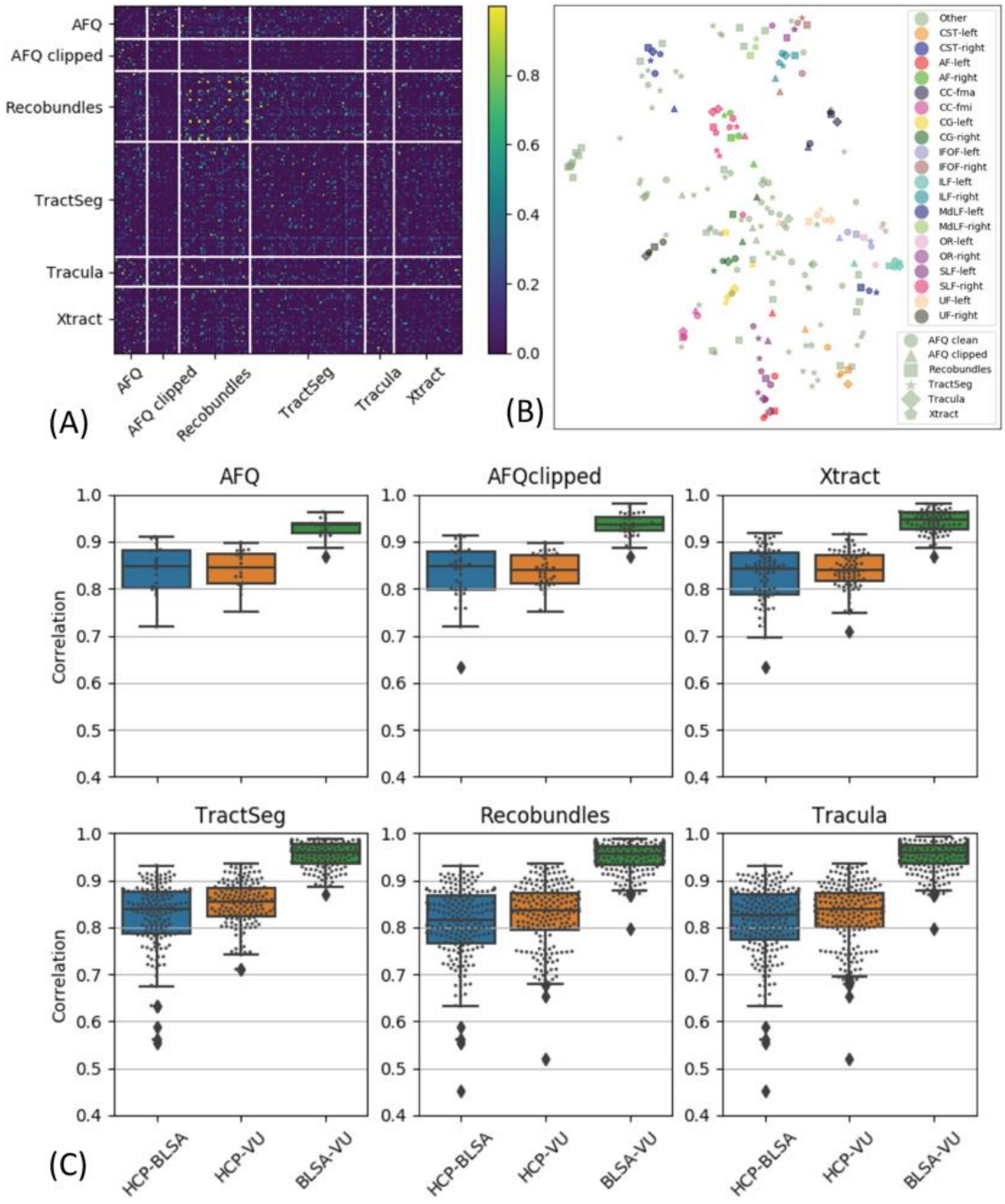


**Figure VII-4. Data validation. (A) Matrix of correlation coefficient of pathways plotted against all others indicates similarities within and across methodologies for bundle dissection. Solid white lines are used to visually separate bundle segmentation methods. (B) UMAP dimensionality reduction projected onto un-scaled 2D plane shows that many WM pathways are similar, but not the same, across methods. Object colors represent specific atlas bundles, with shape indicating segmentation methods. (C) Correlation coefficient of atlases separated by dataset indicates small, but significant, differences between datasets. Together, these justify the inclusion of all tractography methods, as well as separation of atlases by datasets.**

**4. Usage**

Here, we have created and made available the Pandora white matter bundle atlas, that addresses a number of limitations of current human brain atlases by providing a set of population-based volumetric and surface atlases in standard space, based on a large number of subjects, including many pathways from multiple diffusion MRI tractography bundle segmentation methods. We envision the use of these atlases for spatial normalization and label propagation in ways similar to standard usage of volumetric brain atlases. These labels can be used not only for statistical analysis across population and individuals, but also for priors for tractography, relating neuroimaging findings to structural pathways or to inform future methodologies for parcellating and segmenting white matter based on functional, molecular, or alternative contrasts. Similarly, although much less frequently used in the field, the surface-based atlas can also be used to relate functional MRI findings (which are largely applied to cortex, with some evidence for signal contrast in white matter), as priors for cortico-cortical tractography and future bundle segmentations, as a tool for gray matter based spatial statistics, and again for relating alternative neuroimaging findings to structure.

As a simple example workflow. An investigator may be interested in relating tumour localization on a structural image to specific white matter pathways hypothesized to be involved in some functional network. The investigator may choose to register their image to the MNI template, and can either warp their data to template space or apply the inverse transform to get white matter labels into the subject native space. The investigator could then relate tumour location to the probability of given pathways, or could simply threshold the probabilistic maps at a given threshold (for example 0.5) and relate these to the existence/non-existence of the bundle being displaced by the tumour.

We currently recommend the use of the concatenation of all datasets for standard investigative studies unless a population-specific template is required. While differences between datasets are clear and expected, the increased population variability that results from including data from all sources is likely an advantage when investigators are using their own data with possible differences in acquisition, resolution, and subjects. However, future work will investigate creation and dissemination of age-specific white matter analysis, as well as including an age-adjusted surface mesh instead of using the MNI template to generate the surface.

We have chosen to include a large number of algorithms for streamline generation and bundle dissection. Our results (Figure 4) show that even if the same white matter structure is segmented using different techniques, the results are not guaranteed to be the same. This is because different algorithms or workflows may define bundles in

different ways (Kurt G. Schilling et al. 2020), with different approaches taken to segment the structure of interest. Thus, an investigator could use our atlas with the set of protocols that they agree with most, or alternatively, could relate findings to all white matter pathways across all methodologies in our atlas. We note that we have chosen six standard algorithms to create this atlas, although others exist and new ones are continually developed based on improvements in both our understanding of anatomical connections and our ability to reconstruct these connections with tractography. These methods were chosen because they are fully automated, and robust, bundle segmentation techniques that can be easily run on several thousand diffusion datasets.

Inclusion of other tractography and/or segmentation methods are likely additions in future iterations of the atlas, and are easily integrated with existing deformation fields and data organization. The addition of tract orientation maps [280] or orientation-density maps [295, 296] may facilitate the development of bundle segmentation algorithms or act as priors for bundle specific tractography. Finally, future iterations can include variations and concatenations of gray matter and/or regional atlases in the same space, continually adding to the number of features to be investigated with a single dataset in standard space.

## Chapter VIII. Learning white matter subject-specific segmentation from structural MRI

### 1. Introduction

**Note**: This chapter is the result of equal contribution by another author and myself.

Mapping brain WM is essential for building an understanding of brain function and dysfunction [297]. Currently, the only approach for mapping WM relies on dMRI based tractography in vivo [298]. This approach estimates local fiber orientations by measuring the movement of water from dMRI, allowing fiber tracts or streamlines to be computed [299]. The subsequent dissection of streamlines from across the brain, or the whole-brain tractogram, allows for the segmentation of WM pathways, or bundles, which can be used to study brain anatomy [300], development [301], cognition [302], and neurological disease [303].

DMRI appropriate for tractography can be challenging to acquire. These acquisitions often require many gradient directions and high b-value shells that are not commonly acquired in clinical settings and require long scan times [304]. Tractography and WM bundle segmentation are also impossible for retrospective studies without dMRI. The computation and dissection of whole brain tractograms is also time-consuming, which limits the applicability of this technique in time-constrained settings. On the other hand, structural T1 weighted (T1w) MRI acquisitions are widely used in neuroimaging research and in the clinical setting. However, there currently does not exist a method to directly delineate WM pathways from T1w MRI, as contrast within WM is typically poor in T1w MRI.

Fortunately, image registration, an established way of transferring different WM labels from population-based atlases to T1w MRI, can help to solve this issue and isolate different WM regions in T1w MRI. In general, WM atlases from the dMRI community can be divided into two categories: streamline-based atlases [270, 273, 278, 305] and volumetric atlases [265, 268, 306]. Streamline-based WM atlases contain streamlines assigned to various WM pathways derived from dissected whole brain tractograms, while volumetric WM atlases contain labels indicating the pathway assignment(s) of a given voxel. For determining WM labels on T1w MRI, volumetric atlases are more commonly used. One such widely used atlas was proposed by Mori et al. and recognizes 48 different WM labels [265]. Using the same population of subjects, Oishi et al. proposed an atlas to model superficial WM [268]. These atlases have become very popular in neuroimaging analysis but have key limitations. They require that different WM regions are not overlapping and often contain a limited amount of information outside deep WM (Figure 1). To navigate these limitations, Hansen et al. recently proposed the Pandora WM bundle atlases. Those atlases are volumetric atlases,

obtaining 216 overlapping WM pathways from 2300 healthy subjects. This approach has subsequently allowed for both the identification of overlapping pathways and improved WM labeling outside the deep structures on T1w MRI without dMRI [306].

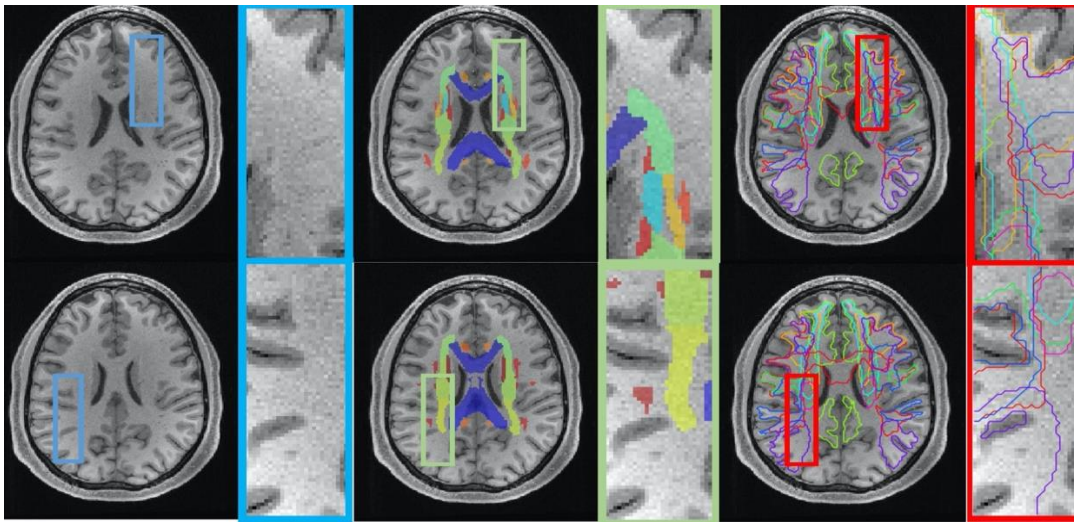

**Figure VIII-1. WM is largely homogenous when imaged using most sources of MRI contrast, for example T1w (left). Traditional WM atlas (center) represents each voxel with one tissue class. Modern approaches at bundle segmentation identify multiple overlapping structures (as shown right). Diffusion tractography offers the ability to capture a multi-label description of WM voxels.**

Despite advancements that have been made in population-based WM atlases in recent years, they all are inherently limited in their subject-specificity as they are built from large cohorts. However, deep convolutional neural networks, which have the potential to capture subject-specific variations. Among convolutional neural networks, the U-net [307] and V-net [308] have obtained impressive results for performing 3D medical image segmentation. Brebisson et al. proposed a deep neural network learning 2D and 3D patches from structural brain MRI to predict the anatomical class of each voxel [120]. DeepNat leverages a hierarchical multi-task network to achieve brain segmentation with 3D patches [122]. SLANT [309], proposed by Huo et al., learns spatially localized 3D patches from structural MRI to achieve brain structure segmentation. Additionally, current deep learning approaches [120, 310] have demonstrated improved performance compared with atlas-based methods on healthy brain segmentation from structural images.

Thus, driven by the need for improved subject-specificity in WM segmentation and inspired by previous work leveraging deep learning to segment the whole brain, we propose a spatially localized patch-wise framework to delineate WM regions from structural T1w MRI. To achieve this, we select six state-of-the-art tractography algorithms

to reconstruct WM pathways from dMRI on a subject-by-subject basis. We then register the bundles and T1w MRI of the same subject to standard place to serve as a ground-truth during supervised training and subsequently produce and characterize six deep learning algorithms to perform subject-specific WM bundle segmentation from T1w MRI. We envision this framework as a tool for researchers to localize white matter regions when dMRI is not available.

## 2. Material and Methods

The aim of this study is to predict WM labels directly from structural T1w MRI with deep learning. To do this, we use T1w MRI as inputs for the proposed deep neural networks. Then, we derive WM bundles from dMRI-derived tractography and convert them to voxel-wise WM bundle labels that can be mapped to T1w MRI on a standard template and serve as a ground truth during the supervised training process. The proposed method includes tractography, registration, normalization, and patch-wise networks (Figure 2). In short, we build six patch-wise U-Nets to predict WM bundles defined by each of six dMRI-based tractography bundle segmentation algorithms from structural T1w MRI. We will use the following names to represent each bundle segmentation algorithm: TractSeg, RecoBundles, XTRACT, Tracula, AFQ and AFQclipped. We divide input T1w MRI into 125 localized patches and feed patches into corresponding U-Nets to obtain output for each neural network. Then, we merge all output to get the final result in the form of an average.

**Table VIII-1. Dataset descriptions. * represents one typical case selected from the VU dataset.**

| Dataset Name | T1w voxel size(mm) | Diffusion voxel size (mm) | B-value | Diffusion volume |
|---|---|---|---|---|
| BLSA | 1.0 ×1.0 × 1.2 | 0.81 × 0.81 × 2.2 | 700 | 1B0+32DWIs |
| HCP | 0.7 × 0.7 × 0.7 | 1.25 × 1.25 × 1.25 | 1000,2000,3000 | (6B0+90DWIs) x3 |
| VU* | 1 × 1× 1 | 2.5 × 2.5 × 2.5 | 1000 | 1B0+64DWIs |
| HCP_LS | 0.8 × 0.8 × 0.8 | 1.5 × 1.5 × 1.5 | 1000,2500 | 5B0+76DWIs |
| IXI | 0.93 × 0.93 × 1.2 | 1.75 × 1.75 × 2.35 | 1000 | 1B0+15DWIs |
| UG | 1 × 1 × 1 | 2 × 2 × 2 | 2000 | 6B0+48DWIs |

### 2.1. Data

We use 2,416 de-identified images from the Baltimore Longitudinal Study of Aging (BLSA) [311], 1,105 images from Human Connectome Project (HCP) S1200 release [20], and 349 images from Vanderbilt University (VU) to train all deep neural networks. We also select three open-source datasets to perform external validation to test the

generalizability of the proposed learning method. We study 26 images from HCP lifeSpan (HCPLS) [20], 394 images from IXI (IXI, http://brain-development.org/ixi-dataset), and 12 images from the Unilateral Glaucoma dataset (UG, https://openneuro.org/datasets/ds001743/versions/1.0.1). All above images include paired T1w MRI and dMRI. The specific voxel size, and diffusion b-values are shown in Table 1.

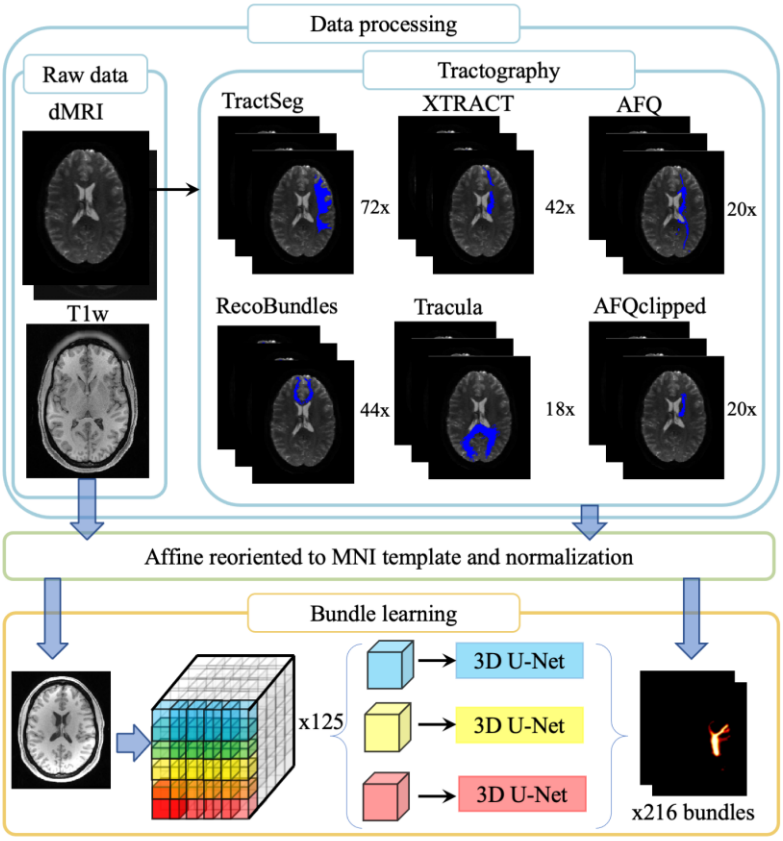


**Figure VIII-2. The pipeline of proposed WM bundle learning is presented, which integrates data processing and registration as well as bundle learning. We extract WM bundles from six different tractography methods. Structural images and corresponding tractograms are reoriented to the MNI template. Patch-wise, spatial-localized neural networks are utilized to learn WM bundle regions from a T1w MRI image. The output of each U-net is concatenated as the final step before segmentation. Representative samples of WM bundles acquired from six automatic tractography methods and the final learning result is visualized.**

### 2.2. Tractography

dMRI is often subject to artifacts, which can deteriorate the accuracy of extracting WM bundles. In order to correct these artifacts, we perform correction for susceptibility distortions, subject motion, eddy currents, and b-tables prior to analysis [312].

We perform tractography on preprocessed dMRI. Diffusion tractography is a tool for extracting WM

pathways non-invasively and in vivo. We select six popular tractography algorithms to recognize pathways and annotate WM bundles. All six algorithms were run using default parameters. (1) TractSeg uses convolutional neural networks to extract bundles from fiber orientation distribution function peaks [280], and obtains 72 pathways per scan. (2) RecoBundles, implemented in the Dipy software package [285], utilizes streamline-based clustering and assigns streamlines to bundles defined by streamline-based atlases [277]. The RecoBundles atlas contains 80 bundles, but we select the most robust 44 for the present study after consulting with the algorithm developers. (3) XTRACT uses a region of interest (ROI)-based protocols to identify specific WM pathways, including seeding area and start and end points applicable to both human and non-human species [281]. XTRACT generates 42 pathways per subject. (4) Tracula [264] is a global probabilistic tractography algorithm that constrains its bundle search space by penalizing connections that do not match anatomical priors, and generates 18 WM bundles per subject. (5) AFQ calculates a whole-brain deterministic tractogram from tensor representations of dMRI and parcellates them into 20 bundles using a fiber tract probability atlas [282]. (6) AFQclipped clips the center of each of the 20 AFQ bundles with ROI-based exclusion criteria. The output bundles of all six algorithms are finally converted into binary masks with their built-in function for each of the WM bundles.

### 2.3. Registration and intensity normalization

To ensure that all inputs have the same image resolution, voxel size, and coordinate space, we register all pathways derived from dMRI through all six bundle segmentation algorithms and T1w MRI, to the Montreal Neurological Institute (MNI) ICBM 152 asymmetric template [313]. First, we rigidly register the b = 0 s/mm2 volume of each dMRI to the T1w MRI of the same subject using FSL [292]. Then, after performing N3 correction of bias field and normalization of white matter intensity by FreeSurfer [291] on raw T1w MRI, corrected T1w MRI is registered to the MNI template with antsRegistrationSyn in ANTs [140]. By linking these registration steps, all pathways are rigidly registered to T1w MRI of the same subject. Then, all pathways are affine reoriented to the MNI template and serve as ground truth. The affine transformation is also applied to the raw T1w MRI.

After registration, we then skull strip all structural images on the MNI template with the bet tool in FSL and clip and normalize the background and the 98th percentile of within-brain intensity to arbitrary intensity units 0 and 1.

### 2.4. Patch-wise network

With all input images and corresponding ground truth WM bundles registered to the MNI template (1mm isotropic, 193 × 229 ×193 voxels), the high-resolution image volume could not fit into the 12G GPU (GTX 1080Ti) memory using current popular network architectures. Inspired by SLANT [309], we designed 125 overlapped 3D U-Nets to cover the entire MNI volume. As input to each 3D U-Net, we subdivide each image into 96×96×96 voxel images or patches. The division strategy of each patch can be shown as below:

$$\mathcal{H}_n = \left[x_{n_{start}} : (x_{n_{start}} + 96), y_{n_{start}} : (y_{n_{start}} + 96), z_{n_{start}} : (z_{n_{start}} + 96)\right] \quad (1)$$

where $\mathcal{H}_n$ represents the $n$th sub-space, $x_{n\_start}$, $y_{n\_start}$,$z_{n\_start}$ represent the corner coordinates of the $n$th sub-space. $x_{n_{start}}$ and $z_{n_{start}} \in [1,25,50,74,98]$ and $y_{n_{start}} \in [1,34,67,101,134]$. The training process for 125 models is time-consuming. Inspired by AssembleNet [314], we adopted transfer learning technology which shares knowledge among neighboring patches and decreases training time to a large extent.

In order to merge the outputs of the U-Nets after training, the pixel-wise output represents an activation value of the neural network rather than specific WM pathways. Thus, the majority vote cannot be directly applied. Instead, the average way is adopted to get the final value:

$$p_{whole}(i) = \frac{1}{n_k} \sum_{k=1}^{n_k} p_k(i) \quad (2)$$

where $p_{whole}$ represents all pixels within the structural image and $p_{whole}(i)$ means the *i*th pixel. $k$ indexes the U-Nets that covers *i*th pixel. $p_k(i)$ represents the final value of *i*th pixel of *kth* U-net. Networks not covering a particular voxel are excluded in the final merge process.

### 2.5. Implementation details

We divided the HCP, BLSA, and VU data into training, validation, and test cohorts evenly based on subjects. We kept the splitting strategy consistent across learning all six diffusion tractography algorithms. To remove data corrupted by registration or failed diffusion tractography algorithms, all registration and tractography results are reviewed to verify alignment and WM segmentations. The resultant number of scans for the training, validation, and testing cohorts is shown in Table 2. The number of scans in the external datasets is shown in Table 3.

