



## Cross-scanner and cross-protocol multi-shell diffusion MRI data harmonization: Algorithms and results

Lipeng Ning<sup>a,b,\*</sup>, Elisenda Bonet-Carne<sup>c</sup>, Francesco Grussu<sup>c</sup>, Farshid Sepehrband<sup>d</sup>, Enrico Kaden<sup>c</sup>, Jelle Veraart<sup>e</sup>, Stefano B. Blumberg<sup>c</sup>, Can Son Khoo<sup>c</sup>, Marco Palombo<sup>c</sup>, Iasonas Kokkinos<sup>c</sup>, Daniel C. Alexander<sup>c</sup>, Jaume Coll-Font<sup>f,b</sup>, Benoit Scherrer<sup>f,b</sup>, Simon K. Warfield<sup>f,b</sup>, Suheyla Cetin Karayumak<sup>a,b</sup>, Yogesh Rath<sup>a,b</sup>, Simon Koppers<sup>g</sup>, Leon Weninger<sup>g</sup>, Julia Ebert<sup>g</sup>, Dorit Merhof<sup>g</sup>, Daniel Moyer<sup>d</sup>, Maximilian Pietsch<sup>h</sup>, Daan Christiaens<sup>h,o</sup>, Rui Azeredo Gomes Teixeira<sup>h</sup>, Jacques-Donald Tournier<sup>h</sup>, Kurt G. Schilling<sup>i</sup>, Yuankai Huo<sup>j</sup>, Vishwesh Nath<sup>j</sup>, Colin Hansen<sup>j</sup>, Justin Blaber<sup>j</sup>, Bennett A. Landman<sup>i,j,k</sup>, Andrey Zhylka<sup>l</sup>, Josien P.W. Pluim<sup>l</sup>, Greg Parker<sup>m</sup>, Umesh Rudrapatna<sup>m</sup>, John Evans<sup>m</sup>, Cyril Charron<sup>m</sup>, Derek K. Jones<sup>m,n</sup>, Chantal M.W. Tax<sup>m</sup>

<sup>a</sup> Brigham and Women's Hospital, Boston, United States

<sup>b</sup> Harvard Medical School, Boston, United States

<sup>c</sup> University College London, London, United Kingdom

<sup>d</sup> Stevens Neuroimaging and Informatics Institute, Keck School of Medicine of USC, University of Southern California, Los Angeles, United States

<sup>e</sup> New York University, New York, NY, United States

<sup>f</sup> Boston Children's Hospital, Boston, United States

<sup>g</sup> RWTH Aachen University, Aachen, Germany

<sup>h</sup> Centre for the Developing Brain, School of Biomedical Engineering and Imaging Sciences, King's College London, London, United Kingdom

<sup>i</sup> Institute of Imaging Science, Vanderbilt University, Nashville, TN, United States

<sup>j</sup> Department of Electrical Engineering & Computer Science, Vanderbilt University, Nashville, TN, United States

<sup>k</sup> Department of Biomedical Engineering, Vanderbilt University, Nashville, TN, United States

<sup>l</sup> Eindhoven University of Technology, Eindhoven, Netherlands

<sup>m</sup> Cardiff University Brain Research Imaging Centre (CUBRIC), Cardiff University, Cardiff, United Kingdom

<sup>n</sup> School of Psychology, Australian Catholic University, Melbourne, Australia

<sup>o</sup> Department of Electrical Engineering (ESAT/PSI), KU Leuven, Leuven, Belgium

### ARTICLE INFO

#### Keywords:

Multi-shell diffusion MRI

Harmonization

Spherical harmonics

Deep learning

Regression

### ABSTRACT

Cross-scanner and cross-protocol variability of diffusion magnetic resonance imaging (dMRI) data are known to be major obstacles in multi-site clinical studies since they limit the ability to aggregate dMRI data and derived measures. Computational algorithms that harmonize the data and minimize such variability are critical to reliably combine datasets acquired from different scanners and/or protocols, thus improving the statistical power and sensitivity of multi-site studies. Different computational approaches have been proposed to harmonize diffusion MRI data or remove scanner-specific differences. To date, these methods have mostly been developed for or evaluated on single b-value diffusion MRI data. In this work, we present the evaluation results of 19 algorithms that are developed to harmonize the cross-scanner and cross-protocol variability of multi-shell diffusion MRI using a benchmark database. The proposed algorithms rely on various signal representation approaches and computational tools, such as rotational invariant spherical harmonics, deep neural networks and hybrid biophysical and statistical approaches. The benchmark database consists of data acquired from the same subjects on two scanners with different maximum gradient strength (80 and 300 mT/m) and with two protocols. We evaluated the performance of these algorithms for mapping multi-shell diffusion MRI data across scanners and across protocols using several state-of-the-art imaging measures. The results show that data harmonization algorithms can reduce the cross-scanner and cross-protocol variabilities to a similar level as scan-rescan variability using the same

\* Corresponding author. Brigham and Women's Hospital, Boston, United States

E-mail address: [lning@bwh.harvard.edu](mailto:lning@bwh.harvard.edu) (L. Ning).

<https://doi.org/10.1016/j.neuroimage.2020.117128>

Received 1 October 2019; Received in revised form 26 June 2020; Accepted 29 June 2020

Available online 13 July 2020

1053-8119/© 2020 Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

scanner and protocol. In particular, the LinearRISH algorithm based on adaptive linear mapping of rotational invariant spherical harmonics features yields the lowest variability for our data in predicting the fractional anisotropy (FA), mean diffusivity (MD), mean kurtosis (MK) and the rotationally invariant spherical harmonic (RISH) features. But other algorithms, such as DIAMOND, SHResNet, DIQT, CMResNet show further improvement in harmonizing the return-to-origin probability (RTOP). The performance of different approaches provides useful guidelines on data harmonization in future multi-site studies.

## 1. Introduction

Diffusion magnetic resonance imaging provides information to characterize tissue microstructure by probing the diffusive displacements of water molecules. It is increasingly used to investigate the structural connections in human brains and tissue abnormalities related to disorders (Kubicki et al., 2007; Pasternak et al., 2018; Mueller et al., 2005). Recent advances in acquisition protocols with multiple b-values provide imaging measures that are even more sensitive and/or specific than standard approaches based on single b-value data (Jensen et al., 2005; Özarslan et al., 2013; Ning et al., 2015). On the other hand, dMRI scans have intrinsic variability caused by various factors including but not limited to scanner field and gradient strength and acquisition protocols (Vollmar et al., 2010; Grech-Sollars et al., 2015; Veenith et al., 2013; Landman et al., 2011). In particular, the variability of imaging measures between datasets acquired from different scanners with different protocols could be much more significant than the intra-scanner variability between two data repetitions acquired from the same subject using the same protocol at a different time (Karayumak et al., 2019). Therefore, it is imperative to reduce the cross-scanner and cross-protocol variability in order to reliably aggregate multi-site databases for increasing statistical power and sensitivity of studies.

In view of the significance of cross-scanner and cross-protocol variability, several types of methods have been recently developed to reliably combine dMRI datasets in multi-site studies (Karayumak et al., 2019; Moyer et al., 2020; Blumberg et al., 2018, 2019; Mirzaalian et al., 2018; Fortin et al., 2017; Mirzaalian et al., 2016; Pohl et al., 2016). Some approaches are able to enhance the spatial or angular resolution or other imaging features of dMRI data with low-b values to match the data quality of the state-of-the-art scanners and protocols (Grech-Sollars et al., 2015; Zhu et al., 2011; Vollmar et al., 2010). However, these approaches were designed or implemented to retrospectively harmonize the statistical differences in imaging measures between groups of subjects from different sites with matched age, gender, handedness and other socio-economic factors. A more ideal approach to evaluate cross-scanner and cross-protocol variability is to use datasets from the different scanners of the same subjects. This enables accurate measurement of the intra-subject reproducibility and cross-scanner variability that is not provided by retrospective analysis using group data. In recent years, the intra-scanner reproducibility has been investigated using test and retest scans from the same subjects using the same scanner (Shahim et al., 2017; Duan et al., 2015; Boekel et al., 2017) or using scanners with different field strengths (1.5T vs 3T and 3T vs 7T). We here specifically seek to understand scanner and/or protocol related variability in current multi-site studies based on 3T scanners.

To facilitate research on cross-scanner and cross-protocol variability, a benchmark database was provided in (Tax et al., 2019) with data acquired from 3 different scanners using 5 different acquisition protocols. An open competition was organized during the Computational Diffusion MRI (CDMRI) workshop at 20th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2017) which invited participants to develop algorithms to solve a dMRI harmonization task which is to find a mapping between datasets acquired using different scanners and/or protocols that were acquired with the same b-value. The result of the competition was reported in (Tax et al., 2019). Most algorithms, e.g. (Koppers et al., 2018), used spherical harmonics to represent single-shell dMRI signals and applied convolutional

neural networks (CNNs) to learn mapping between datasets.

While the majority of previous works focused on the harmonization of single-shell diffusion MRI, this work presents the evaluations of harmonization algorithms for multi-shell diffusion data. In particular, this paper is a continuation and extension of the results in (Tax et al., 2019), and reports on the “Multi-Shell Harmonization and enhancement Challenge” (MUSHAC) (Ning et al., 2019), part of CDMRI at MICCAI 2018. We report the summary and results of the second open competition on harmonizing multi-shell dMRI datasets. The database used in this competition consists of four acquisitions of the same 15 healthy participants scanned on two MRI systems with different maximum gradient strength (80 and 300 mT/m) and different protocols. Different from the earlier work that focused on single-shell data, this paper reports the results of multi-shell comparison from a significantly wider range of algorithms (19 algorithms from 9 research groups). Such systematic evaluations (Klein et al., 2009) have the potential to provide important guidelines for further multi-site studies on diffusion MRI.

Results demonstrate that state-of-the-art computational tools, e.g. CNN and bio-statistical approaches, enable accurate harmonization of multi-shell diffusion MRI across scanners and protocols, reducing between-scanner/between-protocol variability to levels comparable to scan-rescan variability. These results pave the way towards reliable combination of multi-center data in large clinical study, which is key in the area of big data.

## 2. Methods

The acquisition parameters and preprocessing approaches are described in Section 2.1. Section 2.2 introduces the tasks and rules of the MUSHAC challenge on multi-shell dMRI harmonization. Section 2.3 provides the evaluation strategies and metrics. Section 2.4 summarizes the algorithms participating in the open competition.

### 2.1. Data

The database was acquired from 15 healthy volunteers, which includes all the subjects reported in Table 1 of (Tax et al., 2019). The data acquisition was approved by Cardiff University School of Psychology ethics committee. All subjects were scanned using three different scanners with different maximum gradient strength. This work focuses on the multi-shell database acquired from the two scanners with relative higher maximum gradient strength: a) 3T Siemens Prisma (80 mT/m), and b) 3T Siemens Connectom (300 mT/m). The average time between acquisitions on scanners a) and b) was 1 month. The scanners had no software upgrades during the course of the study.

Two types of protocols were acquired from each scanner: 1) a more ‘standard’ (ST) protocol with acquisition parameters matched to a typical clinical protocol; and 2) a more advanced or ‘state-of-the-art’ (SA) protocol where the superior hardware and software specifications were utilized to increase the number of acquisitions and spatial resolution per unit time. The ST data from both scanners have an isotropic resolution of 2.4 mm, TE = 89 ms and TR = 7.2 s. Both ST datasets have 30 directions at b = 1200, 3000 s/mm<sup>2</sup>. On the other hand, the Prisma-SA data has a higher isotropic resolution of 1.5 mm, TE = 80 ms, TR = 7.1s and 60 directions at the same b-values while the Connectom-SA data has the highest resolution of 1.2 mm with TE = 68 ms, TR = 5.4 s and 60 directions. Reversed phase encoding, i.e. AP and PA, b0 images were

acquired in both protocols. Additional  $b = 0 \text{ s/mm}^2$  images were also acquired with matched TE and TR across scanners. Moreover, structural T1-weighted (T1w) images were acquired on both scanners using an MPRAGE sequence with  $1 \text{ mm}^3$  isotropic voxel. More detailed information about the acquisition parameters can be found in Table 2 of (Tax et al., 2019).

The data preprocessing approach includes the following step: The b0 volumes were corrected for EPI distortions by applying FSL TOPUP on reversed phase-encoding pairs (Andersson et al., 2003). The data was corrected for eddy current induced distortions and gradient-nonlinearity distortions (Glasser et al., 2013) with FSL TOPUP/eddy and in-house software kindly provided by Massachusetts General Hospital (MGH) (Andersson and Sotiropoulos, 2016). Spatiotemporally varying b-vectors and b-values due to gradient nonlinearities of the Connectom scanner were made available (Bammer et al., 2003; Sotiropoulos et al., 2013; Rudrapatna et al., 2018). The dMRI of Prisma-SA and Connectom-SA were affinely registered to Prisma-ST using the corresponding fractional anisotropy (FA) maps with FSL FLIRT (Jenkinson and Smith, 2001) and using sinc interpolation and appropriate b-vector rotation. Prisma-ST was used as the reference space since additional geometric distortions can still be present in Connectom data because of gradient nonlinearities (Setsompop et al., 2013; Jones et al., 2018; Tax et al., 2019). The T1w images were registered to the FA of Prisma-ST to avoid extra interpolation of the dMRI data.

## 2.2. Competition tasks

The processed data from 10 randomly selected subjects were used as training datasets which were released to the participants of the MUSHAC. The Prisma-ST scans and the corresponding T1w images of the remaining

5 subjects were used as a testing set. The tasks were to predict the other three datasets using the provided Prisma-ST scans. In particular, the following two tasks were evaluated:

- 1) **Harmonization:** the prediction of dMRI signals of the Connectom-ST dMRI scans using the provided Prisma-ST data;
- 2) **Enhancement:** the prediction of dMRI signals of the state-of-the-art dMRI scans, including Prisma-SA (Task 2a) and Connectom-SA (Task 2b) scans, using the provided Prisma-ST data.

All participants of the dMRI harmonization competition were required to at least complete Task 1, but Task 2 was optional. The preliminary results of the 12 algorithms were announced during the CDMRI workshop at MICCAI 2018 (Ning et al., 2019), as this provided useful feedback regarding the algorithm performance and capabilities to the participants. Then, the participants were allowed to re-submit their updated results or new results after the workshop. New participants were also allowed to enter the competition without providing preliminary evaluations of their algorithms. By the final deadline of the competition, 19 algorithms implemented by 9 different research groups were submitted for evaluation on the test datasets.

## 2.3. Evaluation

Harmonization algorithms had to predict the dMRI signals of three missing datasets by using the provided b-values, b-vectors and the Prisma-ST scans. From the predicted and the underlying ground-truth datasets, diffusion tensor models were estimated using the dMRI signal at  $b = 0, 1200 \text{ s/mm}^2$  using a weighted least-squares estimate provided by the dtfit command in FSL (Jenkinson et al., 2012). The fractional

**Table 1**  
Summary of harmonization algorithms evaluated.

Type	Algorithm name	Additional processing	Signal representation	Algorithm details	Tasks
Interpolation	SHInterp	Nonlinear registration, 4th order SH	SH	Laplace-Beltrami regularization, cubic interpolation	1, 2a, 2b
	DIAMOND-a	Noise filtering	DIAMOND	Sinc interpolation, central and non-central	1, 2a, 2b
	DIAMOND-b			DIAMOND models, constrained estimation	
	DIAMOND-c				
	DIAMOND-d				
Regression	DIAMOND-a-NL	Noise filtering	DIAMOND	Same as above but with spatially varying gradients	1, 2a, 2b
	DIAMOND-b-NL				
	DIAMOND-c-NL				
	SHOREInterp	Interpolation	SHORE	Signal prediction using SHORE model	1, 2a, 2b
	LinearRISH	unringing, 6th order SH	SH	Subject-specific training data	1, 2a, 2b
Convolutional Neural Networks	TWRF-a	Tissue segmentation	DTI	Random forest regression with full and reduced feature sets	1
	TWRF-b		Mean signals		
	SH-1Layer	Interpolation, 4th order SH	SH	A single convolutional layer	1, 2a, 2b
	SHResNet-a	Nonlinear registration, tissue segmentation	SH	ResNet, Anatomical constraints, Spatially varying gradients (-NL)	1, 2a, 2b
	SHResNet-a-NL				
	SHResNet-b0	Interpolation, 4th order SH	SH	ResNet, One-hot orthogonal vectors of tissue labels, Spatial coordinates	1, 2a, 2b
	SHResNet-b1				
	SHResNet-b2				
	SHResNet-b3				
	SHResNet-b4				
Networks	SphericalNet	Nonlinear registration	SH	Local Spherical Convolution Network, Anatomical constraints,	1, 2a, 2b
	SphericalNet-NL	Tissue segmentation			
	Tinypatch	Tissue segmentation	SH	Patch-wise feed forward with auto-encoder pre-training	1, 2a, 2b
	DIQT	4th order SH	SH	Patch selection based on center voxels	1, 2a, 2b
	CMResNet-a	Unringing, inhomogeneity corrections, normalization,	SHARD	ResNet networks,	1
	CMResNet-b	nonlinear registration		Two decoder networks	1, 2a

**Table 2**Summary of the percentage APE values, i.e. APE<sub>100%</sub>, of reported measures in global evaluations for Task1.