**Table VIII-2. The train, validation and test size for all six learning algorithms**

| | Scans of train | Scans of validation | Scans of test |
|---|---|---|---|
| TractSeg | 2803 | 213 | 754 |
| RecoBundles | 2789 | 211 | 754 |
| XTRACT | 2786 | 211 | 751 |
| Tracula | 2538 | 189 | 693 |
| AFQ | 2730 | 201 | 726 |
| AFQclipped | 2730 | 201 | 726 |

**Table VIII-3. The size of external dataset for all six algorithms**

| | Number of scans |
|---|---|
| TractSeg | 431 |
| RecoBundles | 430 |
| XTRACT | 427 |
| Tracula | 428 |
| AFQ | 367 |
| AFQclipped | 367 |

We use a baseline U-Net [307] as the convolutional neural network to learn patches from anatomical images. Each input patch size is 96 × 96 × 96 and we set batch size to 1. The output channel depends on the number of WM bundles recognized by the bundle segmentation algorithm. We set a learning rate of 0.0001 and do not perform learning rate decay during the training process. We adopt the sum binary cross-entropy for each effective WM bundle as a loss function and train all models using the Adam optimizer. In order to save time training 125 models for each tractography method, nine out of the 125 models are trained with ten epochs until validation loss is converged. The corner coordinate of 9 models and index from the corner coordinate are shown in Table 4. The model with index [3,3,3] is the central part of the brain, which is added into pre-training since it contains more anatomical structure compared with peripheral models. The other eight models are distributed evenly over cube vertices centered on the central model. The final weight would be loaded as initial weights for neighboring neural networks. With transfer learning, the neighboring neural networks are trained with 3 epochs. When we infer the WM regions based on deep neural networks, we append a sigmoid function to the output of each patch-wise neural network to map the final merged output to [0,1].

### 2.6. Atlas-based method

We compare the quantitative performance of transferring labels with the traditional atlas-based approach as the baseline method. Here, we use the Pandora atlas [306], which is a 4D collection population-based atlases. The Pandora atlas used the same cohorts and diffusion tractography algorithms to generate each corresponding WM bundle within the atlas that we learn in this study. All volumes of the Pandora atlas are on the same MNI template as we use here. Each volume of the 4D atlas is in the form of a probability map indicating a probability of a pixel being in a specific WM bundle. In order to compare WM segmentation results performed by label propagation of the Pandora atlas to those produced here, the input T1w image is matched to MNI template through affine and deformable registration by ANTs. Then, the atlas is reoriented to the MNI template by the inverse deformable field as the final probability map after transformation.

**Table VIII-4. Corner coordinates of pre-trained nice models out of 125 models, indexed starting at one.**

| Corner coordinate index | Corner coordinate (x,y,z) |
|---|---|
| 2,2,2 | 25,34,25 |
| 2,4,2 | 25,101,25 |
| 4,2,2 | 101,25,25 |
| 4,4,2 | 101,101,25 |
| 3,3,3 | 50,67,50 |
| 2,2,4 | 25,34,74 |
| 2,4,4 | 25,101,74 |
| 4,2,4 | 101,25,74 |
| 4,4,4 | 101,101,74 |

### 2.7. Metrics

To evaluate the accuracy of our proposed method, we compare the segmentation results against the ground truth provided by diffusion tractography. Additionally, we compare the accuracy of the proposed method against the accuracy achieved with the use of the population-based Pandora atlas. To quantify the agreement between segmentation and truth, we use four measures: Dice coefficient (DSC), average symmetry surface distance, bundle overlap, and bundle overreach.

We use DSC as the main evaluation measurement for different bundle segmentation algorithms by comparing binary WM bundle prediction against the ground truth voxel-by-voxel:

$$DSC = \frac{2|R \cap T|}{|R| + |T|} = \frac{2|TP|}{2|TP| + |FP| + |FN|} \quad (3)$$

where $TP$ is true positive, $FP$ is false positive, $FN$ is false negative, $R$ represents the segmentation result generated by the proposed method or atlas-based method and $T$ represents the corresponding ground-truth.

Average symmetry surface distance [315] is given in millimeters and based on surface vertices between the proposed or atlas-based segmentation, R, and the ground-truth segmentation, T. For each vertex on the surface of R, (S(R)), the Euclidean distance to closest surface vertices of truth (S(T)) can be defined in $d(S_R, S(T))$ :

$$d(S_R, S(T)) = \min_{s_R \in S(R)} || S_R - S(T)||$$

$$ASSD = \frac{1}{|S(R)| + |S(T)|}( \sum_{s_R \in S(R)} d(s_R, S(T)) + \sum_{s_T \in S(T)} d(s_T, S(R))) \quad (4)$$

where $|S(R)|$ represents the number of vertices of the resulting surface and $|S(T)|$ represents the number of vertices on the ground-truth surface. $S_R$ represents a vertex from the atlas-based or proposed segmentation. $S_T$ represents a vertex from the ground-truth segmentation.

Bundle overlap is the proportion of voxels that contain the ground truth region that is also overlapped by the results of the learning- and atlas- based methods.

$$OL = \frac{R \cap T}{T} = \frac{|TP|}{|TP| + |FN|} \quad (5)$$

Bundle overreach is the number of voxels containing results from proposed and atlas-based methods that are outside of the ground truth volume divided by the total number of voxels within the ground truth.

$$OR = \frac{R \setminus T}{T} = \frac{|FP|}{|TP| + |FN|} \quad (6)$$

where operator \ denotes the relative complement operation.

The non-parametric Wilcoxon signed-rank test [316] for paired distributions was used to calculate test significance when comparing learning-based results with corresponding atlas-based results.

## 3. Results

### 3.1. Fine-tune binary threshold

The outputs of the atlas-based and proposed methods have been mapped to [0,1] and represent a probability that a given voxel is included in the WM pathway. The binary threshold to convert the probability to a yes or no is important and influences the performance of both the atlas-based and proposed methods. Starting from 0, we sweep

thresholds until 1 with a step size of 0.01, using the validation datasets to calculate mean DSC across all WM pathways of all scans. The curves of the relationships between mean DSC and binary threshold for the atlas- and learning- based methods are shown in Figure 3. The optimal thresholds are the values where the mean DSC across all pathways from all scans are highest for atlas- and learning- based methods. The optimal threshold values are shown in Table 5.

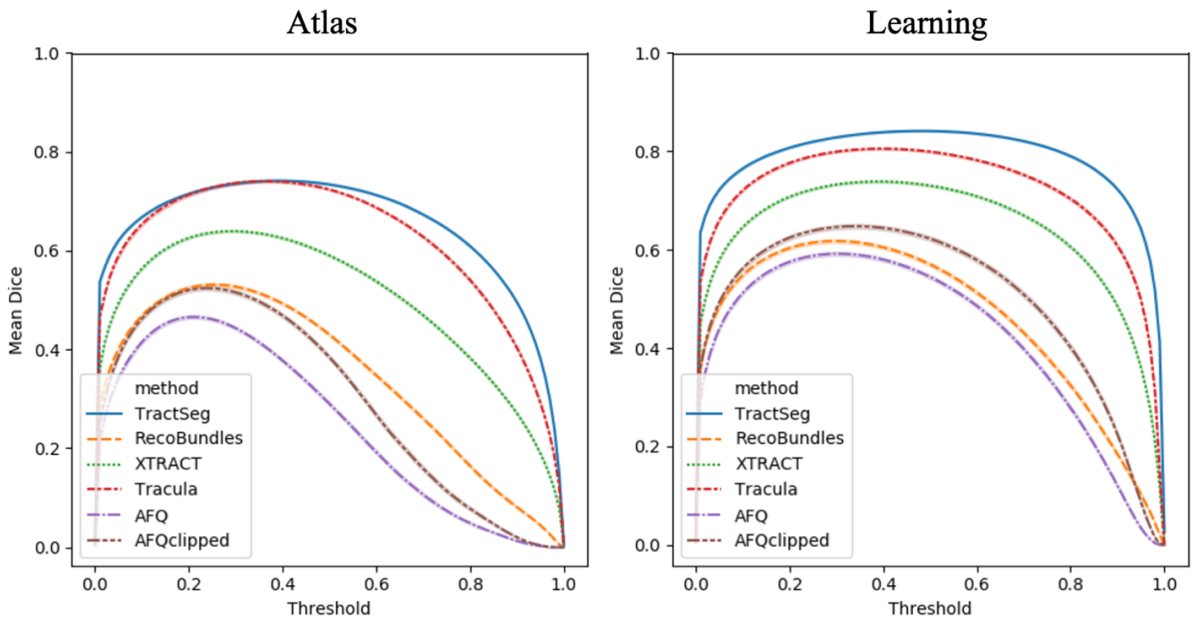


**Figure VIII-3. Each curve represents the average DSC of all WM bundles of all validation dataset scans per diffusion tractography algorithms for atlas- and learning- based methods at different threshold values.**

**Table VIII-5. The optimal threshold values for the atlas-based and proposed method fine-tuned on validation and external**

| | Validation dataset | | External dataset | |
|---|---|---|---|---|
| | Atlas | Learning | Atlas | Learning |
| TractSeg | 0.39 | 0.48 | 0.46 | 0.50 |
| RecoBundles | 0.25 | 0.30 | 0.20 | 0.31 |
| XTRACT | 0.30 | 0.39 | 0.56 | 0.77 |
| Tracula | 0.36 | 0.40 | 0.20 | 0.24 |
| AFQ | 0.21 | 0.30 | 0.30 | 0.43 |
| AFQclipped | 0.24 | 0.34 | 0.31 | 0.46 |

### 3.2. Qualitative results

We select one scan from the HCP test cohort to visualize the left corticospinal tract (CST) across all six bundle segmentation algorithms to see an intra-subject variance of bundle segmentation algorithms and visualize the difference between results derived from T1w images and ground truths from dMRI. We use the optimal threshold values calculated in Table 5 to binarize each output, using marching cube [317] to extract and render the CST surface. 3D visualization is shown in Figure 4.

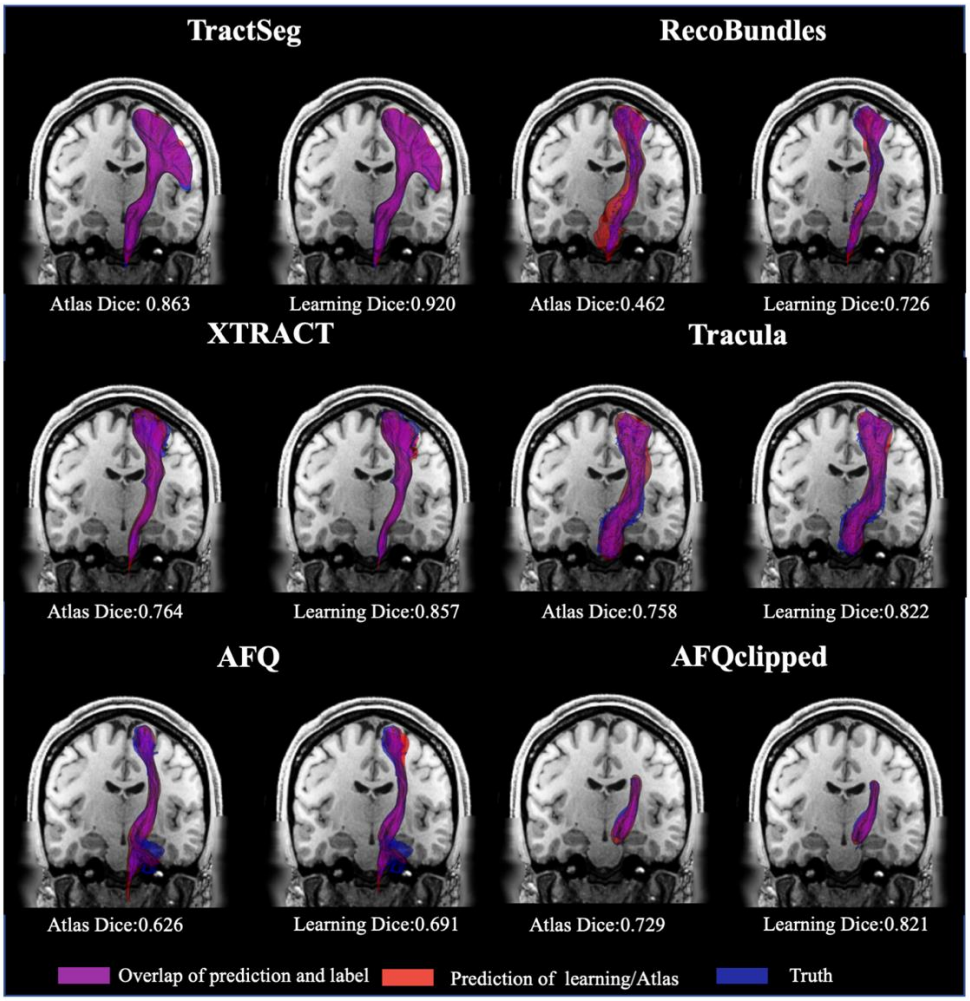


**Figure VIII-4. 3D visualization of atlas- and learning- based results across six diffusion tractography algorithms by reconstruction of the left corticospinal tract (CST) surface on an affine reoriented coronal T1w MRI slice. The text below each image is quantitative DSC for each case.**

From Figure 4, we find the learning-based method per bundle segmentation algorithm has a higher overlap compared with the atlas-based methods according to the areas of magenta overlap for this subject. In this case, the learning-based method performs best on the TractSeg bundle segmentation algorithm. The atlas-based method performs worse on the RecoBundles method because of low threshold leading to over-segmentation. All group truth CSTs have some common parts but those are not exactly the same across from all six algorithms. The CST from TractSeg has the largest volume starting from the brainstem and fanning out through the corona radiata, almost reaching the cortex, while CST from AFQclipped has much smaller volume compared with TractSeg.

### 3.3. Quantitative results

We used the optimal threshold values fine-tuned from the validation datasets to binarize the output on the testing datasets. To examine their overall performance, we evaluated all 216 bundles using the DSC and average symmetry surface distance (Figure 5).

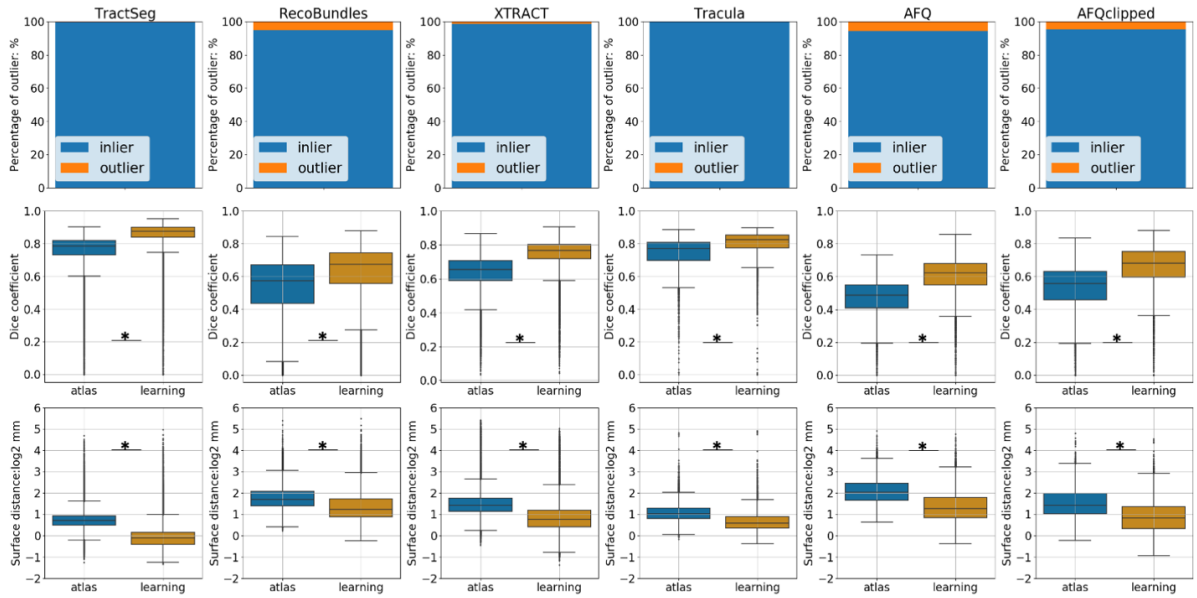


**Figure VIII-5. Quantitative results of atlas-based method and proposed learning methods on test cohorts from HCP, BLSA, and VU. The outlier percentage (top row) of all six algorithms on test cohort is shown in bar plot. Two measures are used to assess the overlap between algorithms deriving fiber mask from T1w and truth from dMRI: Dice (middle row) and surface distance (lower row). Each column presents the result of a different bundle segmentation algorithm and shows the proposed method against an atlas-based registration. Each boxplot includes each pathway of the bundle segmentation algorithm per every scan in the test cohort. The difference between methods was significant (p <0.005, Wilcoxon signed-rank test, indicated by *)**

From Figure 5, the blue bar plot represents the percentage of pathways that successfully passed the human reviewing process across the whole test cohort. All learning-based methods perform statistically better than the atlas-

based methods. When using ground truths derived from TractSeg, the atlas- and learning- based methods achieve the highest median DSC of 0.78 and 0.87 and smallest average symmetry surface distance 1.62 mm and 0.92 mm respectively. Compared with the atlas method, the learning method shows the largest improvement in median DSC for AFQ from 0.48 to 0.62 and reduces median average symmetry surface distance from 4.08 mm to 2.40 mm.

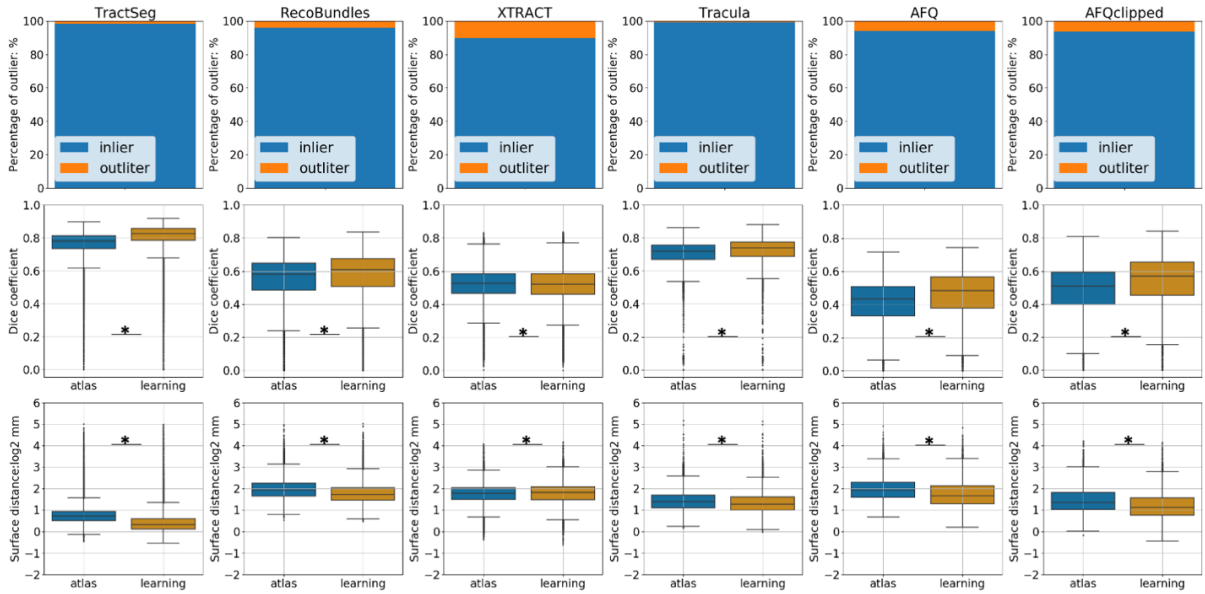


**Figure VIII-6. Quantitative results of atlas-based methods and proposed learning methods on external datasets. The outlier percentage (top row) of all six algorithms is shown in the bar plots. Two measures are used to assess the overlap between algorithms deriving fiber mask from T1w MRI and truth from dMRI: Dice (middle row) and surface distance (lower row). Each column presents the result of a bundle segmentation algorithm and shows the proposed method against an atlas-based registration. Each boxplot includes each pathway of bundle segmentation algorithm per every scan in the external dataset. The difference between methods was significant ($p < 0.005$, Wilcoxon signed-rank test, indicated by *)**

In Figure 6, the blue bar plot represents the percentage of pathways that successfully passed the human reviewing process across the external dataset. All learning-based methods perform statistically better compared with atlas-based methods except for XTRACT. However, the difference between the atlas- and learning- based methods is less pronounced. The median DSC of the learning-based method on XTRACT is 0.522, lower than 0.527 of the atlas-based method. Compared with atlas-based methods, the learning-based method makes the largest improvement on AFQclipped, increasing median DSC from 0.51 to 0.57 and decreasing median average symmetry surface distance from 2.55 mm to 2.17 mm.

We analyze the relationship between the measures of overlap and overreach and the threshold used for

binarization on external datasets. Starting from 0, we sweep the threshold until 1 with step 0.01. At each level, we calculated the two measures for all 216 WM pathways defined from all six bundle segmentation algorithms per scan for both the atlas- and learning- based method respectively. We use the left CST to show the variance of bundle overlap and bundle overreach along with thresholds across all six bundle segmentation algorithms (Figure 7).

From Figure 7, all learning- and atlas- based methods for all six diffusion tractography algorithms identified WM bundles with a high overlap but suffer from high overreach except for Tracula. As for AFQ, when the overlap value reaches about 0.9, the proposed learning method suffers from overreach by as much as 5 – 6 times the actual ground truth volume. The atlas-based method suffers from overreach by as much as 7 – 8 times the actual ground truth volume if the overlap is also about 0.9.

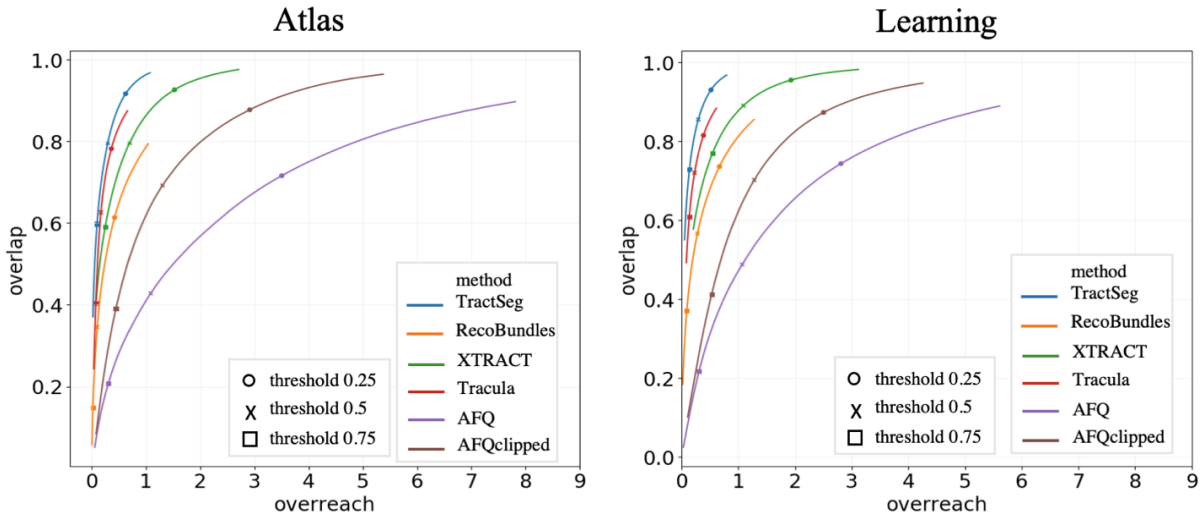


**Figure VIII-7. Plots of overlap versus overreach for the left CST across all bundle segmentation algorithms for atlas- and learning- based methods are shown. The markers on each curve to represent the overlap and overreach values at specific threshold values. The range of overreach for atlas-based methods is [0,9]. The range of overreach for the learning-based method is [0,6]**

We already calculated the overall binary thresholds from the validation cohort. Thus, we want to investigate whether the optimal thresholds calculated from validation datasets can generalize to external datasets. We show the curve of relationship between DSC and binary threshold on the external datasets in Figure 8. The thresholds calculated on external datasets are shown in Table 5. Comparing Figure 8 to Figure 3, the biggest difference between thresholds estimated from the validation dataset and the external datasets is in XTRACT. The binary threshold for atlas-based method in XTRACT is shifted from 0.30 to 0.56. The binary threshold for the proposed method in XTRACT is shifted from 0.39 to 0.77.

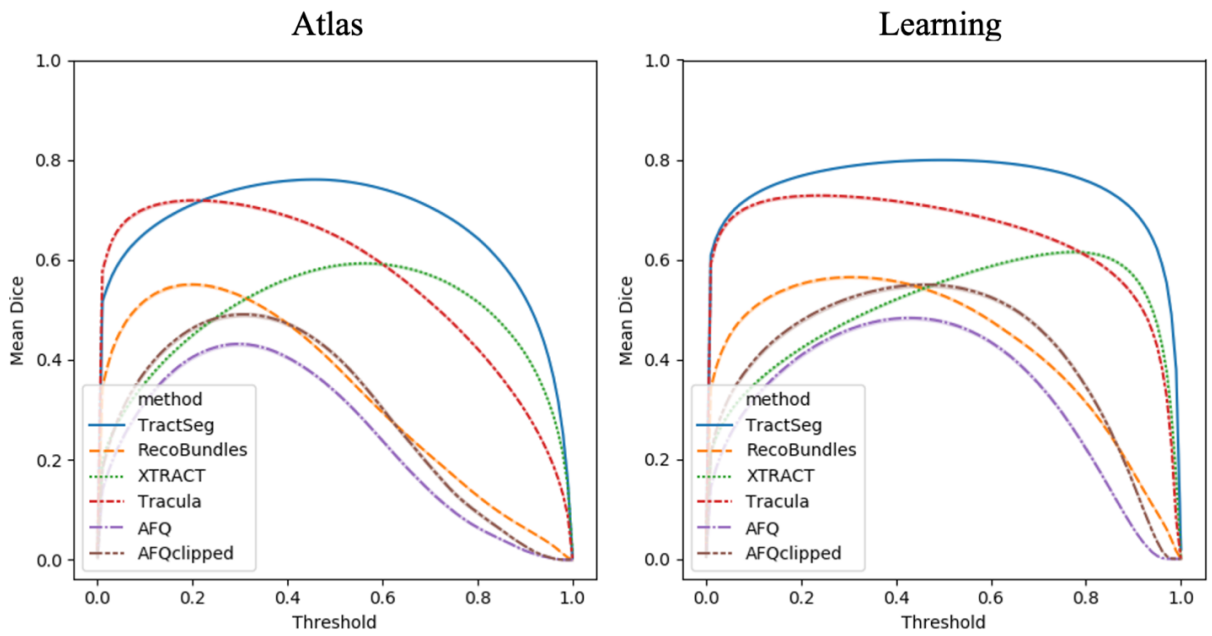


**Figure VIII-8. Each curve represents average DSC of all WM bundles of all external dataset scans per diffusion tractography algorithm for atlas- and learning- based methods.**

**4. Discussion**

In this study, we aim to propose a spatial localized patch-wise framework to segment white matter structure with six different definition schemes only from anatomical image. We envision this framework as a tool to get a coarse WM region of interest rather than segmentation with more details derived from dMRI. The output of the proposed framework is similar to a probability map rather than a binary image, which can provide users with more options to adjust thresholds to adjust bundle overlap or bundle overreach for one specific WM pathway. Different bundle segmentation algorithms do not have exactly the same definition for the same WM tracts. Thus, we provide six definition schemes for users to reconstruct WM in their preferred scheme instead of selecting the best bundle segmentation method through comparison.

As mentioned before, the output threshold is important due to its influence on performance of the segmentation method. When generalizing to the external dataset, we use a previously determined threshold to evaluate performance. The XTRACT is the only method in which the atlas performs statistically better compared with the learning-based method. The estimated thresholds from the validation datasets that are closest to the calculated ones from the external dataset are for TractSeg and RecoBundles for both atlas-based and learning-based methods. For AFQ and AFQclipped, the chosen threshold for the atlas is closer to the true optimal threshold than for the deep

learning method. The optimal threshold of the learning-based method shifts more compared with atlas-based methods since the neural network approach suffers from different scale settings of the raw T1w images.

Currently, the variance of performance of learning-based method is obvious. Inherent definition and way of extracting WM tracts by bundle segmentation methods bring challenges to the proposed learning framework. And then we will fine-tune the optimal hyper-parameters for each learning-based framework to improve the performance in the future.

Apart from the variance in the definition and extraction of WM bundles, we lose information when converting streamlines derived from bundle segmentation algorithm into mask, even with their own built-in function. White matter pathways usually are in the form of streamlines, which can provide connection at the sub-pixel level. However, if we take advantage of the density map to convert streamlines into a mask, we are not aware of those connections anymore. In our current learning pipeline, the original output of TractSeg is in the form of a volume. We haven't made any conversion from streamline to volume and therefore have not incurred any loss of information.

We think this work has potential clinical impact on two kinds of fields. On the one hand, the proposed learning framework can help to map brain tumor to specific WM tract through anatomical images. Previous work has demonstrated that the location of tumor intersecting WM tracts would be associated with differing survival [318]. Thus, knowing specific WM tract in which the brain tumor developed can help to determine patient prognosis. On the other hand, this work might help to map brain lesions to specific white matter regions on anatomical images, which can help to detect modulatory influence on cognitive function when working with functional MRI [319].

## 5. Conclusion

We proposed a spatial localized patch-wise framework to delineate WM structure based on structural T1w images. We use this framework to learn WM regions under six bundle segmentation algorithms and compared the result of the framework to atlas-based methods. When we use the optimal threshold to evaluate scans that have the same acquisition as the training datasets, the learning-based methods are statistically superior to the atlas-based methods.

## Chapter IX. Semi-supervised disentanglement approach to harmonize DW-MRI across single- and multi-shell acquisitions


### Abstract

Diffusion weighted MRI harmonization is necessary for multi-site or multi-acquisition studies. Currently statistical methods address the need to harmonize from one site to another, but do not consider the use of multiple datasets which are comprised of multiple sites, acquisitions protocols, and age demographics. This work explores deep learning methods which can generalize across these variations through semi-supervised and unsupervised learning while also learning to estimate multi-shell data from single-shell data using the MUSHAC and BLSA datasets. We choose to compare disentanglement harmonization models and a CycleGAN harmonization model to the baseline preprocessing and to SHORE interpolation. We find that the disentanglement models achieve superior performance in harmonizing all data while at the same transforming the input data to a single target space across several diffusion metrics (fractional anisotropy, mean diffusivity, mean kurtosis, primary eigenvector).


### 1. Introduction

Diffusion weighted MRI (DW-MRI) is the only non-invasive modality to probe *in vivo* tissue micro-structure and macrostructure [16]. This has opened up new investigations into cognitive neuroscience and brain dysfunction in aging, mental health disorders, and neurological disease [17]. However, clinical adoption is hindered by the variability in DW-MRI measurements caused by differences in the number of head coils, coil sensitivity, imaging gradient non-linearities, magnetic field homogeneity, reconstruction algorithms, and software upgrades [24-28]. These differences are measured in terms of reproducibility across multiple acquisitions and across multiple sites (Figure 1), and the goal of increasing reproducibility is known as harmonization.

Many empirical models have been developed for the purpose of correcting hardware specific effects [29-33], and statistical models have been shown to be effective at harmonizing scalar and vector values [34, 35]. Recently, data driven deep learning approaches have been explored in this arena. Several such methods were proposed for the Multi-shell Diffusion MRI Harmonization Challenge (MUSHAC) which was comprised of different site-to-site harmonization problems [41]. These methods typically choose to learn from spherical harmonic (SH) representations or rotationally invariant features for each shell within the acquisition [42].

These approaches fail to address two issues that limit generalizability. The first of these is the site-to-site

approach requires training data from each site or acquired with each acquisition protocol with matching subjects. Currently, no dataset exists which would cover the possible DW-MRI hardware and acquisition protocols. The second is the use of SH or other single-shell representations to model the diffusion signal. Even when a set of subjects is acquired at multiple sites, the acquisition protocols would need to be comprised of the same diffusion b-values. This work will explore the current models and methods which may overcome these hurdles.

## 2. Related Works

### 2.1. Statistical Models

Through analyzing the effectiveness of several statistical approaches that were developed for other data types, Fortin et al. [35] found that ComBat [2] achieved the best performance. Originally developed for genomics data, ComBat uses an empirical Bayes framework for adjusting data for batch effects that is robust to outliers in small sample sizes. A DTI harmonization technique proposed by Mirzaalian et al. [4] utilizes rotation invariant spherical harmonic (RISH) features and combines the unprocessed DTI images across scanners. A major drawback of these methods is that they require DTI data to have similar acquisition parameters which is often unfeasible in multi-site studies. Although, unlike supervised machine learning methods, acquisitions between sites do not need to be of the same subjects.

### 2.2. Deep Learning Models

Many deep learning approaches have been employed for diffusion harmonization as well. Nath et al. utilized a dual network to incorporate unlabeled paired in-vivo DW-MRI of human subjects along with labeled squirrel monkey DW-MRI with histology ground truth [8]. Koppers et al. designed a residual network specifically for spherical harmonic representations of DW-MRI which predicts the spherical representation at one scanner given the spherical harmonics of another scanner [6]. Given DW-MRI from multiple sources, Moyer et al. uses an unsupervised method based on variational auto-encoders to learn an intermediate representation that is invariant to site and protocol specific effects [7]. Many of these methods are supervised and require matching subjects at all sites, and most of them rely on single-shell representations of the diffusion signal which would limit the models to acquisitions of similar b-values. However, previous work has used neural networks to estimate a second shell of a two shell acquisition given the first shell as input [9]. Figure 2 generalizes the frameworks of these approaches and Table 1 summarizes the features of popular methods.

Various deep learning approaches have been applied to harmonization in other modalities as well. For harmonization between T1 and T2 contrasts, Dewey et al. leverages paired T1 and T2 acquisitions to learn two latent spaces: one which encodes anatomical features and one which encodes acquisition features. The encoder is trained to generate the specified contrast using either sets of anatomical features [320]. CycleGAN has been used to learn style transfer between sites for MRI harmonization as well [321-323]. The cycle consistency loss in this framework ensures anatomical information is retained while the adversarial loss enforces the site-specific changes. In this work we explore the application of these methods for DW-MRI harmonization in a framework that allows for multiple datasets which are not limited by acquisition parameters.

### 2.3. DW-MRI Representations

Both statistical and deep learning approaches to DW-MRI harmonization typically rely on single-shell representations of the signal. SH and RISH represent the data in comparatively few features when considering the number of diffusion volumes acquired in most acquisitions. More importantly, the number of features or coefficients remains constant after choosing the order of the function. However, these representations are still limited by the b-value of the acquisition, so these methods only harmonize multi-site datasets where the b-values are chosen to be the same at all sites. Multi-shell representations can enable multi-site learning across datasets with different b-values. Simple harmonic oscillator based reconstruction and estimation (SHORE) [5, 324] has been shown to generalize diffusion microstructure estimation across multiple b-values [325], and this work will explore the use of SHORE in diffusion harmonization.

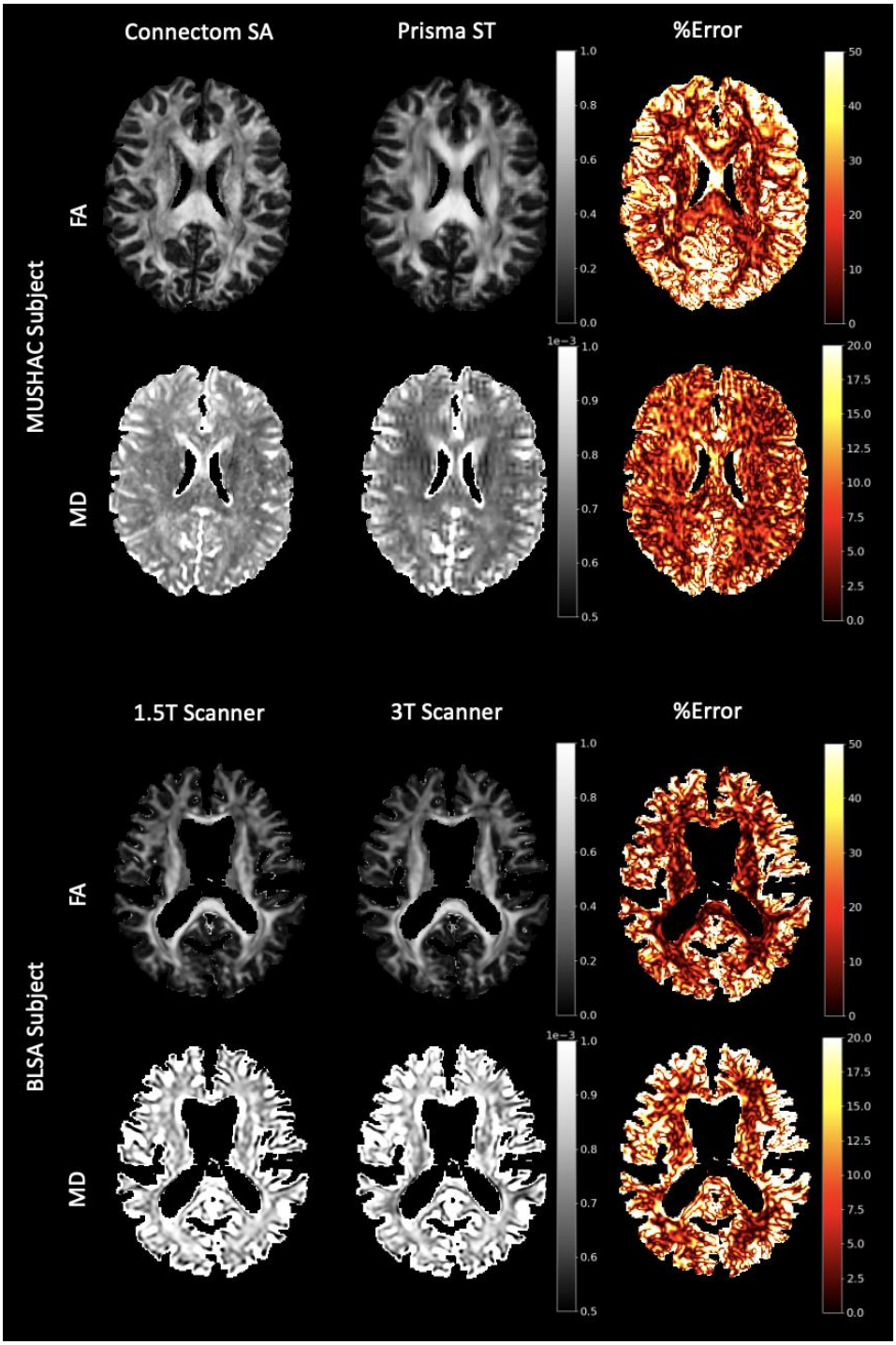


**Figure IX-1. Hardware and protocol differences lead to reproducibility error in DW-MRI metrics. Examples of these differences are shown here for FA and MD for a subject from the MUSHAC dataset (top) as well as the BLSA dataset (bottom). While directly harmonizing between two sites is straightforward, it does not allow for multiple datasets to be jointly analyzed.**

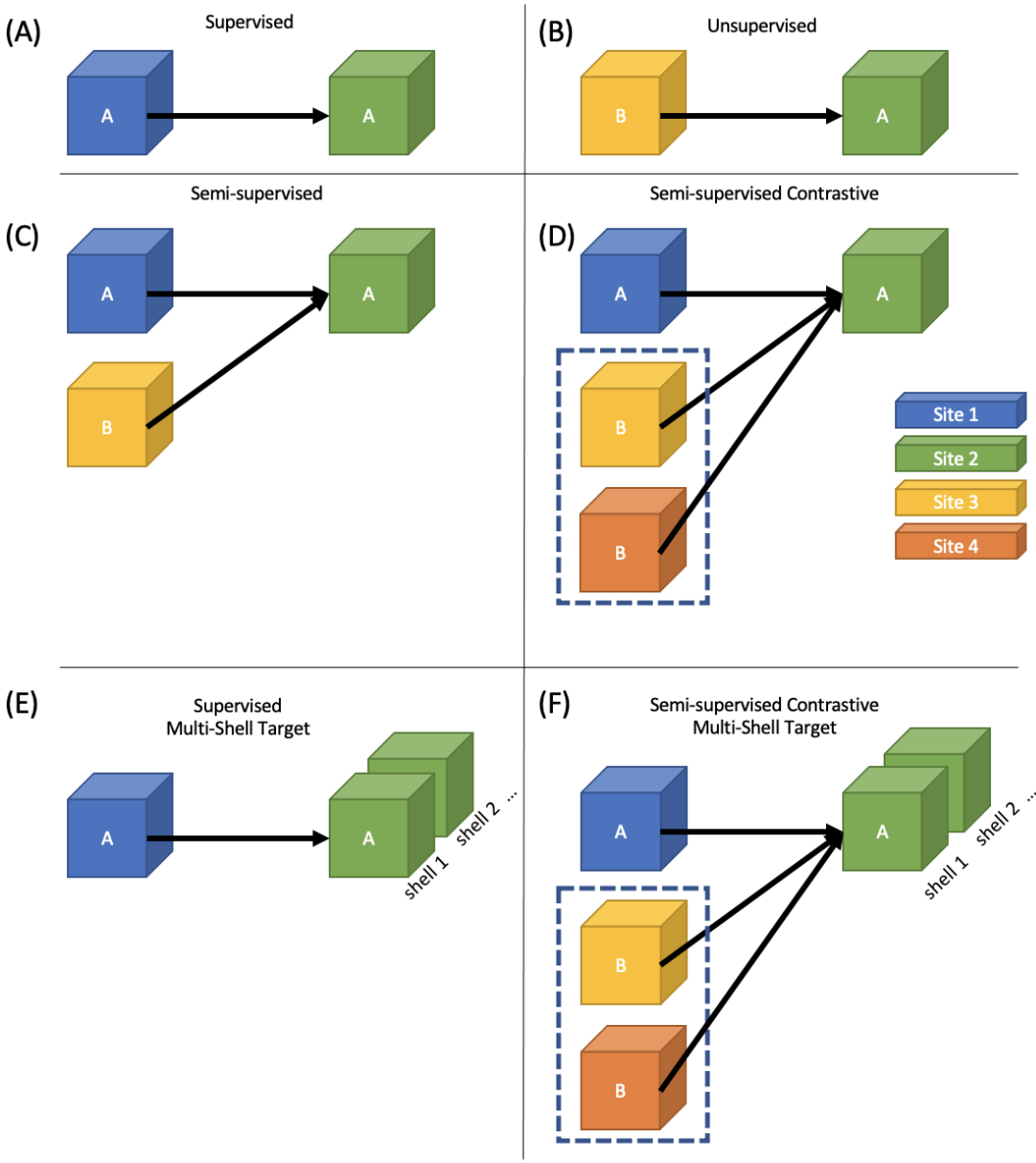


**Figure IX-2. Machine learning approaches in DW-MRI follow the general format of supervised (A) and unsupervised (B). However, there are few approaches which follow the standard semi-supervised approach (C), but a contrastive approach which relies on having paired data across sites or acquisitions (D) has been shown to be effective. A problem more unique to DW-MRI is estimating a multi-shell acquisition from a single-shell acquisition (E). This work focuses on estimating a multi-shell target site from single-shell data in a semi-supervised contrastive learning framework (F).**

**Table IX-1. Statistical methods as well as deep learning methods all depend on b-value specific representations of DW-MRI. SHORE is a multi-shell representation that is not dependent on the b-value and can be used to reconstruct any given acquisition scheme given a set of b-values and directions. We aim to leverage this to create a deep learning framework which could harmonize across datasets without needing to match acquisition parameters across sites.**

| | Regression | Convolutional Neural Net | Supervised | Unsupervised | Semi-supervised | Semi-supervised Contrastive | Multi-shell Target |
|---|---|---|---|---|---|---|---|
| COMBAT [2] | ✓ | X | ✓ | X | X | X | X |
| RISH [4] | ✓ | X | ✓ | X | X | X | X |
| SHORE [5] | ✓ | X | X | X | X | X | X |
| SHResNet [6] | X | ✓ | ✓ | X | X | X | X |
| StarGAN [7] | X | ✓ | X | ✓ | X | X | X |
| NST [8] | X | ✓ | X | X | X | ✓ | X |
| ShellDNN [9] | X | ✓ | ✓ | X | X | X | ✓ |
| This Work | X | ✓ | X | X | X | ✓ | ✓ |

## 3. Methods

### 3.1. Data

The MUSHAC dataset consists of 15 subjects each scanned at two scanners with two different sets of acquisition parameters. The scanners were a 3T Siemens Prisma (80 mT/m) and a 3T Siemens Connectom (300 mT/m) model. A full list of acquisition parameters is provided in Table 2. The two acquisitions at each scanner were designed to be one standard acquisition (ST) and one state-of-the-art acquisition (SA). All acquisitions were acquired with b-values of 1200 and 3000 s/mm$^2$, and the most notable differences between ST and SA are an increase from 30 to 60 directions per b-value and an increase in resolution from a voxel size of 2.4mm isotropic to 1.5mm isotropic in the case of the Prisma scanner and 2.4mm isotropic to 1.2mm isotropic in the case of the Connectom scanner [43].

The BLSA dataset consists of 50 subjects scanned at four scanners: General Electric (GE) Signa 1.5T (A), Philips Ahieva 3T (B), (C), and (D). Every subject was not scanned at all four scanners, but each subject used was

scanned at the 1.5T scanner and one of the 3T scanners. The acquisition parameters have small differences which are provided in Table 3. Unlike the MUSHAC dataset where the average time between acquisitions on scanners was within 2 years, there could be many years between acquisitions in the BLSA data.