	APE (%): Mean (Standard deviation)							
	FA	MD	R0 (b = 1200)	R2 (b = 1200)	R0 (b = 3000)	R2 (b = 3000)	MK	RTOP
Reference	16.7 (2.0)	8.2 (1.2)	15.5 (2.0)	36.4 (3.6)	19.1 (2.1)	40.2 (4.3)	11.0 (1.0)	12.4 (1.4)
SHInterp	16.5 (11.2)	8.1 (5.7)	14.2 (10.0)	32.4 (20.8)	17.5 (12.0)	35.4 (22.8)	10.7 (8.7)	12.4 (8.2)
DIAMOND-a	15.9 (10.5)	6.6 (4.6)	11.5 (7.9)	30.9 (18.5)	16.5 (10.1)	32.3 (20.8)	9.1 (6.5)	11.7 (7.2)
DIAMOND-b	15.9 (10.5)	6.6 (4.6)	11.6 (7.9)	30.8 (18.6)	16.4 (10.1)	32.3 (21.1)	9.0 (6.6)	11.5 (7.2)
SHOREInterp	16.5 (12.1)	8.1 (5.7)	14.2 (10.0)	34.1 (23.8)	17.4 (12.0)	36.8 (25.2)	10.9 (8.9)	12.4 (8.2)
LinearRISH	6.0 (4.1)	2.7 (1.8)	4.7 (3.2)	11.8 (7.8)	6.0 (4.0)	12.9 (8.5)	3.7 (2.9)	14.6 (7.2)
TWRF-a	28.7 (15.9)	10.7 (7.0)	18.4 (11.4)	49.1 (22.8)	21.3 (12.5)	47.6 (22.3)	12.8 (9.7)	15.2 (9.4)
TWRF-b	27.9 (15.7)	9.6 (6.6)	17.1 (11.5)	47.8 (22.9)	19.2 (12.4)	40.9 (22.4)	11.8 (9.3)	13.3 (8.5)
SH-1Layer	14.4 (10.0)	6.0 (4.1)	10.4 (7.1)	28.7 (18.6)	14.0 (9.5)	33.2 (19.3)	8.4 (6.3)	10.2 (6.7)
SHResNet-a	15.7 (10.1)	6.4 (4.4)	11.3 (7.9)	30.0 (17.9)	14.6 (10.1)	32.4 (19.6)	8.4 (6.4)	10.4 (6.9)
SHResNet-b0	19.3 (12.9)	8.8 (6.3)	15.0 (10.4)	35.9 (21.5)	19.6 (12.9)	37.5 (21.3)	9.2 (6.7)	14.3 (9.1)
SHResNet-b1	20.2 (12.8)	8.5 (5.6)	14.8 (10.0)	37.9 (21.5)	20.0 (14.1)	39.0 (24.0)	18.4 (14.2)	14.2 (9.5)
SHResNet-b2	17.0 (11.5)	7.8 (5.5)	13.6 (9.4)	32.9 (20.3)	17.7 (12.5)	35.0 (22.2)	16.8 (13.1)	12.7 (8.5)
SHResNet-b3	20.0 (13.3)	8.3 (5.9)	14.4 (10.0)	36.9 (21.7)	20.1 (13.6)	39.7 (23.5)	16.3 (12.5)	14.5 (9.5)
SHResNet-b4	16.1 (11.0)	7.4 (5.1)	12.8 (8.7)	31.5 (19.8)	16.4 (10.6)	34.8 (20.1)	9.1 (6.7)	11.9 (7.5)
SphericalNet	16.1 (10.2)	6.4 (4.4)	11.2 (7.8)	30.6 (18.0)	14.4 (9.9)	32.5 (19.4)	8.4 (6.5)	10.4 (6.8)
Tinypatch	19.4 (12.7)	6.5 (4.3)	11.4 (7.8)	36.5 (21.3)	15.4 (10.2)	38.4 (21.9)	9.2 (6.5)	11.2 (7.2)
DIQT	13.8 (9.4)	6.6 (4.4)	11.4 (7.1)	26.6 (16.7)	13.7 (8.9)	29.6 (18.2)	8.6 (6.8)	9.8 (6.2)
CMResNet-a	17.1 (10.7)	6.6 (4.6)	11.6 (7.9)	32.2 (18.3)	15.6 (10.0)	35.1 (19.5)	8.1 (5.9)	11.6 (7.2)
CMResNet-b	15.9 (10.6)	6.1 (4.1)	10.9 (7.5)	30.4 (18.2)	14.0 (9.5)	33.2 (19.6)	7.6 (5.7)	10.0 (6.5)

anisotropy (FA) and mean diffusivity (MD) measures were subsequently extracted for each voxel. Moreover, two multi-shell dMRI measures, including the mean kurtosis (MK) provided by the diffusion kurtosis imaging (DKI) model (Jensen et al., 2005) and the return-to-origin probability (RTOP) measure from MAP-MRI (<https://paperpile.com/c/XYw0oa/FzQnK> Özarlan et al., 2013), were computed using multi-shell diffusion data. The FA, MD, MK and RTOP measures were all computed using the Dipy toolbox (Garyfallidis et al., 2014). In addition, the zeroth and second order rotationally invariant spherical harmonic (RISH) features, R0 and R2 (Mirzaalian et al., 2016), were computed for the normalized dMRI signals at b = 1200, 3000 s/mm<sup>2</sup> to measure the angular frequency of dMRI signals at each b-value shell. The RISH features were computed using an in-house MATLAB (Mathworks, Natick, MA) toolbox.

The estimated imaging measures of the predicted datasets were compared against the corresponding features derived from the true acquired data. The performances of algorithms were evaluated in brain regions specified by brain masks excluding the cerebellum, which was not always included within the foot-head field-of-view coverage. Such an evaluation mask was obtained with the Geodesic Information Flow (GIF) algorithm (Cardoso et al., 2015), which was used to segment different brain areas in the anatomical T1w images. The evaluation masks were provided to the participants. Then, the absolute-value of the percentage error (APE) between the predicted and ground-truth measures, i.e.  $APE = |(\text{predicted} - \text{ground-truth})/\text{ground-truth}|$ , were computed on a voxel level. Subsequently, the mean and standard deviation of APE were computed globally in a brain mask excluding cerebellum, and regionally in brain regions provided by FreeSurfer (Fischl, 2012).

## 2.4. Algorithms

Participants of the open competition developed 19 algorithms to solve the harmonization challenge (Task 1: the prediction of the Connectom-ST data using the Prisma-ST data). A subset of 19 algorithms were also implemented to solve the enhancement challenge (Task 2: the prediction of Prisma-SA data (Task: 2a) and Connectom-SA data (Task: 2b) using the Prisma-ST data). For Task 1, the Prisma-ST and the Connectom-ST scans have exactly the same acquisition parameters, including the spatial resolution and the gradient directions. Therefore, the voxel-wise differences in imaging features between the two datasets can be directly computed without using any interpolation algorithms. The corresponding APE value reflects the true cross-scanner differences which can be considered as the *reference* when evaluating the

performance of harmonization algorithms since no harmonization algorithms were applied. Since the voxel size of Prisma-ST scans are different from those of Prisma-SA and Connectom-SA scans, no natural reference is available for Task 2a and Task 2b. But the performance of interpolation-based algorithms can be considered as the baseline to evaluate other more complex algorithms. A detailed description of the participating algorithms is provided below and a brief summary is presented in Table 1.

### 2.4.1. Spherical-harmonic based interpolation (SHInterp)

Additional nonlinear registrations based on ANTs (Avants et al., 2011) were applied to align the Prisma-ST datasets to the other three scans with B-spline interpolation methods. Then the MPAGE images were applied to obtain brain segmentations using the FSL FAST (Jenkinson et al., 2012) and to obtain brain masking using FSL BET command (Jenkinson et al., 2005). The normalized diffusion signals in white matter (WM) and gray matter (GM) voxels within the brain mask were transformed to SH coefficients with order four and Laplace-Beltrami regularization of  $\epsilon = 0.006$ .