**Table IX-2. The MUSHAC dataset consists of 14 subjects across two sites each with two sets of acquisition parameters. For each site, there is a standard (ST) and a state-of-the-art (SA) acquisition where the most noticeable difference is the voxel resolution and the number of directions per b-value.**

| **Scanner (MUSHAC)** | **Siemens 80 mT/m (Prisma)** | | **Siemens 300 mT/m (Connectom)** | |
|---|---|---|---|---|
| **Protocol** | **Standard (ST)** | **State-of-the-art (SA)** | **Standard (ST)** | **State-of-the-art (SA)** |
| **Diffusion weighted images** | | | | |
| Sequence | PGSE | PGSE | PGSE | PGSE |
| b-values [s/mm2] | 1200, 3000 | 1200, 3000 | 1200, 3000 | 1200, 3000 |
| # directions per b-value | 30 | 60 | 30 | 60 |
| TE [ms] | 89 | 80 | 89 | 68 |
| TR [ms] | 7200 | 4500 | 7200 | 5400 |
| Δ/δ [ms] | 41.4/26.0 | 38.3/19.5 | 41.8/28.5 | 31.1/8.5 |
| Phase encoding direction | AP | AP | AP | AP |
| Reconstructed voxel size | 1.8 × 1.8 × 2.4 | 1.5 × 1.5 × 1.5 | 1.8 × 1.8 × 2.4 | 1.2 × 1.2 × 1.2 |
| Matrix size | 96 × 96 | 154 × 154 | 96 × 96 | 180 × 180 |
| # slices | 60 | 84 | 60 | 90a |
| Head coil | 32 channel | 32 channel | 32 channel | 32 channel |
| **b0 images** | | | | |
| TE [ms] | 89, 80, 89 | 80, 80, 89 | 89, 68, 89 | 68, 68, 89 |
| TR [ms] | 7200, 7200, 13000 | 4500, 7200, 7200 | 7200, 7200, 13000 | 5400, 7200, 7200 |
| Phase encoding direction | AP, PA | AP, PA | AP, PA | AP, PA |

DW-MRI from both datasets are preprocessed using standard techniques including EPI distortion correction using FSL TOPUP, and eddy current distortion correction using FSL eddy [134, 326]. Using a b0 image, the DW-MRI are registered to a T1 of the subject suing FSL epi_reg, and then the T1 and the DW-MRI are registered to the MNI152 template using FSL flirt [326]. The template image has a voxel resolution of 1mm isotropic and the volume

dimensions are 193×223×193. Anatomical segmentations as defined by BRAINCOLOR [123] are generated using SLANT [309].

**Table IX-3. The chosen 50 subjects from the BLSA dataset are acquired across four scanners. All subjects have at least a one scan on the 1.5T scanner (A) and at least one scan at one or more of the 3T scanners (B, C, D). The number of directions per b-value are spread across two scans acquired in a single session. There are small differences between acquisitions, but the parameters were not intentionally chosen such that there were differences between scanners.**

| **Scanner (BLSA)** | **A (1.5T)** | **B (3T)** | **C (3T)** | **D (3T)** |
|---|---|---|---|---|
| **Diffusion weighted images** | | | | |
| Sequence | PGSE | PGSE | PGSE | PGSE |
| b-values [s/mm2] | 700 | 700 | 700 | 700 |
| # directions per b-value | 30 | 32 | 32 | 32 |
| TE [ms] | 80 | 75 | 75 | 75 |
| TR [ms] | 6210 | 6801 | 6801 | 7454 |
| Δ/δ [ms] | 39.2/15.1 | 36.3/16 | 36.3/16 | 36.3/13.5 |
| Phase encoding direction | APP | APP | APP | APP |
| Reconstructed voxel size | 0.94 × 0.94 × 2.5 | 0.83 × 0.83 × 2.2 | 0.83 × 0.83 × 2.2 | 0.81 × 0.81 × 2.2 |
| Matrix size | 96 × 96 | 96 × 95 | 96 × 95 | 116 × 115 |
| Reconstruction matrix size | 256 × 256 | 256 × 256 | 256 × 256 | 320 × 320 |
| # slices | 50 | 65 | 65 | 70 |
| Head coil | Philips 8-ch | Philips 8-ch | Philips 8-ch | Philips 8-ch |
| **b0 images** | | | | |
| TE [ms] | 80 | 75 | 75 | 75 |
| TR [ms] | 6210 | 6801 | 6801 | 7454 |
| Phase encoding direction | APP | APP | APP | APP |

For the purposes of this work, we select the Connectom state-of-the-art acquisition within the MUSHAC dataset as the target site. We utilize the MUSHAC data as labeled data where each target has three distinct inputs: Prisma ST, Prisma SA, and Connectom ST. The BLSA dataset is used as unlabeled data. Five subjects from each dataset are withheld for testing. Our baseline method is simply taking the data as is and calculating the DW-MRI metrics. The goal of each method is to harmonize both datasets by removing site specific effects and biases and adding features specific to the target site.

### 3.2. SHORE

SHORE has been shown to capture multi-shell DW-MRI with minimal reconstruction error [5] while ensuring the same when modelling single-shell DW-MRI. The normalized DW-MRI signal can be represented as:

$$E(q) = \sum_{n=0}^{N} \sum_{l=0}^{n} \sum_{m=-l}^{l} c_{nlm} G_{nl}(q, \zeta)\, Y_l^m(u) \quad (1)$$

where $c$ are the coefficients, $G$ is the radial basis, and $Y$ is the SH basis. The radial basis $G$ is expressed as:

$$G_{nl}(q, \zeta) = K_{nl} \left(\frac{q^2}{\zeta}\right)^{\frac{l}{2}} \exp\left(-\frac{q^2}{\zeta}\right) L_{n-\frac{l}{2}}^{l+\frac{1}{2}}\left(\frac{q^2}{\zeta}\right) \quad (2)$$

where $\zeta$ is the scale parameter, $q$ is the radius of the diffusivity value, and $L$ is the associated Laguerre polynomial. Here we use the default parameters of shore as recommended by DIPY [285], so SHORE is estimated at 6$^{th}$ order, $\zeta$ is et as 700, and regularization constants are set as $1\mathrm{e}-8$. This results in 50 estimated coefficients. However, though SHORE achieves minimal reconstruction error in both single-shell and multi-shell estimation, it can not reconstruct multi-shell data from coefficients modelled with single-shell data. Therefore, the input data in the MUSHAC dataset are only modelled using the b-value 1200 s/mm$^2$ shell.

### 3.3. Disentanglement Model

We repurpose the model designed by Dewey et al. to harmonize between sites rather than between contrasts (Figure 3). This method consists of learning two things: the disentanglement between acquisition specific and anatomical specific features and the transformation to the target acquisition. Because cross-site same subject pairs exist within the input data, we can use a pair of scans from different scanners or acquisitions to learn the disentanglement, and we can use the labeled MUSHAC data to learn the transformation from acquisition free latent space. The model is comprised of an anatomical encoder $E_{anat}$, an acquisition encoder $E_{acq}$, an acquisition decoder $D_{acq}$, and a target decoder $D_{targ}$. The architectures of $E_{anat}$, $D_{acq}$, and $D_{targ}$ are modified 3D U-Nets [95, 307] which do not downsample the spatial dimensions of the input. The architecture of $E_{acq}$ is a 3D convolutional neural network which encodes the input into a 1×256 vector that contains acquisition specific features. The architectures are modified for $32\times 32 \times 32$ patches as well as 193×223×3 slabs of axial slices. The specifics of these architectures are shown in Figure 4.

For each step in training, a volume from one of the three input sites of the MUSHAC dataset $x_i$ as well as a pair of sites from either the BLSA or MUSHAC $[x_j, x_k]$ are selected. $E_{anat}$, $E_{acq}$, and $D_{acq}$ are trained using the paired data $[x_j, x_k]$in a similar fashion to Dewey et al. The SHORE coefficients of the input are fed to $E_{anat}$ and $E_{acq}$ for each $x_j$ and $x_k$ resulting in subject features $\beta_j$ and $\beta_k$ and acquisition features $\theta_j$ and $\theta_k$. For each $\beta$ feature map, the feature map is randomly taken from $\beta_j$ or $\beta_k$ to form $\beta_{jk}$. This encourages the model to represent subject features the same across acquisition. $D_{acq}$ is then given the pairs of $[\beta_{jk}, \theta_j]$ and $[\beta_{jk}, \theta_k]$ with the goal of reconstructing the acquisition specified by $\theta$. $D_{targ}$ is given only $\beta_i$ with the goal of generating the associated target image $y_i$. $E_{anat}$, $D_{acq}$, and $E_{acq}$ are trained using the paired data $[x_j, x_k]$, while $D_{targ}$ is trained separately using the labeled data $[x_i, y_i]$. The loss functions employ L1 loss ($L1$), structural similarity index measure loss ($SSIM$), total variation loss ($TV$), and Sobel edge detection ($Sobel$):

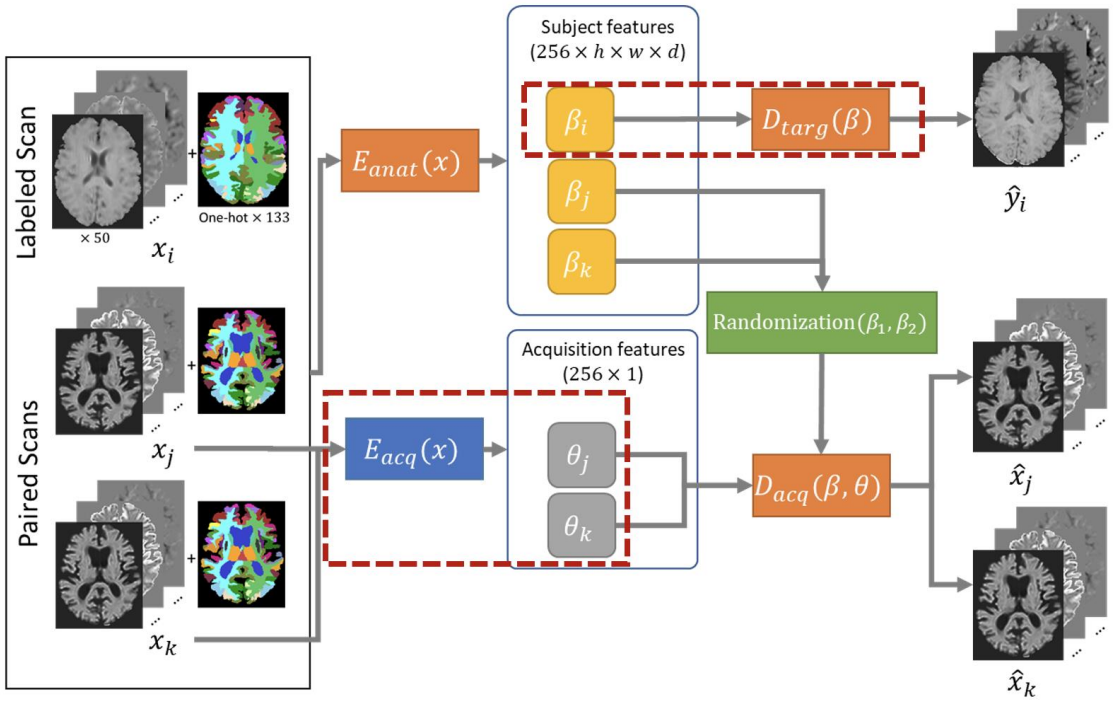


**Figure IX-3. We follow the work of Dewey et al., but where before the goal was to harmonize between T1 and T2 acquisitions, our goal is to harmonize between many DW-MRI acquisitions as well move all data to a single target space. Changes to the method are indicated by red boxes. To account for the much broader range of acquisition possibilities, we use an acquisition encoder which represents the acquisition using a vector of size 256 rather than a single value which only needed to indicate contrast. In a similar manner, we use paired subject data from different acquisitions and encourage the network to encode a latent space which represents only the subject specific feature free from scanner or acquisition bias, and then reconstruct the acquisition indicated by the acquisition encoding vector using subject features from either scan. A second decoder was added to learn from the acquisition free latent space to a target space using the supervised data.**

$$L_{image}(x, \hat{y}, y) = SSIM(\hat{y}, y) + L1(\hat{y}, y) + TV(\hat{y}) + L1(Sobel(x), Sobel(\hat{y})) \quad (3)$$

where the $SSIM$ and $L1$ terms encourage the predicted $\hat{y}$ to have the same information as $y$, the $TV$ term regularizes neighborhood consistency within $\hat{y}$, and another $L1$ term between the edge features of the input $x$ and the prediction $\hat{y}$ enforces that the structure remains the same. For $E_{anat}$, $D_{acq}$, and $E_{acq}$, the loss function is:

$$L = L_{image}(x_j, \hat{x}_j, x_j) + L_{image}(x_k, \hat{x}_k, x_k) + \lambda_E L1(\beta_j, \beta_k) \quad (4)$$

where the final term is a contrastive term which further encourages the structural features to be similar across acquisitions and $\lambda_E$ controls the contribution of this term and is empirically set to 10. The loss for $D_{targ}$ is:

$$L = L_{image}(x_i, \hat{y}_i, y_i) \quad (5)$$

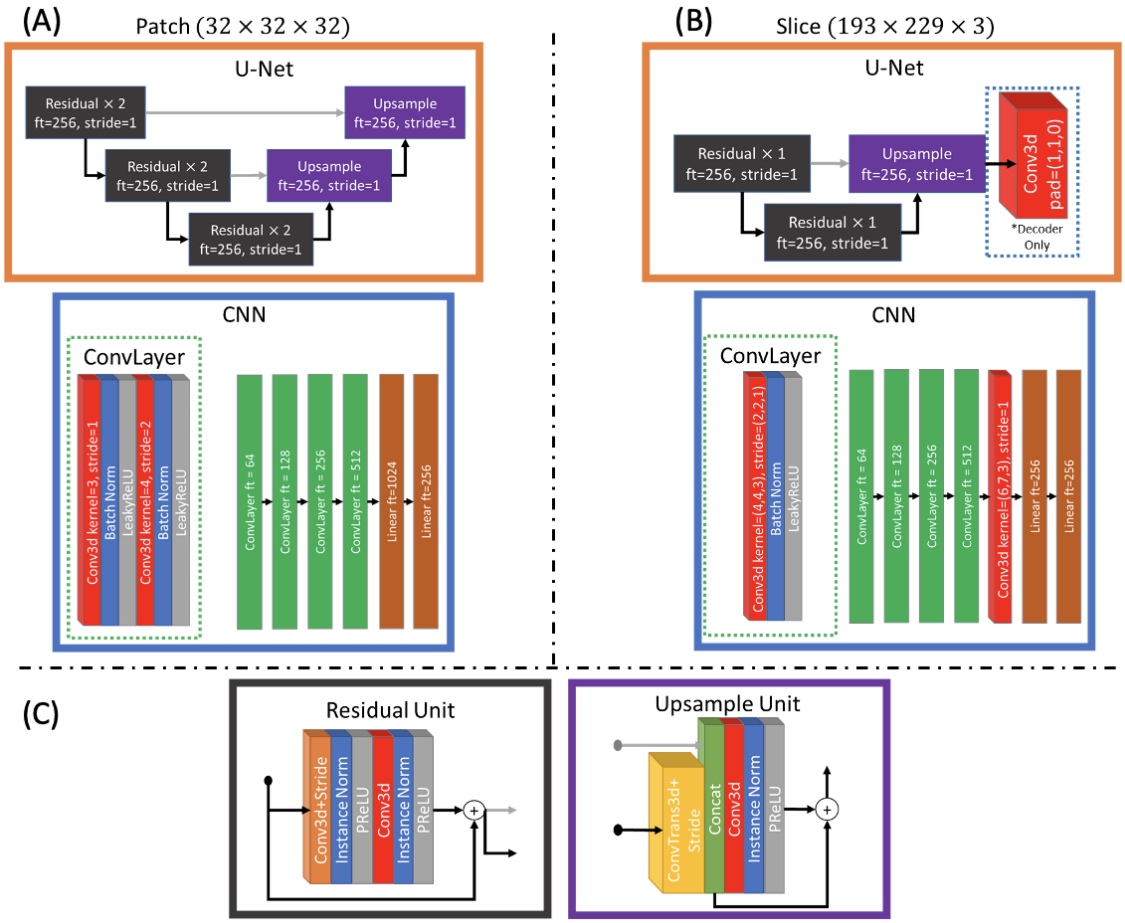


**Figure IX-4. The architectures used in the disentanglement model are modified for 32 by 32 by 32 patches (A) as well as 193 by 229 by 3 axial slabs (B). The acquisition encoder is defined by a CNN which results in a vector of size 256 while the structural encoder and the two decoders are defined by U-Nets which preserve the original size of the input. The U-Nets use the same residual and upsample units (C).**

The implementation of this model requires two optimizers, one responsible for each of these losses. We choose each of these to be an Adam optimizer with a learning rate of 1e-5. Each model was trained until convergence on a validation set which was approximately 75 epochs each consisting of approximately 1500 samples of patches or axial slabs. Where the patch-based model was evaluated in 3D, the slice-based model was evaluated only on the middle of the three axial slices.

### 3.4. CycleGAN Model

As a baseline representing unsupervised learning approaches, we modify the CycleGAN model according to Bashyam et al. (Figure 5). This involves adding an encoder ($E_{acq}$) which encodes the acquisition specific features of images from the input domain to compensate for potentially many different styles coming from many different acquisitions. Similar to the disentanglement model, the generator which goes from the target domain to the input domain ($G_B$) is parameterized by a target domain image and the acquisition features which indicate the specific

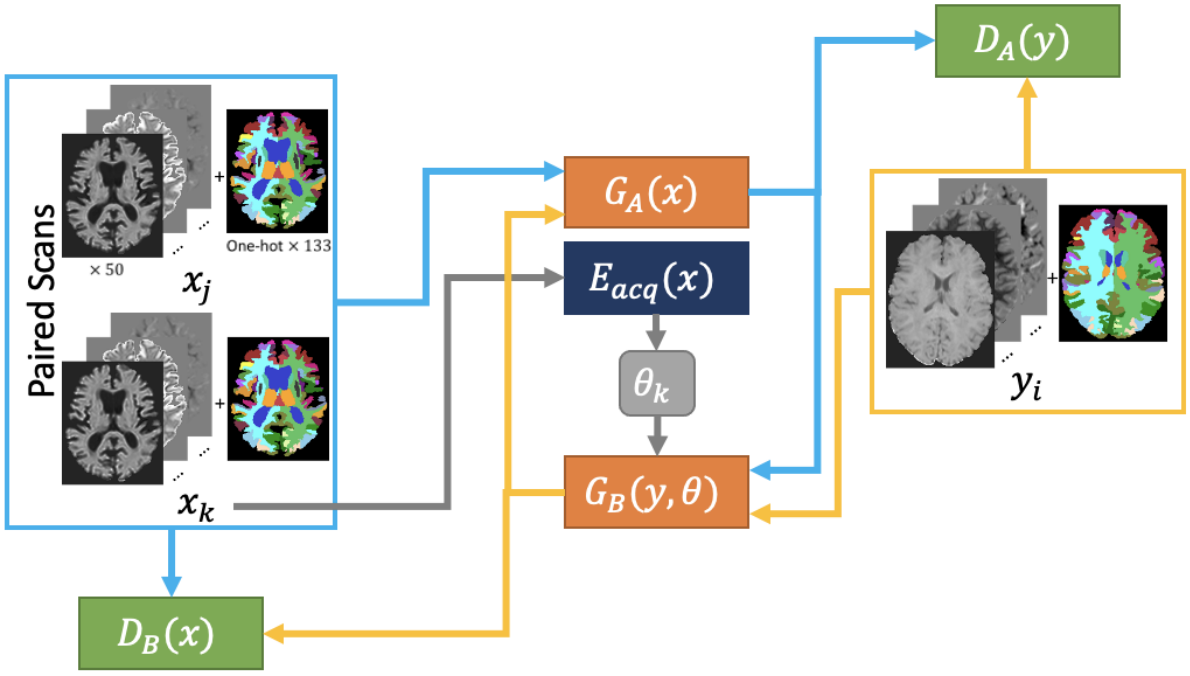


**Figure IX-5. As a baseline, a CycleGAN framework is constructed from two U-Net generators, one which takes an axial slice of SHORE coefficients and one hot encoded SLANT segmentation from the input domain and generates the target domain and vice versa, as well as two patch discriminators, one which tries to classify whether or not the input is from the input domain and one which does the same for the target domain. Due to the input domain being composed of multiple sites and acquisitions, an autoencoder is used to extract acquisition specific information $\theta$ from the input image which is then used as input when trying to generate an input domain image to specify what scanner or acquisition the generated image should resemble.**

acquisition that should be generated. Also, in following Bashyam et al., we pass our input as axial slices rather than slabs or patches. To avoid putting this model at a disadvantage, we also modify the loss to account for paired data which may provide useful information:

$$L_G = L2\left(D_A\left(G_A(x_j)\right),1\right) + \lambda_A L1\left(G_B\left(G_A(x_j),E_{acq}(x_j)\right),x_j\right) + L2\left(D_B\left(G_B\left(y_i,E_{acq}(x_j)\right)\right),1\right) + \lambda_B L1\left(G_A\left(G_B\left(y_i,E_{acq}(x_j)\right)\right),y_i\right) + \lambda_E L1\left(G_B\left(G_A(x_k),E_{acq}(x_j)\right),x_j\right) \quad (6)$$

$$L_D = L2\left(D_A\left(G_A(x_j)\right),0\right) + L2(D_A(y_i),1) + L2\left(D_B\left(G_B\left(y_i,E_{acq}(x_j)\right)\right),0\right) + L2(D_B(x_j),1) \quad (7)$$

where $L_G$ is the generator loss, $L_D$ is the discriminator loss, $G_A$ is the generator which is parameterized by the input domain and generates the target domain, $D_A$ is the discriminator which classifies between real and fake target domain

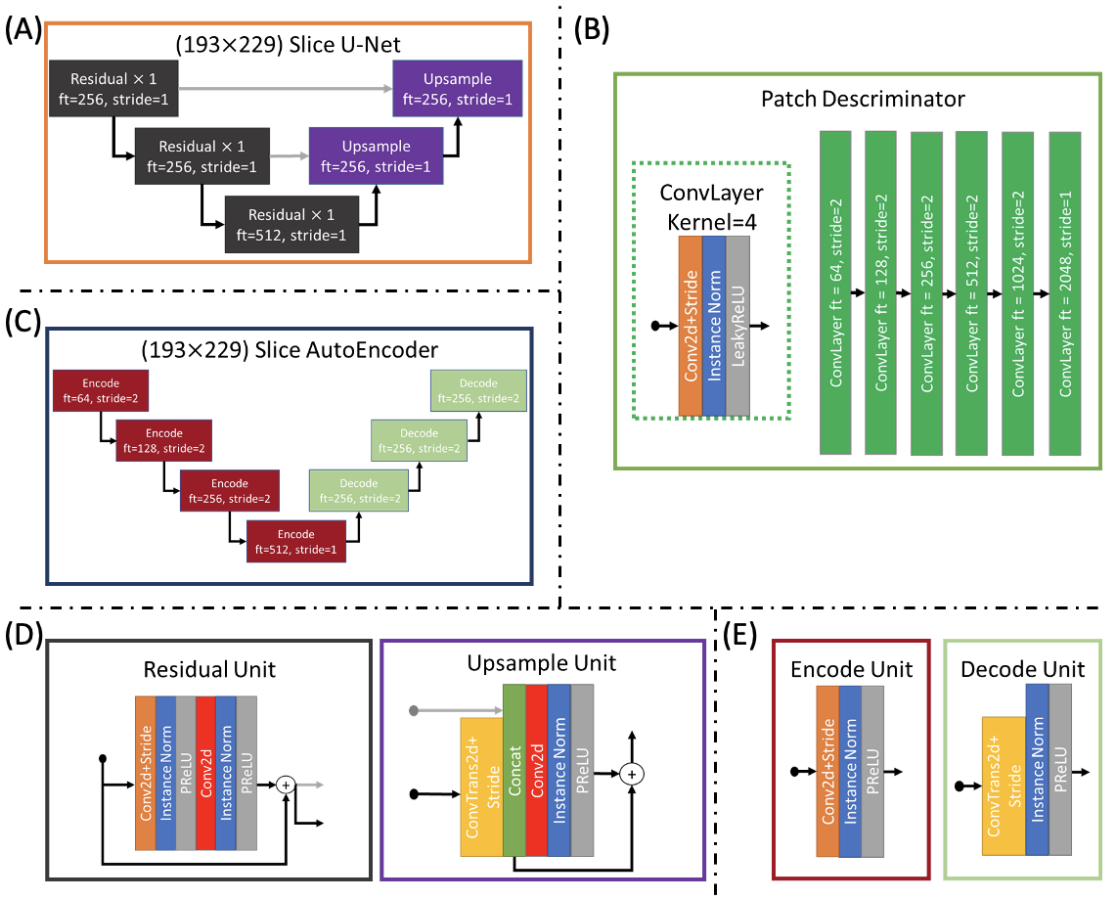


**Figure IX-6. All architectures for the CycleGAN method are designed for 193 by 229 axial slices. The generators are defined as U-Nets (A), the discriminators are defined as patch discriminators (B), and the acquisition encoder is defined as an autoencoder (C). The residual and upsample units for the U-Net are similar to those used in the disentanglement model (D), and the autoencoder uses similar units which lack the skip connection (E).**

images, and $D_B$ is the discriminator which classifies between real and fake input domain images. $\lambda_A$, $\lambda_B$, and $\lambda_E$ are hyperparameters for the cycle loss terms and are empirically set to 25. The final term of the generator loss is similar to the cycle loss except the goal is to recreate $x_j$ using $x_k$ to parameterize $G_A$ to leverage the paired data. Training consisted of 200 epochs where during the first 100 epoch a learning rate of 2e-5 was used and during the last 100 epochs the learning rate was linearly to zero. The architectures used are shown in Figure 6.