The SHInterp approach simply transformed the SH coefficients back to diffusion signals along the set of gradient directions of the to-be-predicted scans. Thus, due to SH fitting, it is analogous to a spatial smoothing or regularization in the angular domain of each shell. Since this approach didn't remove any variability in the data, its performance is expected to reflect the true cross-scanner and/or cross-protocol variability of the datasets. For Task 2, standard cubic interpolation was utilized to increase the spatial resolution.

### 2.4.2. DIAMOND-based approach (DIAMOND-a -b, -c and -d)

This approach applied two additional preprocessing steps for noise filtering and data up-sampling. In particular, noise filtering was done by using the low-rank matrix approximation approach in (Veraart et al., 2016) followed by correcting for the Gibbs ringing approach (Kellner et al., 2016). The up-sampling approach was applied to interpolate the data with isotropic voxels of size 1 mm using the ITK with "sinc" interpolation (Meijering et al., 1999).

The signal prediction approach was developed based on the hybrid biophysical and statistical DIAMOND method (Scherrer et al., 2016, 2017), which fits a probabilistic model that characterizes diffusion signals with a multi-compartment model. DIAMOND models each compartment in each voxel with a continuous statistical distribution of diffusion tensors (Scherrer et al., 2016), the expectation of which characterizing the average 3-D diffusivity rates of the compartment and the



concentration of which capturing the intra-compartment microstructural heterogeneity. The number of fascicle compartments at each voxel was automatically determined using AICU model order selection method (maximum: 3 fascicle compartments). All voxels also included an isotropic diffusion compartment to characterize free-water diffusion.

Two algorithms, denoted by DIAMOND-a and DIAMOND-b, based on the central DIAMOND (Scherrer et al., 2016, 2017) and the non-central DIAMOND models (Scherrer et al., 2016, 2017) were used to characterize diffusion signals, the latter allowing for the separate modeling of axial and radial heterogeneity. In both algorithms, the sum of model coefficients was constrained to be 1. The participant team also submitted two other algorithms based on central DIAMOND and non-central DIAMOND without the constraint on the sum of coefficients. But their experimental results are very similar to the results of DIAMOND-a and DIAMOND-b. Thus, the results without the constraints are not reported in this work.

Based on the estimated model coefficients using the Prisma-ST dataset, the same signal expressions were applied to predict diffusion signals using the b-vectors and b-values of the other three scans. No additional steps were applied to remove potential variability among the datasets from different scanners and protocols.

#### 2.4.3. SHORE-based interpolation (SHOREInterp)

First, the Prisma-ST scans were interpolated to match the spatial resolution of the other three scans to be predicted. Then the Simple Harmonic Oscillator based Reconstruction and Estimation (SHORE) model (Merlet and Deriche, 2013; Cheng et al., 2011) was used to represent the normalized multi-shell diffusion signals. More specifically, SHORE was used at the 4th order and regularized least squares were used for fitting as described in (Merlet et al., 2013; Nath et al., 2019). The regularization hyper-parameters were both set to  $1e-8$  and 'zeta' at  $700 (\text{mm}^{-2})$  as suggested in (Merlet et al., 2013; Nath et al., 2019). The model parameters estimated using the Prisma-ST datasets were used to reconstruct diffusion signals of the other three scans using the respective b-values and b-vectors. No other approaches were used to further reduce potential variability across scans.

#### 2.4.4. Linear regression based on RISH features (LinearRISH)

This approach, denoted by LinearRISH, first preprocessed the raw diffusion MRI following the HCP pipeline steps using an in-house software. Later, the Gibbs unringing approach (Kellner et al., 2016) was applied to reduce the noise in the Prisma-ST data. Then, the steps described in (Karayumak et al., 2019) were followed: (i) RISH features up to 6th order were estimated separately for diffusion signals at two b-shells with  $b = 1200, 3000 \text{ s/mm}^2$ , respectively, using a Tikhonov regularization based algorithm for estimating SH coefficients with the regularization parameter being 0.001; (ii) The mean templates were created for each RISH feature at  $b = 3000 \text{ s/mm}^2$  using the `antsMultiVariateTemplateConstruction2.sh` command provided by ANTs (Avants et al., 2011); (iii) The scale maps that characterize the corresponding cross-scanner and/or cross-protocol variability were calculated between the mean RISH features from the Prisma-ST datasets and the other three scans. Consequently, the estimated maps were applied to scale the SH coefficients in the subject space. The same steps were also applied to scale the SH coefficients corresponding to  $b = 1200 \text{ s/mm}^2$ ; (iv) The scaled SH coefficients were transformed back to diffusion signals along the provided set of b-vectors for the two b-shells, respectively and the harmonized signal at each shell was estimated in this way. Finally, to preserve the overall integrity of the harmonized signal, the mean harmonized diffusion signal is scaled with the average signal from the other three protocols.

Considering potential inter-subject variability and the limited number of training subjects, the adaptive LinearRISH approach used different subsets of the training dataset to estimate mappings for each test subject. Briefly, from the training dataset, heuristically, the three most similar subjects to each test subject were selected. The similarity criterion was

based on the mean squared error of the RISH features of the Prisma-ST dataset and test subjects in the whole-brain as well as region-wise level. The LinearRISH steps described above were followed using subsets of the training dataset to harmonize each test subject.

#### 2.4.5. Tissue-wise regression forest (TWRF-a and TWRF-b)

The MRAGE images were applied to obtain label maps for WM, GM and CSF regions using FSL FAST (Zhang et al., 2001). For each tissue region, a random forest regression (Breiman, 2001) was trained to predict diffusion signals using different training features. Two sets of training features were extracted for each voxel. The first set included the normalized diffusion signal, the eigenvalues of diffusion tensor, the trace, linearity, planarity and sphericity, which were selected to make the solution comparable to (Alexander et al., 2017), and mean signals from the  $3 \times 3 \times 3$  neighbor voxels, b-values, the b-vectors of both input and the to-be-predicted scans. The second set of features only included the normalized diffusion signals and the mean of the neighborhood. The following parameters were used in the training: the max depth of the decision trees = 10, the number of decision trees (estimators) = 10, criterion was mean squared error, allowed samples per leaf nodes = 1. The *scikit-learn* toolbox (Pedregosa et al., 2011) was used for the training. The prediction results corresponding to the two methods were denoted by TWRF-a (full feature set) and TWRF-b (reduced feature set), respectively. As for training, 6 subjects were used to train the algorithm and the rest subjects were used for validation, implementing the algorithms to the testing datasets.

#### 2.4.6. Single-layer SH network (SH-1Layer)

This approach first interpolated Prisma-ST to match the resolution of the two SA scans. Then, the 4th order SH coefficients of the normalized diffusion signals, i.e.  $s(b, \mathbf{u})/s(0)$  with  $s(b, \mathbf{u})$  being the diffusion signal with b-value = b along the gradient direction  $\mathbf{u}$  and  $s(0)$  being the baseline, were computed separately at the two non-zero b-shells. The harmonization algorithm was based on a network with a single convolutional layer which took the SH volume (both shells concatenated) as input and returns a spherical harmonic volume (again, both shells concatenated) as output. It was trained without validation because the relative lower risk of overfitting for this simple network. The performance of SHNet can be considered as a "baseline" to which other more complex network-based methods should be compared. The network was trained using the Adam optimizer and the mean-squared error (MSE) as the loss function with the learning rate being 0.0001 and the number of epochs being 30.

#### 2.4.7. SH based ResNet networks (SHResNet)

Six methods from two participating teams were developed based on similar ResNet structures (Koppers et al., 2018) using concepts from (He et al., 2016) to learn mappings between SH volumes.

**SHResNet-a:** The method from the first team applied the same preprocessing steps as SHInterp. This algorithm, denoted by SHResNet-a, used the SH coefficient from a  $3 \times 3 \times 3$  voxel neighborhood (input size:  $3 \times 3 \times 3 \times 15$ ) to predict the 15-dimensional SH coefficients in the center voxel using 3D-convolutional networks. The predicted coefficients were converted to diffusion signal using the inverse SH transform. A more detailed information on the structure of this network can be found in Fig. 3 (Tax et al., 2019).