**4. Results**

We evaluate each method in terms of RMSE for the metrics fractional anisotropy (FA), mean diffusivity (MD), mean kurtosis (MK), and the angular error of the primary eigenvector (PE) of the diffusion tensor (Figure 7 & Table 4). For the MUSHAC dataset four subjects were withheld for testing and each of the input from the three acquisitions PrismaSA, PrismaST, and ConnectomST are evaluated by their similarity to the target acquisition ConnectomSA. For the BLSA dataset five subjects were withheld for testing and each scan from the 1.5T scanner (A) and the 3T scanners (B, C, or D) are evaluated by their similarity to an average which is calculated for each method. The Wilcoxon signed-rank test was used to test statistical significance of each method (p-value<0.01).

In the MUSHAC data the baseline and SHORE methods are generally similar with some improvement over the baseline in MK and angular error. The disentanglement methods outperform all other methods on average, however, the 25$^{th}$ percentile is ~0.010 higher than the SHORE and baseline methods in white matter FA error. Because the PrismaSA acquisition initially is much closer to the target than the PrismaST or ConnectomST acquisitions, the model sacrifices some ability to make a small adjustment in data that is already close to the target in favor of generalizing to data that is further from the target. Additionally, without the second diffusion shell, the model cannot rely on the identity transform to achieve similar results as the baseline. A visualization of the error reveals the differences in the Patch and Slice method (Figure 8). While the Patch method has a small advantage in gray matter regions, the Slice method outperforms all other methods in white matter, and this is especially evident in FA. The CycleGAN method performs poorly at this task, and it can be seen where the model fails to generate the correct anatomy in the sample subject.

In the BLSA data, the difference between methods is less distinct, but similar trends appear (Figure 9). Although, the SHORE method has much different behavior due to only being fit with a single-shell. The large increase in reproducibility error in MD in both gray and white matter and in FA in gray matter suggest that the method is not particularly stable for single-shell data. Additionally, there is no baseline or SHORE method for MK due the data

only being acquired with a single-shell. Again, the disentanglement methods obtain the lowest error for all metrics. However, the angular error for the slice method is greater than the Patch and CycleGAN methods. Visually it can be seen again that the Slice method has lower error in white matter and that the CycleGAN method is inconsistent.

As an ablation, we also train the Patch method with data augmented by random Gaussian noise to test the stability of the model (Patch Noisy) and without the anatomical parcellation as a prior (Patch w/o SLANT) to assess the contribution of the T1 derived information (Figure 10). We find that the model performance tends to be similar with and without random Gaussian noise suggesting that the model is robust to these small perturbations. We also find that the model generally achieves lower error when the anatomical parcellation is provided as a prior.

**Table IX-4. The mean and standard deviation of the RMSE across scans is reported for each dataset in the white matter and gray matter. The lowest RMSE across FA, MD, MK, and Angular error is achieved by the Patch or Slice disentanglement method.**

| | Gray Matter | | | | White Matter | | | |
|---|---|---|---|---|---|---|---|---|
| Method | FA RMSE | MD RMSE (1e-05) | MK RMSE | Angular RMSE | FA RMSE | MD RMSE (1e-05) | MK RMSE | Angular RMSE |
| **MUSHAC** | | | | | | | | |
| Baseline | 0.087±0.017 | 21.88±4.91 | 0.25±0.04 | 39.02±2.95 | 0.099±0.029 | 9.05±2.15 | 0.17±0.04 | 23.79±4.28 |
| SHORE | 0.087±0.017 | 21.78±4.95 | 0.25±0.04 | 38.42±3.15 | 0.099±0.029 | 9.02±2.17 | 0.16±0.04 | 22.48±4.27 |
| Patch | 0.073±0.012 | **15.78±1.84** | **0.19±0.03** | **33.38±1.56** | 0.084±0.011 | 6.90±0.42 | **0.10±0.01** | **20.42±2.92** |
| Slice | **0.073±0.011** | 16.57±1.85 | **0.19±0.03** | 34.08±1.86 | **0.081±0.010** | **6.66±0.39** | 0.11±0.01 | 21.52±3.14 |
| CycleGAN | 0.107±0.006 | 31.51±2.28 | 0.31±0.03 | 43.76±1.59 | 0.126±0.013 | 11.79±0.59 | 0.29±0.03 | 26.77±2.81 |
| **BLSA** | | | | | | | | |
| Baseline | 0.044±0.010 | 22.39±6.75 | NaN±NaN | 43.39±3.97 | 0.053±0.012 | 10.85±3.02 | NaN±NaN | 37.90±7.66 |
| SHORE | 0.070±0.012 | 91.8±24.63 | NaN±NaN | 42.93±3.81 | 0.052±0.011 | 35.19±9.77 | NaN±NaN | 37.91±7.57 |
| Patch | **0.033±0.008** | 12.59±3.57 | **0.05±0.01** | **37.21±4.55** | **0.046±0.010** | **3.48±0.97** | **0.04±0.01** | **33.24±6.76** |
| Slice | 0.039±0.009 | **12.32±3.12** | 0.06±0.01 | 39.49±4.27 | 0.049±0.011 | 3.65±0.93 | **0.04±0.01** | 36.7±7.63 |
| CycleGAN | 0.057±0.009 | 16.10±3.67 | 0.14±0.02 | 37.73±3.82 | 0.066±0.012 | 8.72±1.60 | 0.13±0.02 | 36.15±6.10 |

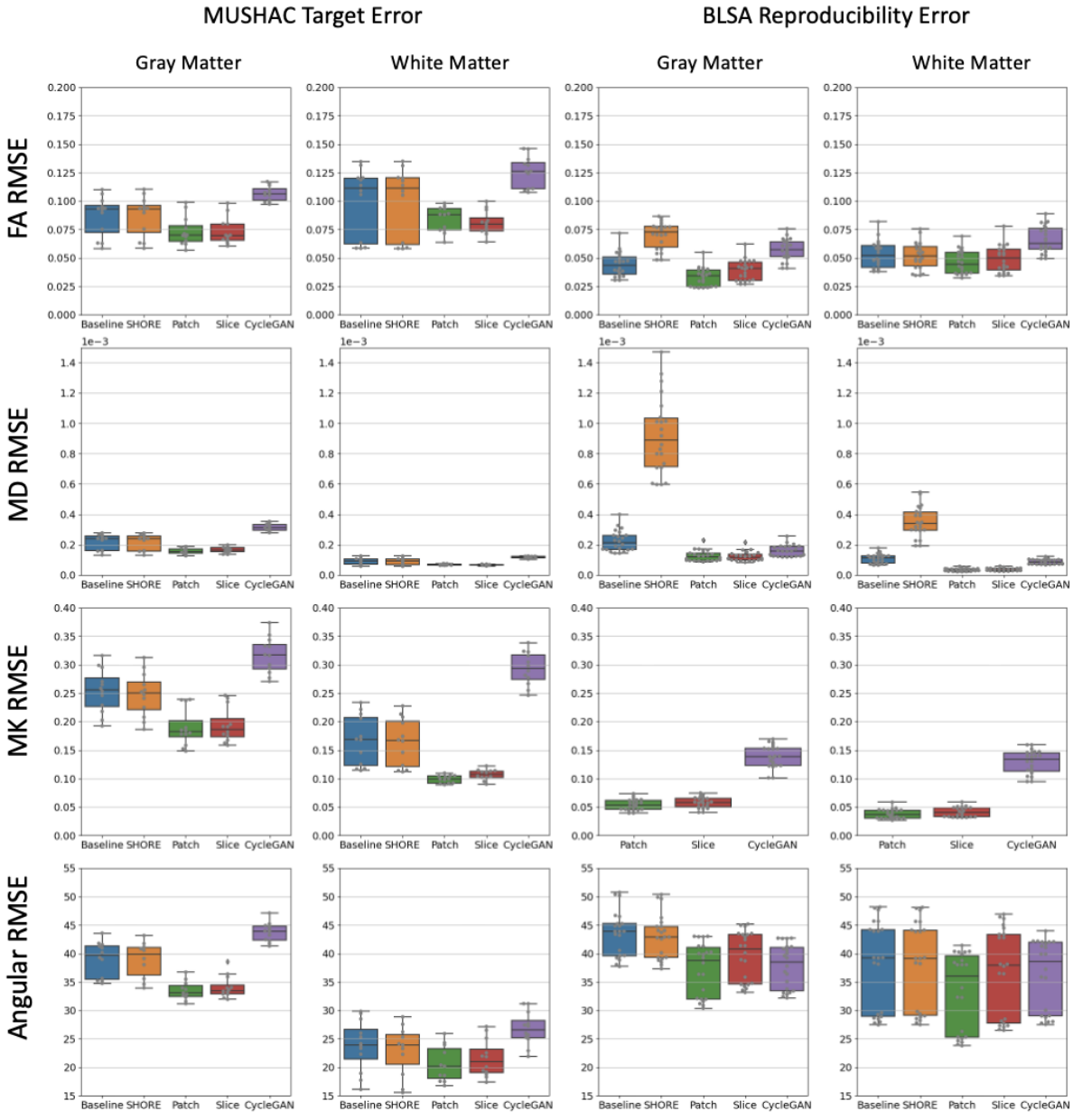


**Figure IX-7. Here the methods are compared in terms of RMSE of FA, MD, MK, and angular error for each input scan. The baseline and SHORE methods use all available shells while all other methods are given on the first shell of a lower b-value. On average, the Patch and Slice disentanglement models perform better in white and gray matter for both datasets across metrics. Notably the improvement in MK indicates the estimation of the second shell is successful. Wilcoxon signed-rank test shows that all methods are statistically significant (p-value<0.01).**

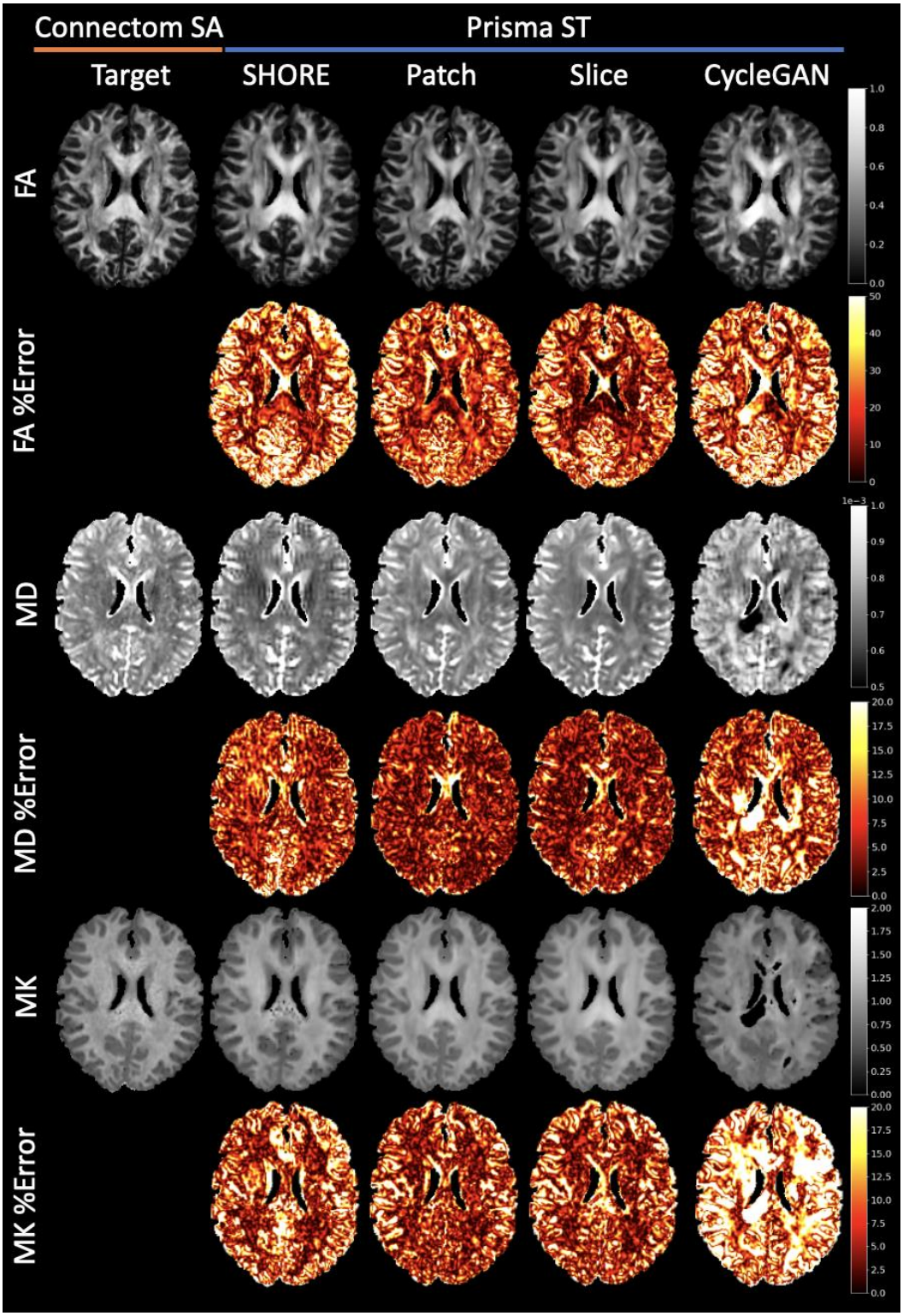


**Figure IX-8. For a single MUSHAC subject an axial slice of FA, MD, and MK and the precent error is shown for each method excluding the baseline. For the disentanglement methods, the error generally improves in both gray and white matter. However, the Slice method shows greater error reduction in white matter.**

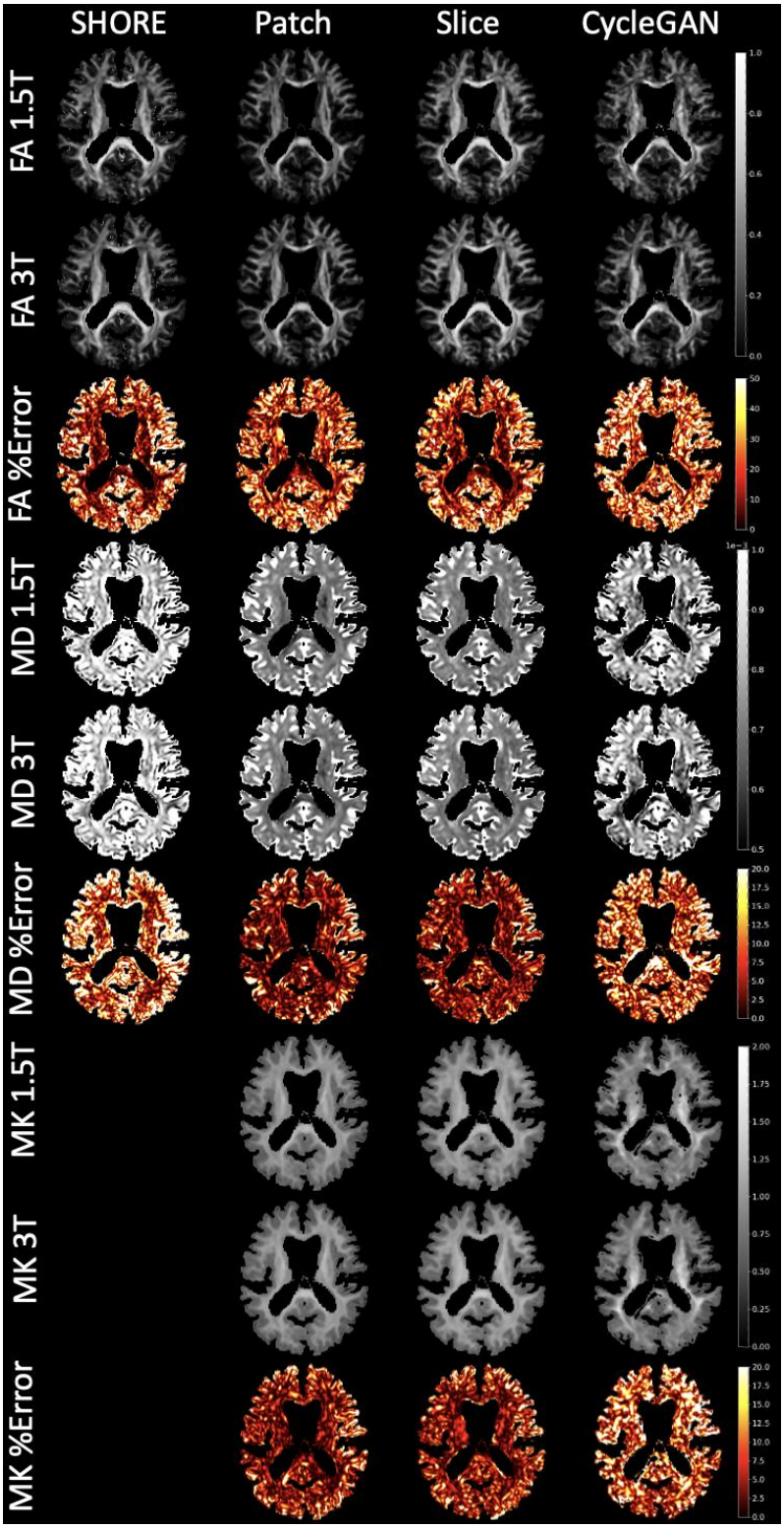


**Figure IX-9. Here we look at the reproducibility error for each method for a BLSA subject using a scan acquired at the 1.5T scanner (A) and a scan acquired at a 3T scanner (B). Here the difference between the Patch and Slice Disentanglement models is clear in FA where the error in white matter is much lower for the Slice method.**

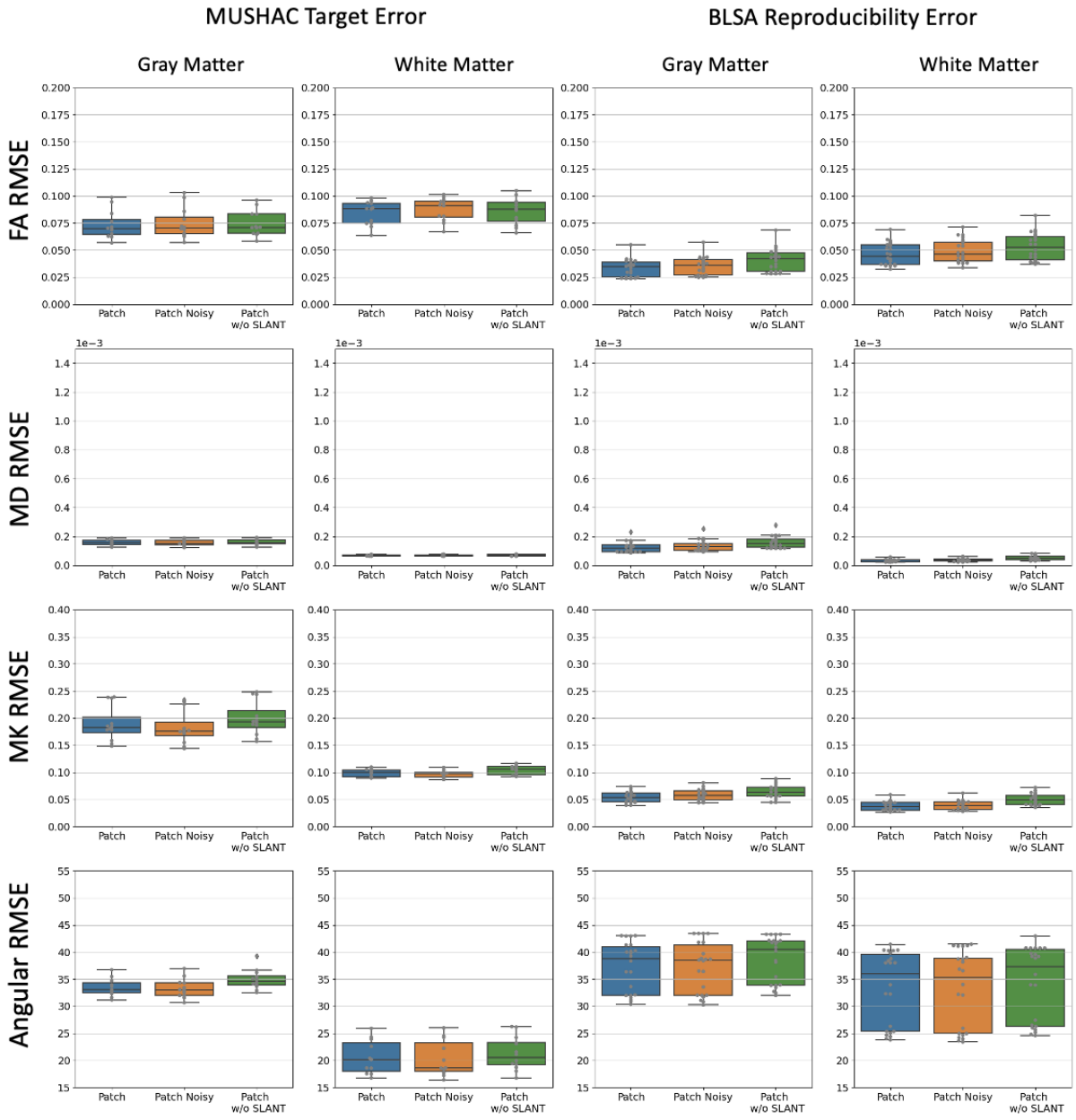


**Figure IX-10. We modify the Patch disentanglement model to 1) test the model's robustness when trained with data augmented with Gaussian noise and 2) test the model's response to removing the anatomical segmentation priors. While the model seems to have a small response to adding noise, removing the anatomical priors generally decreases performance. Wilcoxon signed-rank test shows that all methods are statistically significant (p-value<0.01).**

## 5. Discussion

The chosen datasets are each unique and present different aspects of harmonization. The MUSHAC dataset was specifically designed to have two very different acquisition protocols which tends to be the main contribution of reproducibility error compared to the bias introduced by differences in hardware. The BLSA dataset was not intended to have different acquisition parameters, but throughout the course of the study, there were inevitably replacements made resulting in four different hardware and slightly altered acquisition parameters. The aging aspect of the data introduces reproducibility error resulting from actual changes in anatomy rather than scanner or acquisition bias which should be preserved rather than removed or altered. In trying to harmonize all these data to a single target space includes dealing with the differences in b-value, the number of shells, and the anatomical differences in an aging cohort and a young adult cohort.