**SHResNet-b:** The methods from the second team includes five variants, i.e. SHResNet-b0 to -b4, that all applied the same preprocessing approaches as SH-1Layer. The first approach, denoted by SHResNet-b0, used the same network structure and training method as SHResNet-a. The difference between SHResNet-a and SHResNet-b0 is related to the differences in preprocessing approaches, i.e. nonlinear registration v.s. interpolation. Methods SHResNet-b1, SHResNet-b2, and SHResNet-b3 included a one-hot orthogonal vector of SLANT labels (Huo et al., 2018) as input to reduce anatomically dependent variability across scanners. In particular, SHResNet-b1 used 133 labels to denote different anatomical

regions. SHResNet-b2 simplified the hierarchy of the labels into 23 unique regions, (i.e. combine cortical lobes, combine cerebral WM, cerebellum, etc, using the hierarchical algorithm in (Asman and Landman, 2014)). SHResNet-b3 further smoothed the labels using a Gaussian filter. Instead of using anatomical labels, Method SHResNet-b4 included a one-hot orthogonal vector of voxel location in x, y, and z coordinates (normalized from  $-1$  to  $1$ ) as input for reducing FOV/location dependent artifact. All the models, SHResNet-b0 to SHResNet-b4, were trained using the same parameters and cost functions as SH-1Layer in Section 2.4.6.

#### 2.4.8. Spherical network (SphericalNet)

This method combines the novel local spherical convolution layer (LSC) (Koppers and Merhof, 2018) with the SHResNet (Koppers et al., 2018). The LSC layers are utilized as pre- and postprocessing layer of the network. First, it separates the two input-shells into ten linear combinations, based on the neighboring q-space signals, while the last layer combines the output of the ResNet to predict the two shells of the three to-be-predicted scans. Further, each residual block was extended to be applicable to multi-shell data.

#### 2.4.9. Tiny-patch network (Tinypatch)

In pre-processing, this method, named Tinypatch, applied FSL FAST to obtain label maps for white-matter, gray-matter, and CSF regions using the MPAGE images. The diffusion data were projected to an 8th order Spherical Harmonics (SH) basis, using an L2-minimal projection. These projections were concatenated voxelwise with the tissue labels and then fed into a patch-wise feed-forward fully connected neural network, which output estimates of a corresponding patch from the to-be-predicted site.

Patches were constructed using a center voxel and the 6 immediately adjacent voxels. The input was then vectorized, and then fed into the fully connected network. The neural network consisted of 2 encoder layers, one center layer, and 2 decoder layers, with (128,64), (32), and (64,128) hidden units respectively, using tanh non-linear activations at each hidden unit. The encoder and center layers were pre-trained using an auto-encoder task, attempting to reconstruct the base site using a temporary set of decoder layers. The full network was then trained using corresponding patches between scans; for the higher resolution scans (Task 2a and Task 2b) the spatial offsets were concatenated to the activations at the center layer. Loss was computed by projecting SH estimates back to the subject specific b-vectors. This loss was then propagated back to the rest of the network.

A separate network was trained for each task site (Task 1, Task 2a, and Task 2b). Each network was trained to convergence on 9 of the training subjects, leaving one subject for validation. The best performing weight set on the validation subject was then run on the test dataset, and the outputs submitted for evaluation.

#### 2.4.10. Deeper image quality transfer (DIQT)

Raw data were pre-processed to extract signal features by spherical harmonics (SH) deconvolution (Tournier et al., 2004). The pre-processing was performed using an home-made python script, also including some functions from DIPY (Garyfallidis et al., 2014). For each of the 2 diffusion gradient shells, the 15 coefficients of the 4th order SH deconvolution were estimated from the normalized raw signal at different diffusion gradients directions, using real and symmetric SH basis. These 30 SH coefficients (15 coefficients of the 4th order SH deconvolution per each of the 2 shells) computed from the available subjects in the training set, were then used as input to a CNN, to predict the corresponding 30 SH coefficients for the unseen subjects in the test set. From the predicted 30 SH coefficients, the normalized raw signal for the subjects in the test set was computed by evaluating the SH in the provided set of directions.

The DIQT approach used supervised learning on 3D input-target patches that were extracted from the Prisma-ST scans. The center voxel of each input patch was contained within the brain mask, and each patch

was of physical size  $11 \times 2.4^3 \text{ mm}^3$ . Before extracting the patches, the mean and standard deviation were computed for each input channel, i.e. SH coefficient across training subjects. Then the training (and validation) data was normalized to have zero mean and the standard deviation one. The set of input patches were then used to predict image patches in the target space where the size of the target patch was chosen so that, when considered in the same physical space, the target patches were contained within the input patches.

The performance validation of this algorithm was initially explored on two fixed subjects. Then cross-validation was utilized, where the hold-outs were different to the data exploration subjects, to estimate the generalized error-rate. 10% of the training data was allocated as a pre-validation set, for early stopping of model training. The final validation/submission was performed by multiple reconstruction of a subject's brain. For a reconstruction, a subject's input brain was divided into non-overlapping input patches, where the patch predictions were combined to create a whole brain prediction. This procedure was performed 64 times and the reconstructions averaged. Variants of the DIQT Network (Blumberg et al., 2018; Blumberg et al., 2019) were used for the three tasks, where the shuffle in (Blumberg et al., 2018) was removed in Task1. In Tasks 2 and 3, the shuffle in was replaced with a transposed convolutional layer. The model hyperparameters of these networks, including the number of reversible network layers per block (a proxy for the network's depth), size of input/target patch, learning rate and decay, sampling scheme for training data, were explored in (Blumberg et al., 2019). The code for the extension of this challenge and (Blumberg et al., 2019) is available at <https://github.com/sbb-gh/>.

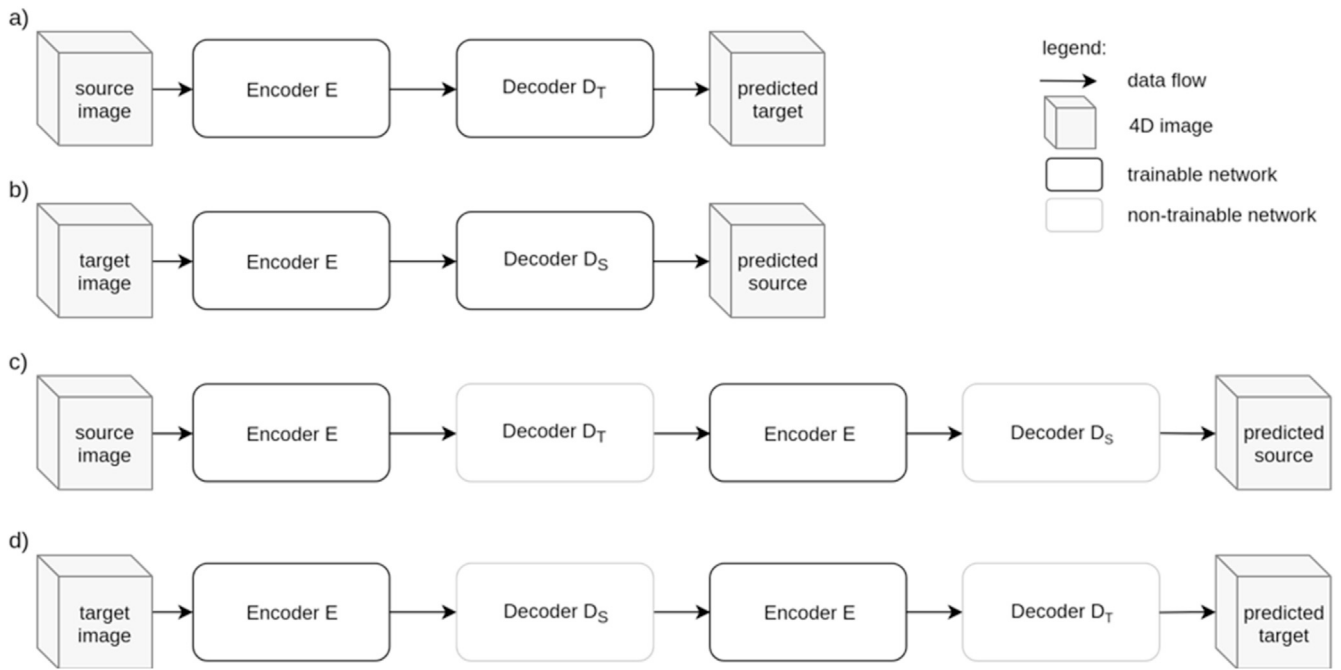
#### 2.4.11. Spherical harmonics and a Radial Decomposition (SHARD) based ResNet (CMResNet-a and CMResNet-b)

As preprocessing, this approach applied the Gibbs unringing toolbox (Kellner et al., 2016) to suppress noise and the ANTs N4 tool (Tustison et al., 2010) to correct for intensity inhomogeneity. Then, the Connectom scans were diffeomorphically registered the Prisma data using multi-modal registration as implemented in MRtrix3 (*mrregister*) (Raffelt et al., 2011), with sum-of-squared differences cost function between FA, MD and histogram-matched b0.

Next, this method used the Spherical Harmonics and a Radial Decomposition (SHARD) proposed in (Christiaens et al., 2018) as rotation-invariant representation of multi-shell dMRI data. SHARD learns the optimal low-rank representation of a given dataset (single subject or group), using a singular value decomposition across shells and spherical harmonic bands. In this work, a groupwise basis was learned for each protocol (*Prisma-ST* and *Connectom-ST*) from the 6 training subjects (subjects A-F). The data of all (training, validation, and test) subjects is then projected onto the basis, thus representing each scan by a 4-D image consisting of 57 vol of SHARD coefficients in the basis of its protocol. The SHARD images were estimated in the original space of Connectom scans then warped to the space of Prisma data.

Two ResNet networks, denoted by CMResNet-a and CMResNet-b, were used to learn mappings between the SHARD images, where each included an encoder and two decoders. The encoder transforms the source SHARD image to a common embedding. The two decoder networks  $D_S$  and  $D_T$  transform from embedding to source and target domain, respectively. The networks consist of multiple depth-wise separable convolution layers, residual blocks, and a Squeeze-and-Excitation connection (Hu et al., 2017) that uses channel-wise average pooling inside the brain mask as features for adaptive channel scaling.

The structure of the two networks, CMResNet-a and CMResNet-b are shown in Fig. 1 to Fig. 3. The networks use a multiply connection between input and encoder output and expand the  $N = 57$  SHARD volume channels to  $C = 72$  channels. The networks were trained using a mean squared loss on the SHARD coefficients with either source or target image as input. The transformed image from the single pass through encoder and decoder was evaluated against the same subject scanned with the other protocol (see Fig. 1a) and b)). Additionally, using two passes



**Fig. 1.** The network uses the encoder E to transform the source and/or target image to a common embedding that the decoders  $D_T$  and  $D_S$  transform to the target and source domain, respectively. The weights are shared across all networks with identical names. The rows show the methods for training the encoder and decoder networks of the networks to predict the target SHARD image given the source SHARD image a), the inverse mapping b), and the full cycle from source to predicted source c) and from target to predicted target d).

through the network, the input image was compared to the prediction given itself as input as shown in Fig. 1c) and d). In the first two training methods the loss was evaluated between domains, in the latter, it was evaluated using a single image from either domain. Chaining E,  $D_S$ , E and  $D_T$  allows training the network using a cycle-consistency loss (Zhou et al., 2016; Liu et al., 2017) of source images without the need for well aligned matching target data. To avoid learning an identity mapping in this setting, either  $D_T$  or  $D_S$  are held constant during cycle-consistency training steps and the network is trained simultaneously with paired and unpaired loss.

### 3. Results

#### 3.1. Task 1: matched resolution cross-scanner harmonization

##### 3.1.1. Global evaluation

Fig. 4(a) and (b) illustrate the average APE of the FA and MD measures over all voxels within the evaluation masks for 19 algorithms that completed Task 1. These DTI features were computed using only diffusion signals at the b-shell with  $b = 1200 \text{ s/mm}^2$ , since DTI only captures the second order cumulant, and higher-order cumulants become more prominent at higher b-values. Similarly, Fig. 5(a) to 5(d) show the APE of the zeroth and second order RISH features, R0 and R2, for diffusion signals at the two b-shells, respectively. Fig. 6(a) and (b) illustrate the APE of MK and RTOP computed using multi-shell diffusion signals. Algorithms of the same type, i.e. interpolation-, regression- or CNN-based, are coded by the same color.

The bar plots in Figs. 4 to 6 show the 10th to 90th percentile of the distribution of APE values over all voxels. The blue circle illustrates the median value of APE distributions. The 90th percentile bars are significantly further away from the median values than the 10th percentile bars, indicating that the underlying distributions have long tails which may be driven by outliers. Thus, the mean values, indicated by the red squares, were computed using all the measures lower than the 90th percentile to reduce the effect of outliers, i.e. it is the truncated mean.

The leftmost barplot in Figs. 4 to 6, denoted by “Reference”, shows

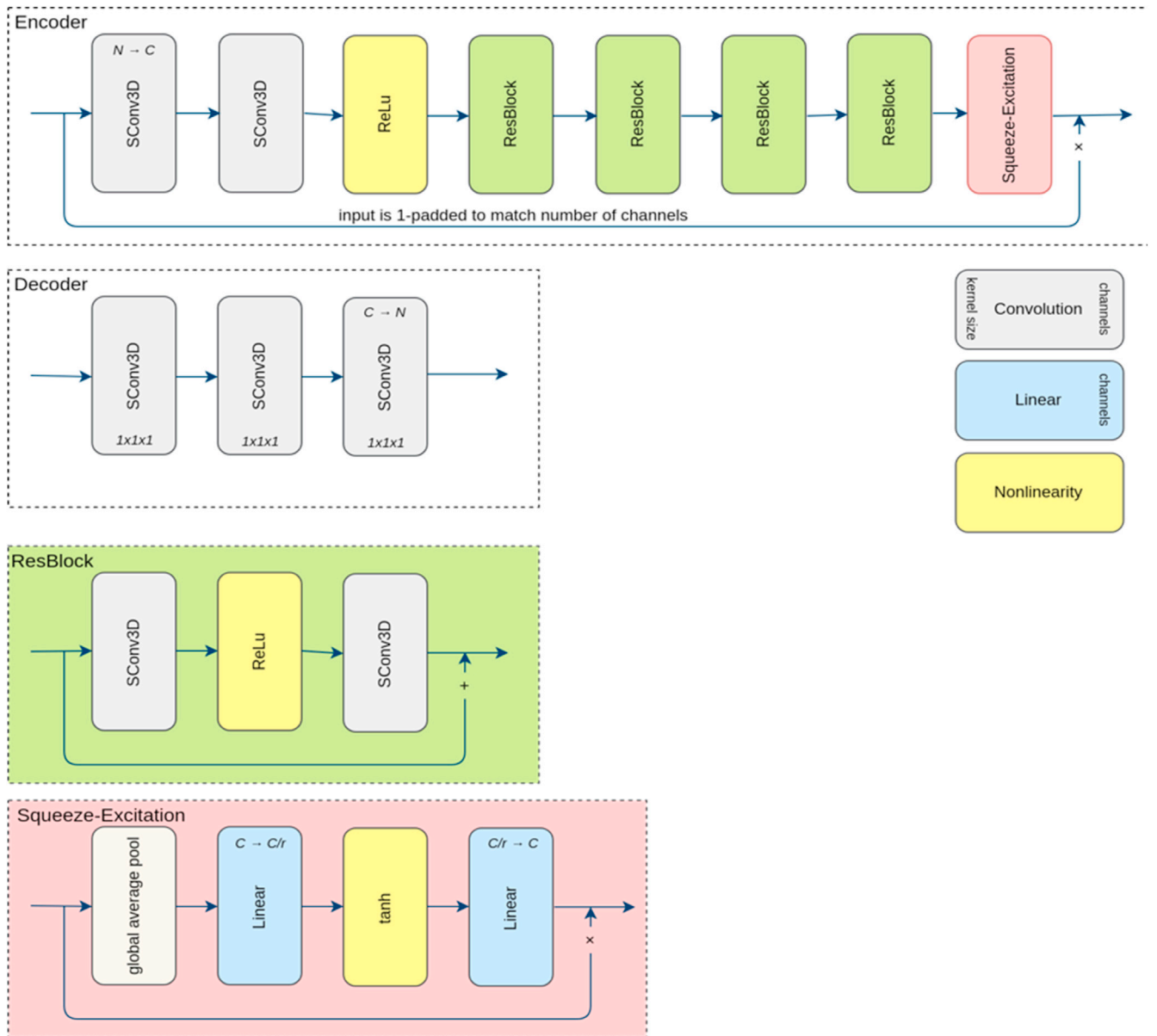
the APE values between imaging measures from the Prisma-ST and Connectom-ST datasets. Since the two datasets have similar scanning parameters and matched spatial resolution, the APE values were computed by voxel-wise differences between the imaging measures. The reference distribution shows the true cross-scanner variability if no harmonization algorithm is applied. The results of a good harmonization algorithm are expected to have lower mean and median APE values than the Reference, which are indicated by the two dashed lines.