By choosing a method that represents the diffusion signal as the same set of coefficients regardless of the acquisition, the simplified input space can potentially be dealt with a single model. However, the differences between the estimated SHORE coefficients from single-shell data and multi-shell data creates two distinct tasks for the model to learn. This becomes particularly difficult when the unlabeled set of data only contains single-shell representations, and the labeled data only contains multi-shell representations as semi-supervised learning relies on the supervised term to form a good approximation to start learning from the unlabeled data. We choose to only use the first shell for our deep learning methods to avoid this, but it is a considerable limitation. Despite only using a single-shell, the disentanglement models achieve lower MK error overall which is derived from diffusion kurtosis imaging (DKI) and requires multiple shells to be estimated. This suggests that the multi-shell SHORE representation can be approximated well, and that given the choice, harmonizing with a single-shell is better than using multi-shell data as it is.

An interesting aspect of the disentanglement model used here, is that the harmonization between the input takes place entirely in the encoders which are tasked with extracting either the anatomical or the acquisition specific information from the data. By preventing the gradients from the decoder $D_{targ}$ which is tasked with estimating the target site from being backpropagated to the encoder, we ensure that the encoder $E_{anat}$ is fully self-supervised along with the decoder $D_{acq}$. $D_{targ}$ can generate an image when parameterized by the unlabeled BLSA data even though it only learned from the labeled MUSHAC data, because the anatomical latent space which it learns from is free from acquisition specific features.

Though deep learning is a promising approach to harmonization, the advantage of such approaches lies in extracting information from large datasets. As was shown in the challenge associated with the MUSHAC dataset, with only 10 subjects in a supervised training set, regression models can outperform convolutional neural networks with millions of parameters. Additionally, a model trained to transform one site to another is only useful for those who need to harmonize data acquired on those specific scanners or those who have a large enough cohort to retrain the model. Where deep learning can be the most useful in harmonization is in developing a model, which can generalize well to unseen diffusion acquisitions. Because a fully supervised dataset of many subjects covering many acquisition parameters and scanner hardware is time consuming and expensive, it would be beneficial to leverage semi-supervised approaches to bring together many different datasets. Though this work takes a step towards this goal, it is limited to datasets which contain paired data which contain differences in acquisition or hardware between them.

The framework provided in this work can reduce error introduced by differences in acquisition and hardware in two unique datasets and can potentially be extended to many datasets provided they contain paired data. We advocate for further development of harmonization models which generalize across many datasets and account for the various differences in acquisition protocols in DW-MRI.

## Chapter X. Conclusion & Future Work

### 1. Conclusion

This dissertation focuses on the harmonization of DW-MRI through reducing site and acquisition specific bias. Relying on DW-MRI phantoms and empirically driven methodology has proven to be an effective approach to measuring, modelling, and correcting certain aspects of variability introduced by these biases. This form of modelling is convenient due to it being directly related to a physical property of the scanning system. It is then simple to argue in favor of the method being adapted for all processing pipelines just as we argue that the spatially varying effects of signal drift and gradient nonlinearity in DW-MRI need to be addressed in DW-MRI processing (chapters III and IV). However, the inconvenience lies in the practicality of the approach. Because these rely on some empirical measurement, it is at the discretion of each study to incorporate these measurements in their acquisitions. While this can range from a change in the acquisition parameters to a suite of entirely different acquisitions to be performed on each scanning system, the more important aspect of this is the considerable amount of data that were not acquired with the needed acquisitions or may have been acquired on now decommissioned systems. In some cases, such as in gradient nonlinearity in MRI, the knowledge of the physical properties of the system is sufficient to apply a correction. However, manufacturers are often hesitant to share such information, and while some problems are well known, the severity of their effect would need to be broadly acknowledged by researchers and clinicians before manufacturers will include the corrections within their hardware or software. Future studies can incorporate the required empirical methodology with each new discovery, but data driven approaches are necessary to harmonize existing datasets.

The promising field of deep learning has quickly spread to become a key tool in medical image processing. The large variability in human anatomy, while often intractable for traditional models, can be tackled with large models consisting of millions of parameters. The task of harmonization must account for the variability in anatomy while trying to solve for the variability introduced by differences in acquisition parameters, number of head coils, sensitivity of the coils, the imaging gradient non-linearity, the magnetic field homogeneity, the algorithms used to reconstruct the data, and the changes made during software upgrades. Thus, it is no surprise that deep learning approaches are being proposed for DW-MRI harmonization. Assuming is it possible to match the expected representation of the input the model was trained with, it is desirable to deploy the learned network to a new dataset. However, these models often suffer from a lack of understandability and are vulnerable to variations that were not

accounted for in the training set. Of the three main paradigms of deep learning, supervised learning is arguably the most powerful but also the least suited to many datasets. In medical imaging datasets, the barriers to annotation are the hourly cost of expert raters and the cost of matched acquisitions which in case of harmonization could mean a subject travelling around the globe. This is the driving rationale in using automated tractography methods for defining white matter bundle regions in chapters VII and VIII. Without the need of an expert rater, we were able to generate a large, labeled dataset for our atlas and our CNN. The unsupervised learning approach is the least restricted, and a few promising works have utilized unsupervised GAN models in DW-MRI harmonization. However, the Nash equilibrium found in training a GAN does not necessarily guarantee anatomically correct images as we saw with CycleGAN in chapter IX.

Semi-supervised learning leverages any available labeled and unlabeled data which is often the case for medical imaging datasets. A large suite of semi-supervised approaches exists for classification tasks, but these generally do not translate to image generation tasks. Nullspace tuning, which is explored in depth in this dissertation (chapters V and VI), is a form of semi-supervised contrastive learning which is well suited to medical imaging due to the presence of repeat acquisitions in medical datasets. While chapters V and VI focus on classification tasks (as they are easier to benchmark), the concept was first developed for harmonizing voxel-wise histology and DW-MRI derived FODs, and we show its use in chapter IX in harmonizing patches and slices from SHORE representations of DW-MRI. Where contrastive learning relies on an anchor, a positive sample (same class as the anchor), and a negative (different class from the anchor) sample, nullspace tuning is only concerned with the situation where only the anchor and the positive sample are provided. The contrastive term which constrains the model to provide similar results for paired data is effective for a large variety of tasks and data. While the novelty of the methodology is debatable, the semi-supervised use case we posit had not been completely investigated or established. However, nullspace tuning does not address the common situation in medical imaging where there could be many years between scans which can introduce anatomical changes due to age or disease. This will need to be a consideration for longitudinal datasets.

The use of structural T1 information in DW-MRI was essential for the construction of our white matter bundle atlas as well as the automated white matter bundle segmentation method (chapters VII and VIII). It is common practice to use a T1 acquisition to register DW-MRI data or metrics to a template space which is necessary step in atlas creation. However, the choice to solely rely on T1 information to extract white matter bundle regions was supported by the intuition that the general shape of a white matter bundle could be estimated by experts using only a

T1 as reference. In a similar manner, the white matter segmentation model estimates high level information about DW-MRI tractography. In chapter IX, we further rely on T1 derived information to supply global context to the patch or slice-based model which would otherwise have no indication as to the location of the brain the input originated. Because we find that removing the T1 information as a prior to the model negatively impacts performance, we know that this information is important when using a limited window of the brain.

Many deep learning methods presented for DW-MRI harmonization, while interesting and high performing with respect to some benchmark or metric, are typically built to be re-trained for each new study. This may require having some minimum number of subjects acquired on each scanner used in the study and matching subjects would be required for supervised methods. Additionally, models trained using spherical harmonics representations of the DW-MRI signal typically rely on a homogenous dataset where the acquisition parameters are similar or at least the b-values are the same. Chapter IX presents a semi-supervised framework which can incorporate multiple datasets with the requirement that these data be paired to enable nullspace tuning as well as the disentanglement model that was chosen. We rely on a representation which models DW-MRI using the same coefficients regardless of the acquisition parameters to allow for many different datasets. In doing so, we lay the groundwork for a generalizable learning scheme which could potentially be applied to unseen data given a large and representative training set.

**2. Contributions and Future Work**

Empirical and data driven methods for harmonizing DW-MRI and extracting DW-MRI information are important tools for diffusion related studies. This is especially true for studies which involve multiple acquisition protocols or imaging sites. In chapters III and IV, we investigate the use of empirically derived corrections for signal drift and gradient nonlinearity correction for DW-MRI which are acquisition and scanner specific biases. Next, we turn towards using data driven methods which are suited for extracting information in DW-MRI and identifying scanner effects which may looked over in empirical methods. We explore the use of a DW-MRI harmonization method called nullspace tuning in general machine learning tasks as well as other medical imaging domains in chapters V and VI. In chapters VII and VIII using a large MRI and DW-MRI dataset, we construct a set of white matter bundle atlases as well as an automated white matter bundle segmentation method which extracts information from only a T1 image. Using this knowledge of nullspace tuning and diffusion information present in T1, chapter IX presents a method for DW-MRI harmonization using T1 derived anatomical segmentations priors, nullspace tuning using paired

information, and acquisition independent diffusion signal representations.

## 2.1. Empirical DW-MRI Harmonization

### *2.1.1. Summary*

Previous to this work, signal drift correction in DW-MRI did not consider spatial variation in the drift during acquisitions with some hardware. We characterize this spatial variation along with the temporal drift in DW-MRI phantom acquisitions and show that there are instances in which this is necessary for correction. Additionally, we endeavored to find a solution for gradient nonlinearity correction in DW-MRI when manufacturer specifications are not available. We found that an oil carboy phantom could be used to empirically estimate the fieldmaps of an MRI system.

### *2.1.2. Main Contribution/Results*

Spatial-temporal signal drift correction was shown to be necessary in some acquisitions where the signal drift in some ROIs did not follow the global trend. This further shows the importance of signal drift correction as a standard preprocessing step which requires interspersed b0 volumes in DW-MRI. In evaluating the oil carboy phantom's ability to estimate the system's fields, we found the empirically estimated fieldmaps resulted in a corrected FA and MD that were similar to the FA and MD if the manufacturer fields were provided.

### *2.1.3. Future Work*

The main limitation to these empirical methods is the requirement of a certain acquisition. In the case of signal drift correction, scans already acquired cannot be corrected retroactively if interspersed b0s were not acquired. The empirical fieldmapping acquisition can be acquired at any point and then retroactively applied to DW-MRI acquired on the system, but a more convenient approach would be more widely adopted. Machine learning may be used to extract this information from cohorts acquired on a scanner of interest.

## 2.2. Applications of Nullspace Tuning

### *2.2.1. Summary*

Nullspace tuning is a form of semi-supervised contrastive learning which learns only from positive pairs. Originally used for DW-MRI harmonization of voxel-wise FODs, we explore its use for general machine learning

tasks relying on standard datasets and frameworks for comparing against state-of-the-art semi-supervised approaches. Additionally, we explore its use in other medical imaging domains specifically skin lesion classification and lung cancer diagnosis.

#### *2.2.2. Main Contribution/Results*

We found that nullspace tuning could provide significantly improved performance to classification models in natural image datasets and that this improvement could be additive when used in conjunction with state-of-the-art semi-supervised approaches. Unsurprisingly, similar classification results were obtained when the same experiment was performed in natural images of skin lesions. In a transfer learning scheme, we found a small improvement in lung cancer diagnosis when using nullspace tuning over strictly supervised learning.

#### *2.2.3. Future Work*

As longitudinal data is common in medical datasets, further exploration of the use of nullspace tuning should extend to accounting for anatomical changes over the course of time. Because our experiments were limited to classification tasks, further investigation is needed to extend this method to segmentation and image generation in other domains as well.

### 2.3. Uses of Structural T1 and Semi-supervised Learning in DW-MRI

#### *2.3.1. Summary*

With multiple datasets of DW-MRI and T1 acquisitions, we developed a population based white matter bundle atlas using six state-of-the-art automated white matter tractography methods. Rather than a parcellation, this atlas defines each white matter bundle within a different volume allowing for overlapping tracts. To investigate the ability of deep learning methods to extract white matter bundle information from T1 acquisitions, we train a SLANT model to predict subject specific bundles using only T1 images as input. Lastly, we investigate the use of T1 based parcellations as priors along with a nullspace tuning based approach for DW-MRI harmonization.

#### *2.3.2. Main Contribution/Results*

The resulting white matter atlas provides useful white matter definitions which can be used as priors for tractography, relating neuroimaging findings to structural pathways, or to inform future methodologies for parcellating and segmenting white matter based on functional, molecular, or alternative contrasts. In comparing the performance

of registration to the atlas and the T1 informed SLANT based method, we found that the bundle segmentations estimated from a T1 achieved similar or superior performance to the use of the atlas. For DW-MRI harmonization, we found that nullspace tuning allowed us to harmonize multiple datasets to a single target space and that the use of T1 informed parcellations as priors was beneficial to performance.

#### *2.3.3. Future Work*

Further investigation is needed to determine the difference in information provided between T1 structural and DW-MRI acquisitions. This work only shows that the white matter regions can be identified in T1 data and that parcellation derived from T1 can be informative to DW-MRI methods. However, further insight could be provided as to what if any information about the underlying microstructure can be mined from large T1 datasets.

**References**


1. Huo, Y., et al. *Spatially localized atlas network tiles enables 3D whole brain segmentation from limited data*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2018. Springer.
2. Johnson, W.E., C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical Bayes methods.* Biostatistics, 2007. **8**(1): p. 118-127.
3. Lawrence, R.M., et al., *Standardizing human brain parcellations.* Scientific Data, 2021. **8**(1): p. 1-9.
4. Mirzaalian, H., et al., *Inter-site and inter-scanner diffusion MRI data harmonization.* NeuroImage, 2016. **135**: p. 311-323.
5. Ozarslan, E., et al., *Simple harmonic oscillator based reconstruction and estimation for three-dimensional q-space MRI.* 2009.
6. Koppers, S., et al. *Spherical harmonic residual network for diffusion signal harmonization*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2018. Springer.
7. Moyer, D., et al., *Scanner Invariant Representations for Diffusion MRI Harmonization.* arXiv preprint arXiv:1904.05375, 2019.
8. Nath, V., et al. *Inter-scanner harmonization of high angular resolution DW-MRI using null space deep learning*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2018. Springer.
9. Koppers, S., C. Haarburger, and D. Merhof. *Diffusion MRI signal augmentation: from single shell to multi shell with deep learning*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2016. Springer.
10. Warach, S., *Pitfalls and potential of clinical diffusion-weighted MR imaging in acute stroke.* Stroke, 1997. **28**: p. 481-482.
11. Warach, S., J.F. Dashe, and R.R. Edelman, *Clinical outcome in ischemic stroke predicted by early diffusion-weighted and perfusion magnetic resonance imaging: a preliminary analysis.* Journal of Cerebral Blood Flow & Metabolism, 1996. **16**(1): p. 53-59.
12. Gonzalez, R.G., et al., *Diffusion-weighted MR imaging: diagnostic accuracy in patients imaged within 6 hours of stroke symptom onset.* Radiology, 1999. **210**(1): p. 155-162.
13. Guo, Y., et al., *Differentiation of clinically benign and malignant breast lesions using diffusion-weighted imaging.* Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine, 2002. **16**(2): p. 172-178.
14. Turkbey, B., et al., *Is apparent diffusion coefficient associated with clinical risk scores for prostate cancers that are visible on 3-T MR images?* Radiology, 2011. **258**(2): p. 488-495.
15. Cui, Y., et al., *Apparent diffusion coefficient: potential imaging biomarker for prediction and early detection of response to chemotherapy in hepatic metastases.* Radiology, 2008. **248**(3): p. 894-900.
16. Le Bihan, D., *Looking into the functional architecture of the brain with diffusion*

*MRI.* Nature Reviews Neuroscience, 2003. **4**(6): p. 469-480.
17. Le Bihan, D. and H. Johansen-Berg, *Diffusion MRI at 25: exploring brain tissue structure and function.* Neuroimage, 2012. **61**(2): p. 324-341.
18. Basser, P.J., et al., *In vivo fiber tractography using DT-MRI data.* Magnetic resonance in medicine, 2000. **44**(4): p. 625-632.
19. Jones, D.K., T.R. Knösche, and R. Turner, *White matter integrity, fiber count, and other fallacies: the do's and don'ts of diffusion MRI.* Neuroimage, 2013. **73**: p. 239-254.
20. Van Essen, D.C., et al., *The Human Connectome Project: a data acquisition perspective.* Neuroimage, 2012. **62**(4): p. 2222-2231.
21. Van Essen, D.C., et al., *The WU-Minn human connectome project: an overview.* Neuroimage, 2013. **80**: p. 62-79.
22. Kawas, C., et al., *Age-specific incidence rates of Alzheimer's disease: the Baltimore Longitudinal Study of Aging.* Neurology, 2000. **54**(11): p. 2072-2077.
23. Shock, N.W., *Normal human aging: The Baltimore longitudinal study of aging*. 1984: US Department of Health and Human Services, Public Health Service, National ….
24. Vollmar, C., et al., *Identical, but not the same: intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0 T scanners.* Neuroimage, 2010. **51**(4): p. 1384-1394.
25. Matsui, J.T., *Development of image processing tools and procedures for analyzing multi-site longitudinal diffusion-weighted imaging studies.* 2014.
26. Zhu, T., et al., *Quantification of accuracy and precision of multi-center DTI measurements: a diffusion phantom and human brain study.* Neuroimage, 2011. **56**(3): p. 1398-1411.
27. Jovicich, J., et al., *Multisite longitudinal reliability of tract-based spatial statistics in diffusion tensor imaging of healthy elderly subjects.* Neuroimage, 2014. **101**: p. 390-403.
28. Teipel, S.J., et al., *Multicenter stability of diffusion tensor imaging measures: a European clinical and physical phantom study.* Psychiatry Research: Neuroimaging, 2011. **194**(3): p. 363-371.
29. Rogers, B.P., et al. *Stability of gradient field corrections for quantitative diffusion MRI.* in *Medical Imaging 2017: Physics of Medical Imaging*. 2017. International Society for Optics and Photonics.
30. Bammer, R., et al., *Analysis and generalized correction of the effect of spatial gradient field distortions in diffusion-weighted imaging.* Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 2003. **50**(3): p. 560-569.
31. Tao, A.T., et al., *Improving apparent diffusion coefficient accuracy on a compact 3T MRI scanner using gradient nonlinearity correction.* Journal of Magnetic Resonance Imaging, 2018. **48**(6): p. 1498-1507.
32. Newitt, D.C., et al., *Gradient nonlinearity correction to improve apparent diffusion coefficient accuracy and standardization in the american college of radiology imaging network 6698 breast cancer trial.* Journal of Magnetic Resonance Imaging, 2015. **42**(4): p. 908-919.
33. Malyarenko, D.I., B.D. Ross, and T.L. Chenevert, *Analysis and correction of*

*gradient nonlinearity bias in apparent diffusion coefficient measurements.* Magnetic resonance in medicine, 2014. **71**(3): p. 1312-1323.
34. Fortin, J.-P., et al., *Harmonization of cortical thickness measurements across scanners and sites.* Neuroimage, 2018. **167**: p. 104-120.
35. Fortin, J.-P., et al., *Harmonization of multi-site diffusion tensor imaging data.* Neuroimage, 2017. **161**: p. 149-170.
36. Prohl, A.K., et al., *Reproducibility of structural and diffusion tensor imaging in the TACERN multi-center study.* Frontiers in integrative neuroscience, 2019. **13**: p. 24.
37. Hansen, C.B., et al., *Characterization and correlation of signal drift in diffusion weighted MRI.* Magnetic resonance imaging, 2019. **57**: p. 133-142.
38. Provenzale, J.M., et al., *Analysis of variability of fractional anisotropy values at 3T using a novel diffusion tensor imaging phantom.* The neuroradiology journal, 2018. **31**(6): p. 581-586.
39. Fan, Q., et al., *Validation of diffusion MRI estimates of compartment size and volume fraction in a biomimetic brain phantom using a human MRI scanner with 300 mT/m maximum gradient strength.* Neuroimage, 2018. **182**: p. 469-478.
40. Fan, Q., et al., *A comprehensive diffusion MRI dataset acquired on the MGH Connectome scanner in a biomimetic brain phantom.* Data in brief, 2018. **18**: p. 334-339.
41. Ning, L., et al. *Muti-shell diffusion MRI harmonisation and enhancement challenge (MUSHAC): progress and results*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2018. Springer.
42. St-Jean, S., P. Coupé, and M. Descoteaux, *Non Local Spatial and Angular Matching: Enabling higher spatial resolution diffusion MRI datasets through adaptive denoising.* Medical image analysis, 2016. **32**: p. 115-130.
43. Tax, C.M., et al., *Cross-scanner and cross-protocol diffusion MRI data harmonisation: A benchmark database and evaluation of algorithms.* NeuroImage, 2019. **195**: p. 285-299.
44. Zeng, X., et al., *Segmentation and measurement of the cortex from 3-D MR images using coupled-surfaces propagation.* IEEE transactions on medical imaging, 1999. **18**(10): p. 927-937.
45. Hahn, E.L., *Spin echoes.* Physical review, 1950. **80**(4): p. 580.
46. Carr, H.Y. and E.M. Purcell, *Effects of diffusion on free precession in nuclear magnetic resonance experiments.* Physical review, 1954. **94**(3): p. 630.
47. Torrey, H.C., *Bloch equations with diffusion terms.* Physical review, 1956. **104**(3): p. 563.
48. Stejskal, E.O. and J.E. Tanner, *Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient.* The journal of chemical physics, 1965. **42**(1): p. 288-292.
49. Einstein, A., *On the motion required by the molecular kinetic theory of heat of small particles suspended in a stationary liquid.* Annalen der physik, 1905. **17**(8): p. 549-560.
50. Johansen-Berg, H. and T.E. Behrens, *Diffusion MRI: from quantitative measurement to in vivo neuroanatomy*. 2013: Academic Press.
51. Moseley, M., et al., *Diffusion-weighted MR imaging of acute stroke: correlation*