The 2nd to 5th barplots in Figs. 4 to 6 show the results of interpolation-based algorithms, i.e. SHInterp to SHOREInterp. It is interesting to point out that all techniques outperformed the reference though the cross-scanner and cross-protocol variability is not considered in the underlying algorithms. In particular, SHInterp and SHOREInterp have similar mean and median values in APE of FA, MD MK and RTOP as the reference. The corresponding APE values of RISH features are slightly lower than the reference, which is possibly due to the denoising effect of the algorithms. DIAMOND-a and DIAMOND-b have relative lower APE values for MD, RISH features, MK and RTOP than the reference.

The 6th to 8th barplots in Figs. 4 to 6 show the results of regression-based algorithms, i.e. LinearRISH to TWRF-b. The LinearRISH approach has the lowest values in APE of FA, MD, RISH features and MK. But it has a similar APE of RTOP as the reference. On the other hand, TWRF-a and TWRF-b have relative higher APE values than the reference. TWRF-a has worse performance than TWRF-b which may be caused by the redundant features, such as the b-values and the b-vectors.

The last 12 barplots in Figs. 4 to 6, from SH-1Layer to CMResNet-b, illustrate the results based on convolutional neural networks. The results of SH-1Layer serve as a baseline for other deep-network-based results. It can be seen that several deep-network-based algorithms, such as SHResNet-b0 to SHResNet-b3, have much higher prediction error than SH-1Layer, though the underlying network structures are more complex. Other methods, including SHResNet-a, Tinypatch, DIQT, CMResNet-a and CMResNet-b, all have similar performances as SH-1Layer.

A summary of the mean APE values of the 19 algorithms is provided in Table 2. It can be seen that SHInterp has similar or slightly better performance than the reference. The two DIAMOND model-based



**Fig. 2.** The encoder and decoder networks architecture of CMResNet-a and their building blocks. SConv3D denotes a depth-wise separable convolution consisting of a convolution (with kernel size  $3 \times 3 \times 3$ , unless otherwise stated, stride = 1, zero padded), grouped by input channels, followed by a  $1 \times 1 \times 1$  convolution (stride = 1, no padding) mapping to the number of output channels. The number of channels is omitted if it is unchanged. The Squeeze-and-Excitation reduction factor ( $r$ ) is 4.

algorithms have very similar performances, indicating the different choices of DIAMOND models have only little effect on the results. SHOREInterp has similar performances as SHInterp and Reference. TWRF-b is slightly better than TWRF-a, but both actually increased variability compared to the reference. SHResNet-a is the best among the six SHResNet-based algorithms. DIQT is in general the best CNN based approach, but slightly worse than SH-1Layer or CMResNet-b in some measures. Overall, LinearRISH has the best performance in most evaluation metrics but the RTOP measure, in which LinearRISH is the 2nd highest and is worse than the reference and SHInterp.

### 3.1.2. Regional evaluation

Fig. 7 illustrates the spatial distribution of the mean APE over the ROIs from the Kesikan-Killiany atlas from FreeSurfer (Fischl, 2012) in sagittal, axial and coronal views of a brain for 8 representative algorithms. The value in each ROI is the average APE over all measures and all the underlying voxels, with the size of ROIs not considered in the analysis. In particular, Fig. 7(a) to 7(h) show the APE corresponding to Reference,

DIAMOND-a, LinearRISH, SH-1Layer, SphericalNet, SHResNet-b0, DIQT and CMResNet-b. Fig. 7(a) shows that the frontal and temporal lobes have the highest errors, which may be related to residual, uncorrected distortions and motion-induced effects in these regions. The errors are significantly reduced by several methods in Fig. 7(b) to 7(f), though the frontal and temporal lobes still have relatively higher errors than other regions. It can be seen from Fig. 7(c) that LinearRISH has relatively lower errors than other methods in almost all regions.

Table 3 summarizes the mean APE in the gray matter and white matter regions of all 19 algorithms. It also provides the two ROIs with the highest or lowest APE for each method. For all methods, the APE values in gray matter are all higher than the result of white matter. Moreover, several ROIs, such as the entorhinal cortex, the superiorparietal cortex, and the rostralmiddlefrontal cortex, consistently have the highest errors for most algorithms. On the other hand, white matter regions in the paracentral, the banks of the superior temporal sulcus (bankssts) and the insula have the lowest errors.



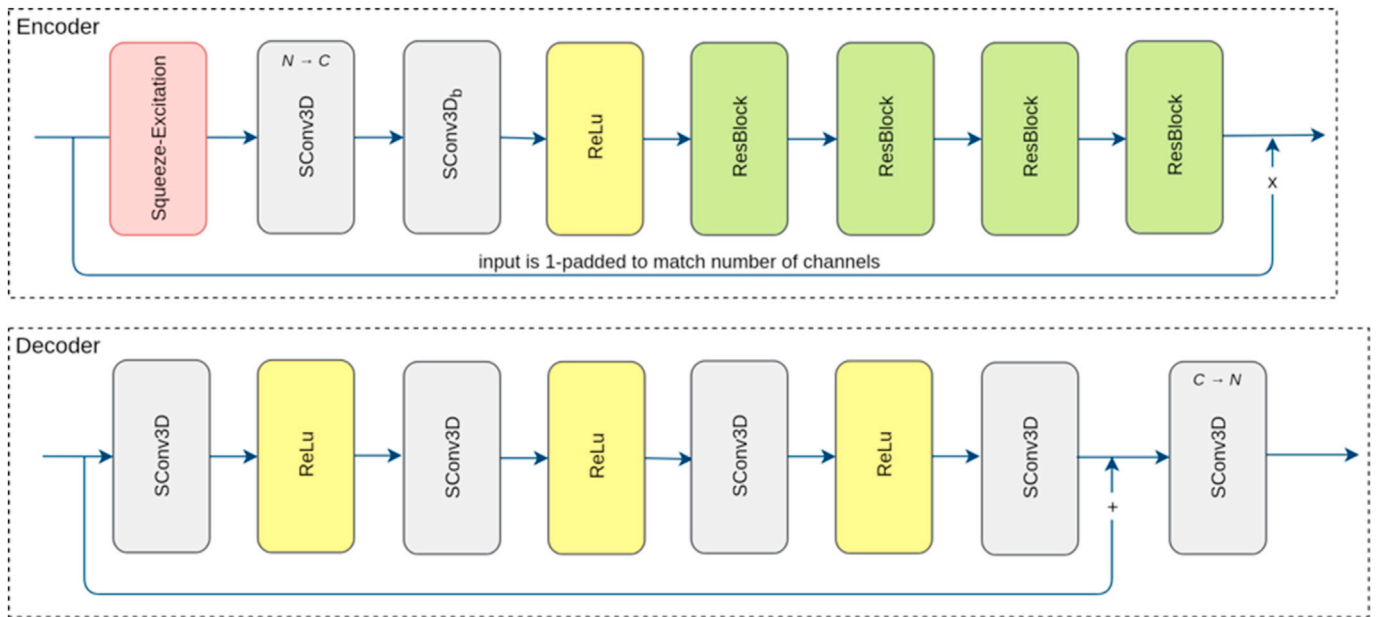


Fig. 3. Encoder and decoder network architecture of the CMResNet-a. See Fig. 2 for the layer legend.