*with T2-weighted and magnetic susceptibility-enhanced MR imaging in cats.* American Journal of Neuroradiology, 1990. **11**(3): p. 423-429.
52. Beaulieu, C., *The basis of anisotropic water diffusion in the nervous system–a technical review.* NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo, 2002. **15**(7-8): p. 435-455.
53. Basser, P.J., J. Mattiello, and D. LeBihan, *MR diffusion tensor spectroscopy and imaging.* Biophysical journal, 1994. **66**(1): p. 259-267.
54. Basser, P.J., J. Mattiello, and D. LeBihan, *Estimation of the effective self-diffusion tensor from the NMR spin echo.* Journal of Magnetic Resonance, Series B, 1994. **103**(3): p. 247-254.
55. Alexander, A.L., et al., *Diffusion tensor imaging of the brain.* Neurotherapeutics, 2007. **4**(3): p. 316-329.
56. Alexander, D., G. Barker, and S. Arridge, *Detection and modeling of non-Gaussian apparent diffusion coefficient profiles in human brain data.* Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 2002. **48**(2): p. 331-340.
57. Behrens, T.E., et al., *Probabilistic diffusion tractography with multiple fibre orientations: What can we gain?* Neuroimage, 2007. **34**(1): p. 144-155.
58. Anderson, A.W., *Measurement of fiber orientation distributions using high angular resolution diffusion imaging.* Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 2005. **54**(5): p. 1194-1206.
59. Anderson, A. and Z. Ding. *Sub-voxel measurement of fiber orientation using high angular resolution diffusion tensor imaging*. in *Book of abstracts: Tenth Annual Meeting of the International Society for Magnetic Resonance in Medicine. Berkeley, CA: ISMRM*. 2002.
60. Tournier, J.-D., et al., *Direct estimation of the fiber orientation density function from diffusion-weighted MRI data using spherical deconvolution.* NeuroImage, 2004. **23**(3): p. 1176-1185.
61. Özarslan, E., et al., *Resolution of complex tissue microarchitecture using the diffusion orientation transform (DOT).* NeuroImage, 2006. **31**(3): p. 1086-1103.
62. Aganj, I., et al., *Reconstruction of the orientation distribution function in single- and multiple-shell q-ball imaging within constant solid angle.* Magnetic resonance in medicine, 2010. **64**(2): p. 554-566.
63. Yeh, F.-C., V.J. Wedeen, and W.-Y.I. Tseng, *Generalized ${q} $-sampling imaging.* IEEE transactions on medical imaging, 2010. **29**(9): p. 1626-1635.
64. Daducci, A., et al., *Quantitative comparison of reconstruction methods for intra-voxel fiber recovery from diffusion MRI.* IEEE transactions on medical imaging, 2013. **33**(2): p. 384-399.
65. Koutsarnakis, C., et al., *A laboratory manual for stepwise cerebral white matter fiber dissection.* World neurosurgery, 2015. **84**(2): p. 483-493.
66. Catani, M. and M.T. De Schotten, *A diffusion tensor imaging tractography atlas for virtual in vivo dissections.* cortex, 2008. **44**(8): p. 1105-1132.
67. Mori, S., et al., *Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging.* Annals of Neurology: Official Journal of the

American Neurological Association and the Child Neurology Society, 1999. **45**(2): p. 265-269.
68. Lazar, M. and A.L. Alexander, *An error analysis of white matter tractography methods: synthetic diffusion tensor field simulations.* Neuroimage, 2003. **20**(2): p. 1140-1153.
69. Jeurissen, B., et al., *Probabilistic fiber tracking using the residual bootstrap with constrained spherical deconvolution.* Human brain mapping, 2011. **32**(3): p. 461-479.
70. Jones, D.K., *Determining and visualizing uncertainty in estimates of fiber orientation from diffusion tensor MRI.* Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 2003. **49**(1): p. 7-12.
71. Behrens, T.E., et al., *Characterization and propagation of uncertainty in diffusion-weighted MR imaging.* Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 2003. **50**(5): p. 1077-1088.
72. Kaden, E., T.R. Knösche, and A. Anwander, *Parametric spherical deconvolution: inferring anatomical connectivity using diffusion MR imaging.* NeuroImage, 2007. **37**(2): p. 474-488.
73. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* nature, 2015. **521**(7553): p. 436-444.
74. Krizhevsky, A., I. Sutskever, and G.E. Hinton. *Imagenet classification with deep convolutional neural networks*. in *Advances in neural information processing systems*. 2012.
75. Farabet, C., et al., *Learning hierarchical features for scene labeling.* IEEE transactions on pattern analysis and machine intelligence, 2012. **35**(8): p. 1915-1929.
76. Tompson, J.J., et al. *Joint training of a convolutional network and a graphical model for human pose estimation*. in *Advances in neural information processing systems*. 2014.
77. Szegedy, C., et al. *Going deeper with convolutions*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
78. Mikolov, T., et al. *Strategies for training large scale neural network language models*. in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. 2011. IEEE.
79. Hinton, G., et al., *Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups.* IEEE Signal processing magazine, 2012. **29**(6): p. 82-97.
80. Sainath, T.N., et al. *Deep convolutional neural networks for LVCSR*. in *2013 IEEE international conference on acoustics, speech and signal processing*. 2013. IEEE.
81. Ma, J., et al., *Deep neural nets as a method for quantitative structure–activity relationships.* Journal of chemical information and modeling, 2015. **55**(2): p. 263-274.
82. Ciodaro, T., et al. *Online particle detection with neural networks based on topological calorimetry information*. in *Journal of physics: conference series*. 2012. IOP Publishing.

83. Adam-Bourdarios, C., et al. *The Higgs boson machine learning challenge*. in *NIPS 2014 Workshop on High-energy Physics and Machine Learning*. 2015.
84. Helmstaedter, M., et al., *Connectomic reconstruction of the inner plexiform layer in the mouse retina.* Nature, 2013. **500**(7461): p. 168-174.
85. Leung, M.K., et al., *Deep learning of the tissue-regulated splicing code.* Bioinformatics, 2014. **30**(12): p. i121-i129.
86. Xiong, H.Y., et al., *The human splicing code reveals new insights into the genetic determinants of disease.* Science, 2015. **347**(6218): p. 1254806.
87. Srivastava, N., et al., *Dropout: a simple way to prevent neural networks from overfitting.* The journal of machine learning research, 2014. **15**(1): p. 1929-1958.
88. Perez, L. and J. Wang, *The effectiveness of data augmentation in image classification using deep learning.* arXiv preprint arXiv:1712.04621, 2017.
89. Zhu, X. and A.B. Goldberg, *Introduction to semi-supervised learning.* Synthesis lectures on artificial intelligence and machine learning, 2009. **3**(1): p. 1-130.
90. Laine, S. and T. Aila, *Temporal ensembling for semi-supervised learning.* arXiv preprint arXiv:1610.02242, 2016.
91. Sajjadi, M., M. Javanmardi, and T. Tasdizen. *Regularization with stochastic transformations and perturbations for deep semi-supervised learning*. in *Advances in neural information processing systems*. 2016.
92. Tarvainen, A. and H. Valpola. *Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results*. in *Advances in neural information processing systems*. 2017.
93. Lee, D.-H. *Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks*. in *Workshop on challenges in representation learning, ICML*. 2013.
94. Miyato, T., et al., *Virtual adversarial training: a regularization method for supervised and semi-supervised learning.* IEEE transactions on pattern analysis and machine intelligence, 2018. **41**(8): p. 1979-1993.
95. Ronneberger, O., P. Fischer, and T. Brox. *U-net: Convolutional networks for biomedical image segmentation*. in *International Conference on Medical image computing and computer-assisted intervention*. 2015. Springer.
96. Hochreiter, S., *The vanishing gradient problem during learning recurrent neural nets and problem solutions.* International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 1998. **6**(02): p. 107-116.
97. Koppers, S. and D. Merhof. *Direct estimation of fiber orientations using deep learning in diffusion imaging*. in *International Workshop on Machine Learning in Medical Imaging*. 2016. Springer.
98. Lancaster, J.L., et al., *Automated Talairach atlas labels for functional brain mapping.* Human brain mapping, 2000. **10**(3): p. 120-131.
99. Mazziotta, J., et al., *A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM).* Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 2001. **356**(1412): p. 1293-1322.
100. Collins, D.L., et al., *Automatic 3-D model-based neuroanatomical segmentation.* Human brain mapping, 1995. **3**(3): p. 190-208.
101. Myers, P.E., et al., *Standardizing human brain parcellations.* BioRxiv, 2019: p.

845065.
102. Glasser, M.F., et al., *A multi-modal parcellation of human cerebral cortex.* Nature, 2016. **536**(7615): p. 171-178.
103. Schaefer, A., et al., *Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI.* Cerebral Cortex, 2018. **28**(9): p. 3095-3114.
104. Thomas Yeo, B., et al., *The organization of the human cerebral cortex estimated by intrinsic functional connectivity.* Journal of neurophysiology, 2011. **106**(3): p. 1125-1165.
105. Rolls, E.T., et al., *Automated anatomical labelling atlas 3.* NeuroImage, 2020. **206**: p. 116189.
106. Tzourio-Mazoyer, N., et al., *Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain.* Neuroimage, 2002. **15**(1): p. 273-289.
107. Mars, R.B., et al., *Diffusion-weighted imaging tractography-based parcellation of the human parietal cortex and comparison with human and macaque resting-state functional connectivity.* Journal of Neuroscience, 2011. **31**(11): p. 4087-4100.
108. Mars, R.B., et al., *Connectivity-based subdivisions of the human right "temporoparietal junction area": evidence for different areas participating in different cortical networks.* Cerebral cortex, 2012. **22**(8): p. 1894-1903.
109. Neubert, F.-X., et al., *Comparison of human ventral frontal cortex areas for cognitive control and language with areas in monkey frontal cortex.* Neuron, 2014. **81**(3): p. 700-713.
110. Makris, N., et al., *Decreased volume of left and total anterior insular lobule in schizophrenia.* Schizophrenia research, 2006. **83**(2-3): p. 155-171.
111. Desikan, R.S., et al., *An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest.* Neuroimage, 2006. **31**(3): p. 968-980.
112. Eickhoff, S.B., et al., *Assignment of functional activations to probabilistic cytoarchitectonic areas revisited.* Neuroimage, 2007. **36**(3): p. 511-521.
113. Oishi, K., et al., *Atlas-based whole brain white matter analysis using large deformation diffeomorphic metric mapping: application to normal elderly and Alzheimer's disease participants.* Neuroimage, 2009. **46**(2): p. 486-499.
114. Gholipour, A., et al., *Brain functional localization: a survey of image registration techniques.* IEEE transactions on medical imaging, 2007. **26**(4): p. 427-451.
115. Lötjönen, J.M., et al., *Fast and robust multi-atlas segmentation of brain magnetic resonance images.* Neuroimage, 2010. **49**(3): p. 2352-2365.
116. Pipitone, J., et al., *Multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates.* Neuroimage, 2014. **101**: p. 494-512.
117. Huo, Y., et al., *Consistent cortical reconstruction and multi-atlas brain segmentation.* NeuroImage, 2016. **138**: p. 197-210.
118. Asman, A.J., et al., *Multi-atlas learner fusion: An efficient segmentation approach for large-scale data.* Medical image analysis, 2015. **26**(1): p. 82-91.
119. Pham, D.L. and J.L. Prince, *Adaptive fuzzy segmentation of magnetic resonance images.* IEEE transactions on medical imaging, 1999. **18**(9): p. 737-752.
120. de Brebisson, A. and G. Montana. *Deep neural networks for anatomical brain*

*segmentation*. in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2015.
121. Mehta, R., A. Majumdar, and J. Sivaswamy, *BrainSegNet: a convolutional neural network architecture for automated segmentation of human brain structures.* Journal of Medical Imaging, 2017. **4**(2): p. 024003.
122. Wachinger, C., M. Reuter, and T. Klein, *DeepNAT: Deep convolutional neural network for segmenting neuroanatomy.* NeuroImage, 2018. **170**: p. 434-445.
123. Klein, A., et al. *Open labels: online feedback for a public resource of manually labeled brain images*. in *16th Annual Meeting for the Organization of Human Brain Mapping*. 2010.
124. Veenith, T.V., et al., *Inter subject variability and reproducibility of diffusion tensor imaging within and between different imaging sessions.* PloS one, 2013. **8**(6).
125. Giannelli, M., et al., *MR scanner systems should be adequately characterized in diffusion-MRI of the breast.* PLoS One, 2014. **9**(1).
126. Cannon, T.D., et al., *Reliability of neuroanatomical measurements in a multisite longitudinal study of youth at risk for psychosis.* Human brain mapping, 2014. **35**(5): p. 2424-2434.
127. Lemkaddem, A., et al., *A multi-center study: intra-scan and inter-scan variability of diffusion spectrum imaging.* Neuroimage, 2012. **62**(1): p. 87-94.
128. Shokouhi, M., et al., *Assessment of the impact of the scanner-related factors on brain morphometry analysis with Brainvisa.* BMC Medical Imaging, 2011. **11**(1): p. 23.
129. Fox, R., et al., *A validation study of multicenter diffusion tensor imaging: reliability of fractional anisotropy and diffusivity values.* American journal of neuroradiology, 2012. **33**(4): p. 695-700.
130. Nyholm, T., et al., *Variability in prostate and seminal vesicle delineations defined on magnetic resonance images, a multi-observer,-center and-sequence study.* Radiation oncology, 2013. **8**(1): p. 126.
131. Han, X., et al., *Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer.* Neuroimage, 2006. **32**(1): p. 180-194.
132. Schilling, K.G., et al., *Empirical consideration of the effects of acquisition parameters and analysis model on clinically feasible q-ball imaging.* Magnetic resonance imaging, 2017. **40**: p. 62-74.
133. Nath, V., et al. *Comparison of multi-fiber reproducibility of PAS-MRI and Q-ball with empirical multiple b-value HARDI*. in *Medical Imaging 2017: Image Processing*. 2017. International Society for Optics and Photonics.
134. Andersson, J.L., S. Skare, and J. Ashburner, *How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging.* Neuroimage, 2003. **20**(2): p. 870-888.
135. Andersson, J.L. and S.N. Sotiropoulos, *An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging.* Neuroimage, 2016. **125**: p. 1063-1078.
136. Smith, S.M., *Fast robust automated brain extraction.* Human brain mapping, 2002. **17**(3): p. 143-155.

137. Jones, D.K., *Diffusion mri*. 2010: Oxford University Press.
138. Fischer, B. and J. Modersitzki. *FLIRT: A flexible image registration toolbox*. in *International Workshop on Biomedical Image Registration*. 2003. Springer.
139. Andersson, J., S. Smith, and M. Jenkinson, *FNIRT–FMRIB's non-linear image registration tool.* Human Brain Mapping, 2008. **2008**.
140. Avants, B.B., N. Tustison, and G. Song, *Advanced normalization tools (ANTS).* Insight j, 2009. **2**(365): p. 1-35.
141. Rogers, B.P., et al. *Phantom-based field maps for gradient nonlinearity correction in diffusion imaging*. in *Medical Imaging 2018: Physics of Medical Imaging*. 2018. International Society for Optics and Photonics.
142. Vos, S.B., et al., *The importance of correcting for signal drift in diffusion MRI.* Magnetic resonance in medicine, 2017. **77**(1): p. 285-299.
143. Thesen, S., G. Kruger, and E. Muller. *Absolute correction of B0 fluctuations in echo-planar imaging*. in *Proceedings of the International Society of Magnetic Resonance in Medicine*. 2003.
144. Benner, T., et al., *Real-time RF pulse adjustment for B0 drift correction.* Magnetic resonance in medicine, 2006. **56**(1): p. 204-209.
145. Asman, A.J. and B.A. Landman, *Hierarchical performance estimation in the statistical label fusion framework.* Med Image Anal, 2014. **18**(7): p. 1070-81.
146. Smith, S.M., et al., *Advances in functional and structural MR image analysis and implementation as FSL.* Neuroimage, 2004. **23 Suppl 1**: p. S208-19.
147. Jeurissen, B., et al., *Investigating the prevalence of complex fiber configurations in white matter tissue with diffusion magnetic resonance imaging.* Human brain mapping, 2013. **34**(11): p. 2747-2766.
148. Reijmer, Y.D., et al., *Improved sensitivity to cerebral white matter abnormalities in Alzheimer's disease with spherical deconvolution based tractography.* PloS one, 2012. **7**(8): p. e44074.
149. Yogarajah, M., et al., *Tractography of the parahippocampal gyrus and material specific memory impairment in unilateral temporal lobe epilepsy.* Neuroimage, 2008. **40**(4): p. 1755-1764.
150. Pierpaoli, C., et al., *Water diffusion changes in Wallerian degeneration and their dependence on white matter architecture.* Neuroimage, 2001. **13**(6): p. 1174-1185.
151. Farrell, J.A., et al., *Effects of signal-to-noise ratio on the accuracy and reproducibility of diffusion tensor imaging–derived fractional anisotropy, mean diffusivity, and principal eigenvector measurements at 1.5 T.* Journal of Magnetic Resonance Imaging, 2007. **26**(3): p. 756-767.
152. Landman, B.A., et al., *Effects of diffusion weighting schemes on the reproducibility of DTI-derived fractional anisotropy, mean diffusivity, and principal eigenvector measurements at 1.5 T.* Neuroimage, 2007. **36**(4): p. 1123-1138.
153. Grech-Sollars, M., et al., *Multi-centre reproducibility of diffusion MRI parameters for clinical sequences in the brain.* NMR in Biomedicine, 2015. **28**(4): p. 468-485.
154. Dumouchel, W. and F. O'Brien. *Integrating a robust option into a multiple regression computing environment*. in *Computing and graphics in statistics*. 1992. Springer-Verlag New York, Inc.
155. Tanabe, J., et al., *Comparison of detrending methods for optimal fMRI*

*preprocessing*. NeuroImage, 2002. **15**(4): p. 902-907.
156. Bandettini, P.A., et al., *Processing strategies for time-course data sets in functional MRI of the human brain.* Magnetic resonance in medicine, 1993. **30**(2): p. 161-173.
157. Skudlarski, P., R.T. Constable, and J.C. Gore, *ROC analysis of statistical methods used in functional MRI: individual subjects*. Neuroimage, 1999. **9**(3): p. 311-329.
158. Mattay, V.S., et al., *Whole-brain functional mapping with isotropic MR imaging.* Radiology, 1996. **201**(2): p. 399-404.
159. Genovese, C.R., *A Bayesian time-course model for functional magnetic resonance imaging data*. Journal of the American Statistical Association, 2000. **95**(451): p. 691-703.
160. Meyer, F.G., *Wavelet-based estimation of a semiparametric generalized linear model of fMRI time-series*. IEEE transactions on medical imaging, 2003. **22**(3): p. 315-322.
161. Macey, P.M., et al., *A method for removal of global effects from fMRI time series.* Neuroimage, 2004. **22**(1): p. 360-366.
162. Glover, G.H. and N.J. Pelc, *Method for correcting image distortion due to gradient nonuniformity*. 1986, Google Patents.
163. Michiels, J., et al., *On the problem of geometric distortion in magnetic resonance images for stereotactic neurosurgery.* Magnetic resonance imaging, 1994. **12**(5): p. 749-765.
164. Sumanaweera, T., et al., *Quantifying MRI geometric distortion in tissue.* Magnetic resonance in medicine, 1994. **31**(1): p. 40-47.
165. Langlois, S., et al., *MRI geometric distortion: a simple approach to correcting the effects of non-linear gradient fields.* Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine, 1999. **9**(6): p. 821-831.
166. LeBihan, D. and R. Tumer, *Diffusion and perfusion, magnetic resonance imaging, Mozbey Year Book*. 1992, Inc.
167. Conturo, T.E., et al., *Diffusion MRI: precision, accuracy and flow effects.* NMR in Biomedicine, 1995. **8**(7): p. 307-332.
168. Bernstein, M.A. and J.A. Polzin, *Method and system for correcting errors in MR images due to regions of gradient non-uniformity for parametric imaging such as quantitative flow analysis*. 2000, Google Patents.
169. Bammer, R., et al., *Assessment of spatial gradient field distortion in diffusion-weighted imaging*. Proceedings of the International Society for Magnetic Resonance in Medicine , Honolulu, HI, 2002: p. 1172.
170. Robson, M. *Non-linear gradients on clinical MRI systems introduce systematic errors in ADC and DTI measurements*. in *Proceedings of the 10th Annual Meeting of ISMRM, Honolulu*. 2002.
171. Basser, P.J., *Inferring microstructural features and the physiological state of tissues from diffusion-weighted images.* NMR in Biomedicine, 1995. **8**(7): p. 333-344.
172. Frank, L.R., *Anisotropy in high angular resolution diffusion-weighted MRI.* Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 2001. **45**(6): p. 935-939.

173. Setsompop, K., et al., *Pushing the limits of in vivo diffusion MRI for the Human Connectome Project.* Neuroimage, 2013. **80**: p. 220-233.
174. Malyarenko, D.I., et al., *Demonstration of nonlinearity bias in the measurement of the apparent diffusion coefficient in multicenter trials.* J Magnetic resonance in medicine, 2016. **75**(3): p. 1312-1323.
175. McNab, J.A., et al., *The Human Connectome Project and beyond: initial applications of 300 mT/m gradients.* Neuroimage, 2013. **80**: p. 234-245.
176. Mesri, H.Y., et al., *The adverse effect of gradient nonlinearities on diffusion MRI: From voxels to group studies.* NeuroImage, 2019: p. 116127.
177. Tan, E.T., et al., *Improved correction for gradient nonlinearity effects in diffusion-weighted imaging.* Journal of Magnetic Resonance Imaging, 2013. **38**(2): p. 448-453.
178. Malyarenko, D.I. and T.L. Chenevert, *Practical estimate of gradient nonlinearity for implementation of apparent diffusion coefficient bias correction.* Journal of Magnetic Resonance Imaging, 2014. **40**(6): p. 1487-1495.
179. Sotiropoulos, S.N., et al., *Advances in diffusion MRI acquisition and processing in the Human Connectome Project.* Neuroimage, 2013. **80**: p. 125-143.
180. Glasser, M.F., et al., *The minimal preprocessing pipelines for the Human Connectome Project.* Neuroimage, 2013. **80**: p. 105-124.
181. Janke, A., et al., *Use of spherical harmonic deconvolution methods to compensate for nonlinear gradient effects on MRI images.* Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 2004. **52**(1): p. 115-122.
182. Doran, S.J., et al., *A complete distortion correction for MR images: I. Gradient warp correction.* Physics in Medicine & Biology, 2005. **50**(7): p. 1343.
183. Tao, S., et al., *Integrated image reconstruction and gradient nonlinearity correction.* Magnetic resonance in medicine, 2015. **74**(4): p. 1019-1031.
184. Tao, S., et al., *NonCartesian MR image reconstruction with integrated gradient nonlinearity correction.* Medical physics, 2015. **42**(12): p. 7190-7201.
185. Tough, R.J. and A.J. Stone, *Properties of the regular and irregular solid harmonics.* Journal Of Physics A: Mathematical General, 1977. **10**(8): p. 1261.
186. Caola, M., *Solid harmonics and their addition theorems.* Journal of Physics A: Mathematical General, 1978. **11**(2): p. L23.
187. Makris, N., et al., *MRI-based anatomical model of the human head for specific absorption rate mapping.* Medical & biological engineering & computing, 2008. **46**(12): p. 1239-1251.
188. Mattiello, J., P.J. Basser, and D. Le Bihan, *The b matrix in diffusion tensor echo-planar imaging.* Magnetic Resonance in Medicine, 1997. **37**(2): p. 292-300.
189. Mattiello, J., P.J. Basser, and D. LeBihan, *Analytical expressions for the b matrix in NMR diffusion imaging and spectroscopy.* Journal of magnetic resonance, Series A, 1994. **108**(2): p. 131-141.
190. Alger, J.R., *The diffusion tensor imaging toolbox.* Journal of Neuroscience, 2012. **32**(22): p. 7418-7428.
191. Pierpaoli, C., et al. *Polyvinylpyrrolidone (PVP) water solutions as isotropic phantoms for diffusion MRI studies*. in *Proc Intl Soc Magn Reson Med*. 2009.
192. Jenkinson, M., et al., *Improved optimization for the robust and accurate linear*

*registration and motion correction of brain images.* Neuroimage, 2002. **17**(2): p. 825-841.
193. Lee, Y., et al., *A comprehensive approach for correcting voxel-wise b-value errors in diffusion MRI.* 2019.
194. Chapelle, O., B. Schölkopf, and A. Zien, *Semi-supervised learning, vol. 2.* Cambridge: MIT Press. Cortes, C., & Mohri, M.(2014). Domain adaptation and sample bias correction theory and algorithm for regression. Theoretical Computer Science, 2006. **519**: p. 103126.
195. Zhu, X., Z. Ghahramani, and J.D. Lafferty. *Semi-supervised learning using gaussian fields and harmonic functions*. in *Proceedings of the 20th International conference on Machine learning (ICML-03)*. 2003.
196. Verma, V., et al., *Interpolation consistency training for semi-supervised learning.* arXiv preprint arXiv:1903.03825, 2019.
197. Cubuk, E.D., et al. *Autoaugment: Learning augmentation strategies from data*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2019.
198. Berthelot, D., et al. *Mixmatch: A holistic approach to semi-supervised learning*. in *Advances in Neural Information Processing Systems*. 2019.
199. Yankelevitz, D.F., et al., *Small pulmonary nodules: evaluation with repeat CT—preliminary experience.* Radiology, 1999. **212**(2): p. 561-566.
200. Freeborough, P.A. and N.C. Fox, *The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI.* IEEE transactions on medical imaging, 1997. **16**(5): p. 623-629.
201. Dirix, P., et al., *Dose painting in radiotherapy for head and neck squamous cell carcinoma: value of repeated functional imaging with 18F-FDG PET, 18F-fluoromisonidazole PET, diffusion-weighted MRI, and dynamic contrast-enhanced MRI.* Journal of Nuclear Medicine, 2009. **50**(7): p. 1020-1027.
202. Huo, Y., et al. *Coronary calcium detection using 3D attention identical dual deep network based on weakly supervised learning*. in *Medical Imaging 2019: Image Processing*. 2019. International Society for Optics and Photonics.
203. Oliver, A., et al. *Realistic evaluation of deep semi-supervised learning algorithms*. in *Advances in Neural Information Processing Systems*. 2018.
204. Chen, T., et al. *A simple framework for contrastive learning of visual representations*. in *International conference on machine learning*. 2020. PMLR.
205. Zhu, X.J., *Semi-supervised learning literature survey*. 2005, University of Wisconsin-Madison Department of Computer Sciences.
206. Kumar, B., G. Carneiro, and I. Reid. *Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
207. Cireşan, D.C., et al., *Deep, big, simple neural nets for handwritten digit recognition.* Neural computation, 2010. **22**(12): p. 3207-3220.
208. Simard, P.Y., D. Steinkraus, and J.C. Platt. *Best practices for convolutional neural networks applied to visual document analysis*. in *Icdar*. 2003.
209. Wan, L., et al. *Regularization of neural networks using dropconnect*. in *International conference on machine learning*. 2013.

210. Sato, I., H. Nishimura, and K. Yokoi, *Apac: Augmented pattern classification with neural networks.* arXiv preprint arXiv:1505.03229, 2015.
211. Ciregan, D., U. Meier, and J. Schmidhuber. *Multi-column deep neural networks for image classification*. in *2012 IEEE conference on computer vision and pattern recognition*. 2012. IEEE.
212. Zhang, H., et al., *mixup: Beyond empirical risk minimization.* arXiv preprint arXiv:1710.09412, 2017.
213. Krizhevsky, A. and G. Hinton, *Learning multiple layers of features from tiny images.* 2009.
214. Netzer, Y., et al., *Reading digits in natural images with unsupervised feature learning.* 2011.
215. Bromley, J., et al. *Signature verification using a" siamese" time delay neural network*. in *Advances in neural information processing systems*. 1994.
216. Chopra, S., R. Hadsell, and Y. LeCun. *Learning a similarity metric discriminatively, with application to face verification*. in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. 2005. IEEE.
217. Hadsell, R., S. Chopra, and Y. LeCun. *Dimensionality reduction by learning an invariant mapping*. in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. 2006. IEEE.
218. Schroff, F., D. Kalenichenko, and J. Philbin. *Facenet: A unified embedding for face recognition and clustering*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
219. McLaughlin, N., J. Martinez del Rincon, and P. Miller. *Recurrent convolutional network for video-based person re-identification*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
220. Chung, D., K. Tahboub, and E.J. Delp. *A two stream siamese convolutional neural network for person re-identification*. in *Proceedings of the IEEE International Conference on Computer Vision*. 2017.
221. Varior, R.R., M. Haloi, and G. Wang. *Gated siamese convolutional neural network architecture for human re-identification*. in *European conference on computer vision*. 2016. Springer.
222. Guo, Y.-F., et al., *Null foley–sammon transform.* Pattern recognition, 2006. **39**(11): p. 2248-2251.
223. Zhang, L., T. Xiang, and S. Gong. *Learning a discriminative null space for person re-identification*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
224. Goodfellow, I., Y. Bengio, and A. Courville, *Deep learning*. 2016: MIT press.
225. Loshchilov, I. and F. Hutter, *Fixing weight decay regularization in adam.* 2018.
226. Zhang, G., et al., *Three mechanisms of weight decay regularization.* arXiv preprint arXiv:1810.12281, 2018.
227. Zagoruyko, S. and N. Komodakis, *Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer.* arXiv preprint arXiv:1612.03928, 2016.
228. Paszke, A., et al., *Automatic differentiation in pytorch.* 2017.
229. Abadi, M., et al. *Tensorflow: A system for large-scale machine learning*. in *12th*

*{USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. 2016.
230. McInnes, L., J. Healy, and J. Melville, *Umap: Uniform manifold approximation and projection for dimension reduction.* arXiv preprint arXiv:1802.03426, 2018.
231. Chapelle, O., B. Scholkopf, and A. Zien, *Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews].* IEEE Transactions on Neural Networks, 2009. **20**(3): p. 542-542.
232. Tschandl, P., C. Rosendahl, and H. Kittler, *The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions.* Scientific data, 2018. **5**: p. 180161.
233. Team, N.L.S.T.R., *The national lung screening trial: overview and study design.* Radiology, 2011. **258**(1): p. 243-253.
234. Roth, H.R., et al. *Anatomy-specific classification of medical images using deep convolutional nets*. in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. 2015. IEEE.
235. Frid-Adar, M., et al., *GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification.* Neurocomputing, 2018. **321**: p. 321-331.
236. Frid-Adar, M., et al. *Synthetic data augmentation using GAN for improved liver lesion classification*. in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. 2018. IEEE.
237. Wen, Y., et al. *A discriminative feature learning approach for deep face recognition*. in *European conference on computer vision*. 2016. Springer.
238. Chen, T., et al., *A simple framework for contrastive learning of visual representations.* arXiv preprint arXiv:2002.05709, 2020.
239. Iandola, F., et al., *Densenet: Implementing efficient convnet descriptor pyramids.* arXiv preprint arXiv:1404.1869, 2014.
240. Li, K.M. and E.C. Li, *Skin lesion analysis towards melanoma detection via end-to-end deep learning of convolutional neural networks.* arXiv preprint arXiv:1807.08332, 2018.
241. Liao, F., et al., *Evaluate the malignancy of pulmonary nodules using the 3-D deep leaky noisy-or network.* IEEE transactions on neural networks and learning systems, 2019. **30**(11): p. 3484-3495.
242. Gao, R., et al. *Distanced LSTM: Time-Distanced Gates in Long Short-Term Memory Models for Lung Cancer Detection*. in *International Workshop on Machine Learning in Medical Imaging*. 2019. Springer.
243. Cabezas, M., et al., *A review of atlas-based segmentation for magnetic resonance brain images.* Computer methods and programs in biomedicine, 2011. **104**(3): p. e158-e177.
244. Toga, A.W., *Brain warping*. 1998: Elsevier.
245. Gee, J.C., M. Reivich, and R. Bajcsy, *Elastically deforming a three-dimensional atlas to match anatomical brain images.* 1993.
246. Yeo, B.T., et al., *The organization of the human cerebral cortex estimated by intrinsic functional connectivity.* Journal of neurophysiology, 2011.
247. Sallet, J., et al., *The organization of dorsal frontal cortex in humans and macaques.* Journal of Neuroscience, 2013. **33**(30): p. 12255-12274.

248. Neubert, F.-X., et al., *Connectivity reveals relationship of brain areas for reward-guided learning and decision making in human and monkey frontal cortex.* Proceedings of the national academy of sciences, 2015. **112**(20): p. E2695-E2704.
249. Labache, L., et al., *A SENtence Supramodal Areas AtlaS (SENSAAS) based on multiple task-induced activation mapping and graph analysis of intrinsic connectivity in 144 healthy right-handers.* Brain Structure and Function, 2019. **224**(2): p. 859-882.
250. Kikinis, R., et al., *A digital brain atlas for surgical planning, model-driven segmentation, and teaching.* IEEE Transactions on visualization and computer graphics, 1996. **2**(3): p. 232-241.
251. Van Baarsen, K., et al., *A probabilistic atlas of the cerebellar white matter.* Neuroimage, 2016. **124**: p. 724-732.
252. Mazziotta, J.C., et al., *A probabilistic atlas of the human brain: theory and rationale for its development.* Neuroimage, 1995. **2**(2): p. 89-101.
253. Behrens, T.E., et al., *Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging.* Nature neuroscience, 2003. **6**(7): p. 750-757.
254. Ewert, S., et al., *Toward defining deep brain stimulation targets in MNI space: a subcortical atlas based on multimodal MRI, histology and structural connectivity.* Neuroimage, 2018. **170**: p. 271-282.
255. Ilinsky, I., et al., *Human motor thalamus reconstructed in 3D from continuous sagittal sections with identified subcortical afferent territories.* eneuro, 2018. **5**(3).
256. Keuken, M.C., et al., *Ultra-high 7T MRI of structural age-related changes of the subthalamic nucleus.* Journal of Neuroscience, 2013. **33**(11): p. 4896-4900.
257. Tziortzi, A.C., et al., *Imaging dopamine receptors in humans with [11C]-(+)-PHNO: dissection of D3 signal and anatomy.* Neuroimage, 2011. **54**(1): p. 264-277.
258. Mazziotta, J., et al., *A four-dimensional probabilistic atlas of the human brain.* Journal of the American Medical Informatics Association, 2001. **8**(5): p. 401-430.
259. Eickhoff, S.B., et al., *Testing anatomically specified hypotheses in functional imaging using cytoarchitectonic maps.* Neuroimage, 2006. **32**(2): p. 570-582.
260. Eickhoff, S.B., et al., *A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data.* Neuroimage, 2005. **25**(4): p. 1325-1335.
261. Talairach, J., *Co-planar stereotaxic atlas of the human brain-3-dimensional proportional system.* An approach to cerebral imaging, 1988.
262. Zhang, F., et al., *An anatomically curated fiber clustering white matter atlas for consistent white matter tract parcellation across the lifespan.* NeuroImage, 2018. **179**: p. 429-447.
263. Varentsova, A., S. Zhang, and K. Arfanakis, *Development of a high angular resolution diffusion imaging human brain template.* Neuroimage, 2014. **91**: p. 177-186.
264. Yendiki, A., et al., *Automated probabilistic reconstruction of white-matter pathways in health and disease using an atlas of the underlying anatomy.* Frontiers in neuroinformatics, 2011. **5**: p. 23.
265. Mori, S., et al., *Stereotaxic white matter atlas based on diffusion tensor imaging*

*in an ICBM template.* Neuroimage, 2008. **40**(2): p. 570-582.
266. Mori, S., et al., *MRI atlas of human white matter*. 2005: Elsevier.
267. Catani, M. and M. Thiebaut de Schotten, *Atlas of human brain connections*. 2015.
268. Oishi, K., et al., *Human brain white matter atlas: identification and assignment of common anatomical structures in superficial white matter.* Neuroimage, 2008. **43**(3): p. 447-457.
269. Román, C., et al., *Clustering of whole-brain white matter short association bundles using HARDI data.* Frontiers in neuroinformatics, 2017. **11**: p. 73.
270. Yeh, F.-C., et al., *Population-averaged atlas of the macroscale human structural connectome and its network topology.* NeuroImage, 2018. **178**: p. 57-68.
271. Yeh, F.-C. and W.-Y.I. Tseng, *NTU-90: a high angular resolution brain atlas constructed by q-space diffeomorphic reconstruction.* Neuroimage, 2011. **58**(1): p. 91-99.
272. Figley, T.D., et al., *Probabilistic white matter atlases of human auditory, basal ganglia, language, precuneus, sensorimotor, visual and visuospatial networks.* Frontiers in human neuroscience, 2017. **11**: p. 306.
273. Chenot, Q., et al., *A population-based atlas of the human pyramidal tract in 410 healthy participants.* Brain Structure and Function, 2019. **224**(2): p. 599-612.
274. Archer, D.B., D.E. Vaillancourt, and S.A. Coombes, *A template and probabilistic atlas of the human sensorimotor tracts using diffusion MRI.* Cerebral Cortex, 2018. **28**(5): p. 1685-1699.
275. Rojkova, K., et al., *Atlasing the frontal lobe connections and their variability due to age and education: a spherical deconvolution tractography study.* Brain Structure and Function, 2016. **221**(3): p. 1751-1766.
276. de Schotten, M.T., et al., *Monkey to human comparative anatomy of the frontal lobe association tracts.* Cortex, 2012. **48**(1): p. 82-96.
277. Garyfallidis, E., et al., *Recognition of white matter bundles using local and global streamline-based registration and clustering.* NeuroImage, 2018. **170**: p. 283-295.
278. Guevara, P., et al., *Automatic fiber bundle segmentation in massive tractography datasets using a multi-subject bundle atlas.* Neuroimage, 2012. **61**(4): p. 1083-1099.
279. Schilling, K.G., et al., *A fiber coherence index for quality control of B-table orientation in diffusion MRI scans.* Magnetic resonance imaging, 2019. **58**: p. 82-89.
280. Wasserthal, J., P. Neher, and K.H. Maier-Hein, *Tractseg-fast and accurate white matter tract segmentation.* NeuroImage, 2018. **183**: p. 239-253.
281. Warrington, S., et al., *XTRACT-Standardised protocols for automated tractography in the human and macaque brain.* NeuroImage, 2020: p. 116923.
282. Yeatman, J.D., et al., *Tract profiles of white matter properties: automating fiber-tract quantification.* PloS one, 2012. **7**(11).
283. Wasserthal, J., et al., *Combined tract segmentation and orientation mapping for bundle-specific tractography.* Medical image analysis, 2019. **58**: p. 101559.
284. Tournier, J.-D., et al., *MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation.* NeuroImage, 2019. **202**: p. 116137.
285. Garyfallidis, E., et al., *Dipy, a library for the analysis of diffusion MRI data.* Frontiers in neuroinformatics, 2014. **8**: p. 8.

286. Destrieux, C., et al., *Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature.* Neuroimage, 2010. **53**(1): p. 1-15.
287. Fischl, B., M.I. Sereno, and A.M. Dale, *Cortical surface-based analysis: II: inflation, flattening, and a surface-based coordinate system.* Neuroimage, 1999. **9**(2): p. 195-207.
288. Dale, A.M., B. Fischl, and M.I. Sereno, *Cortical surface-based analysis: I. Segmentation and surface reconstruction.* Neuroimage, 1999. **9**(2): p. 179-194.
289. Jenkinson, M., et al., *Fsl.* Neuroimage, 2012. **62**(2): p. 782-90.
290. Wakana, S., et al., *Reproducibility of quantitative tractography methods applied to cerebral white matter.* Neuroimage, 2007. **36**(3): p. 630-644.
291. Fischl, B., *FreeSurfer.* Neuroimage, 2012. **62**(2): p. 774-781.
292. Smith, S.M., et al., *Advances in functional and structural MR image analysis and implementation as FSL.* Neuroimage, 2004. **23**: p. S208-S219.
293. Bürgel, U., et al., *White matter fiber tracts of the human brain: three-dimensional mapping at microscopic resolution, topography and intersubject variability.* Neuroimage, 2006. **29**(4): p. 1092-1105.
294. de Schotten, M.T., et al., *Atlasing location, asymmetry and inter-subject variability of white matter tracts in the human brain with MR diffusion tractography.* Neuroimage, 2011. **54**(1): p. 49-59.
295. Rheault, F., et al., *Bundle-specific tractography*, in *Computational Diffusion MRI*. 2018, Springer. p. 129-139.
296. Rheault, F., et al., *Bundle-specific tractography with incorporated anatomical and orientational priors.* NeuroImage, 2019. **186**: p. 382-398.
297. Filley, C.M., *White matter dementia.* Therapeutic advances in neurological disorders, 2012. **5**(5): p. 267-277.
298. Lazar, M., *Mapping brain anatomical connectivity using white matter tractography.* NMR in Biomedicine, 2010. **23**(7): p. 821-835.
299. Bammer, R., B. Acar, and M.E. Moseley, *In vivo MR tractography using diffusion imaging.* European journal of radiology, 2003. **45**(3): p. 223-234.
300. Mandonnet, E., S. Sarubbo, and L. Petit, *The nomenclature of human white matter association pathways: proposal for a systematic taxonomic anatomical classification.* Frontiers in neuroanatomy, 2018. **12**: p. 94.
301. O'Muircheartaigh, J., et al., *White matter development and early cognition in babies and toddlers.* Human brain mapping, 2014. **35**(9): p. 4475-4487.
302. Filley, C.M. and R.D. Fields, *White matter and cognition: making the connection*. Journal of neurophysiology, 2016. **116**(5): p. 2093-2104.
303. Hirono, N., et al., *Impact of white matter changes on clinical manifestation of Alzheimer's disease: a quantitative study.* Stroke, 2000. **31**(9): p. 2182-2188.
304. Xie, S., et al., *How does B-value affect HARDI reconstruction using clinical diffusion MRI data?* PloS one, 2015. **10**(3): p. e0120773.
305. O'Donnell, L.J. and C.-F. Westin, *Automatic tractography segmentation using a high-dimensional white matter atlas.* IEEE transactions on medical imaging, 2007. **26**(11): p. 1562-1575.
306. Hansen, C.B., et al., *Pandora: 4-D white matter bundle population-based atlases derived from diffusion MRI fiber tractography.* Neuroinformatics, 2020: p. 1-14.
307. Çiçek, Ö., et al. *3D U-Net: learning dense volumetric segmentation from sparse*

*annotation*. in *International conference on medical image computing and computer-assisted intervention*. 2016. Springer.
308. Milletari, F., N. Navab, and S.-A. Ahmadi. *V-net: Fully convolutional neural networks for volumetric medical image segmentation*. in *2016 fourth international conference on 3D vision (3DV)*. 2016. IEEE.
309. Huo, Y., et al., *3D whole brain segmentation using spatially localized atlas network tiles.* NeuroImage, 2019. **194**: p. 105-119.
310. Bao, S. and A.C. Chung, *Multi-scale structured CNN with label consistency for brain MR image segmentation.* Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 2018. **6**(1): p. 113-117.
311. Ferrucci, L., *The Baltimore Longitudinal Study of Aging (BLSA): a 50-year-long journey and plans for the future*. 2008, Oxford University Press.
312. Cai, L.Y., et al., *PreQual: An automated pipeline for integrated preprocessing and quality assurance of diffusion weighted MRI images.* Magnetic Resonance in Medicine, 2021. **86**(1): p. 456-470.
313. Fonov, V., et al., *Unbiased average age-appropriate atlases for pediatric studies.* Neuroimage, 2011. **54**(1): p. 313-327.
314. Coupé, P., et al. *AssemblyNet: A novel deep decision-making process for whole brain MRI segmentation*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2019. Springer.
315. Heimann, T., et al., *Comparison and evaluation of methods for liver segmentation from CT datasets.* IEEE transactions on medical imaging, 2009. **28**(8): p. 1251-1265.
316. Wilcoxon, F., *Individual comparisons by ranking methods*, in *Breakthroughs in statistics*. 1992, Springer. p. 196-202.
317. Lorensen, W.E. and H.E. Cline, *Marching cubes: A high resolution 3D surface construction algorithm.* ACM siggraph computer graphics, 1987. **21**(4): p. 163-169.
318. Mickevicius, N.J., et al., *Location of brain tumor intersecting white matter tracts predicts patient prognosis.* Journal of neuro-oncology, 2015. **125**(2): p. 393-400.
319. Karnath, H.-O., C. Sperber, and C. Rorden, *Mapping human brain lesions and their functional consequences.* Neuroimage, 2018. **165**: p. 180-189.
320. Dewey, B.E., et al. *A Disentangled Latent Space for Cross-Site MRI Harmonization*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2020. Springer.
321. Modanwal, G., et al. *MRI image harmonization using cycle-consistent generative adversarial network*. in *Medical Imaging 2020: Computer-Aided Diagnosis*. 2020. International Society for Optics and Photonics.
322. Palladino, J.A., D.F. Slezak, and E. Ferrante. *Unsupervised domain adaptation via CycleGAN for white matter hyperintensity segmentation in multicenter MR images*. in *16th International Symposium on Medical Information Processing and Analysis*. 2020. International Society for Optics and Photonics.
323. Bashyam, V.M., et al., *Medical Image Harmonization Using Deep Learning Based Canonical Mapping: Toward Robust and Generalizable Learning in Imaging.* arXiv preprint arXiv:2010.05355, 2020.
324. Merlet, S.L. and R. Deriche, *Continuous diffusion signal, EAP and ODF*

*estimation via compressive sensing in diffusion MRI.* Medical image analysis, 2013. **17**(5): p. 556-572.
325. Nath, V., et al. *Enabling multi-shell b-value generalizability of data-driven diffusion models with deep SHORE*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2019. Springer.
326. Jenkinson, M., et al., *FSL*. Neuroimage, 2012. **62**: p. 782-90.