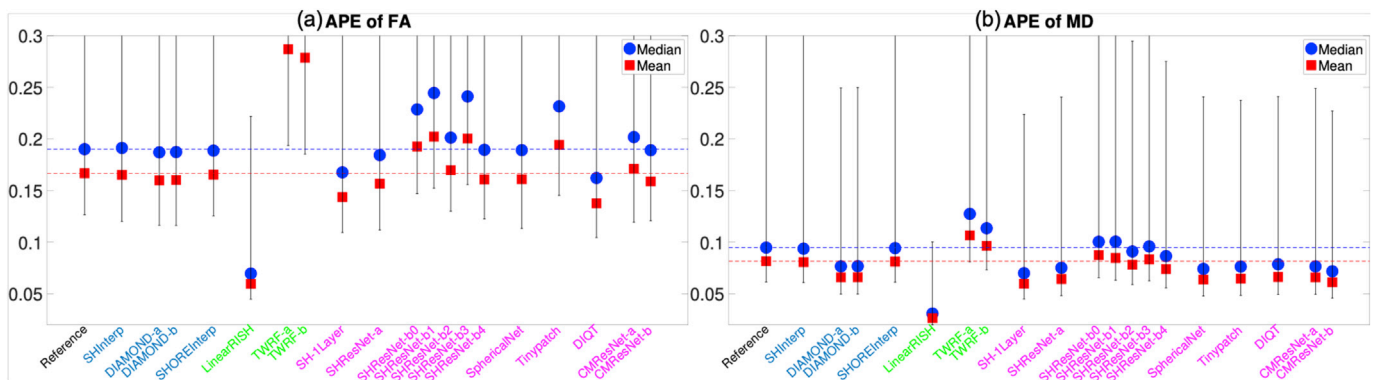


Fig. 4. An illustration of the APE of FA and MD measures computed using diffusion signals at the b-shell with  $b = 1200 \text{ s/mm}^2$  for algorithms evaluated for Task 1. The reference (dashed lines) illustrates the difference between Prisma-ST and Connectom-ST scans with no harmonization. The y-axis does not include the origin in order to emphasize the differences between methods.

### 3.2. Task 2: spatial- and angular-resolution enhancement

#### 3.2.1. Global evaluation

The barplots in Figs. 8 to 10 illustrate the APE of 8 measures for 19 algorithms evaluated for Task 2a. Overall, the average APE is higher than the corresponding results in Task 1 shown in Figs. 4 to 6. The reference value of APE is not available in this task since the Prisma-ST and Prisma-SA scans have different spatial resolutions. But the SHInterp approach can be considered as the “baseline” since it provides harmonization errors as assessed in Task 1 that are similar to the intrinsic between-scanner differences. A summary of the mean APE values for 8 evaluation measures is shown in Table 4.

As in Task 1, the DIAMOND and SHORE-based approaches have better performances than the SHInterp method, indicating that DIAMOND and SHORE are more agnostic to scanner or protocol dependent noise/errors in the measurements. The LinearRISH approach still has the best performances in predicting the FA, MD, RISH and MD measures, though it has higher error in RTOP than SHInterp.

For CNN-based approaches, the simple SH-1Layer approach has better performances than SHResNetb1 to SHResNet-b4 and DIQT, implying that these complex networks need to be further trained to improve the performances. Nonetheless, other deep-networks based methods, such as

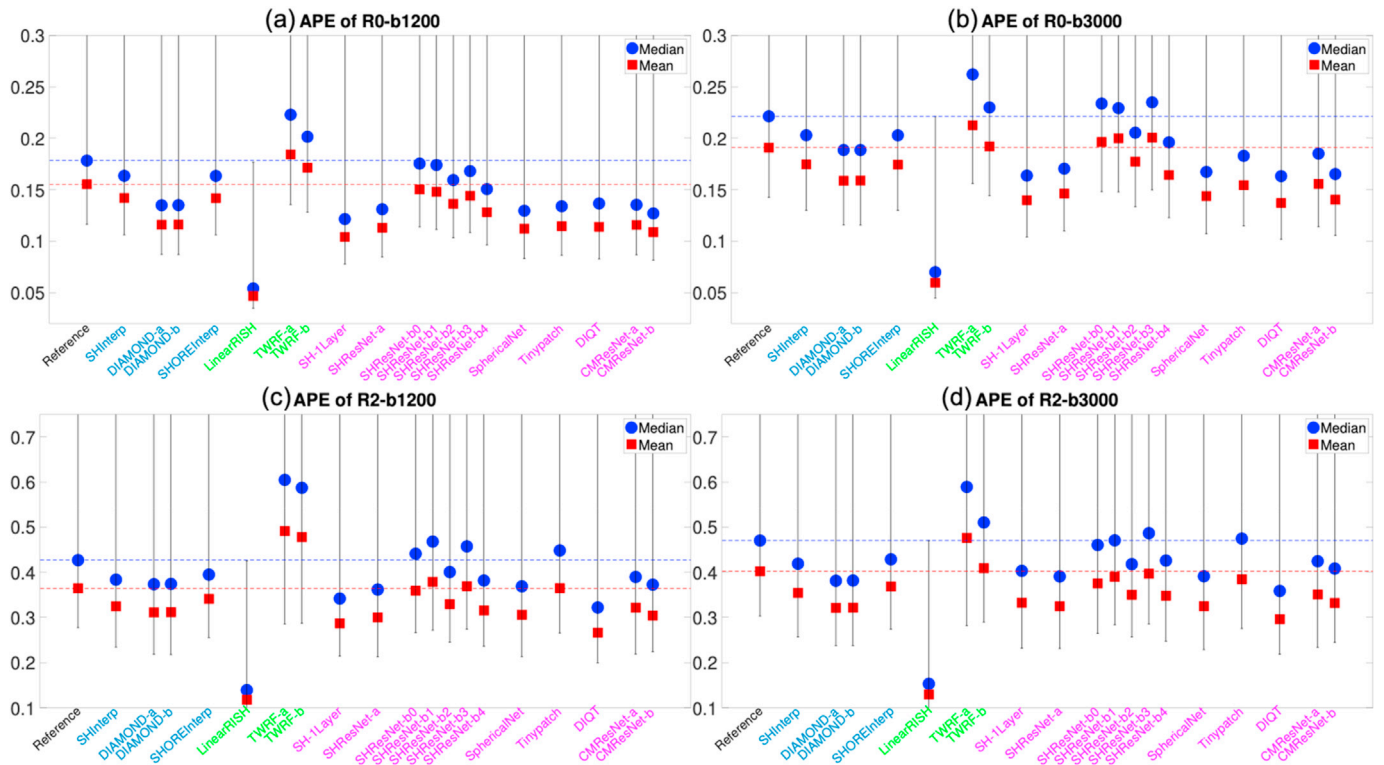
SHResNet-b0 and CMResNet-b, have better performances than SH-1Layer.

The global evaluation results for Task 2b are provided in Figs. S1–S3 and Table S1 in the Supplementary Materials. Overall, similar results as in Task 2a were found for this task. The LinearRISH approach still has the best performances in most metrics but the RTOP measure. The DIAMOND-based methods and the SH-1Layer approach have similar performances. SHResNet-b0 and DIQT are two of the best CNN-based algorithms.

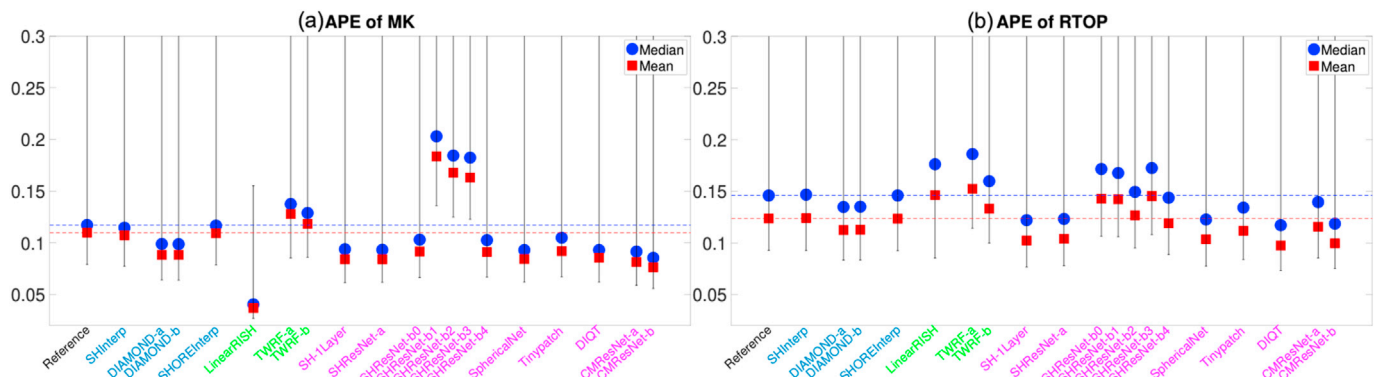
#### 3.2.2. Regional evaluation

The spatial distribution of APE values has similar patterns as in Task 1. In particular, the frontal and the temporal lobes have relatively higher errors than other regions which may be caused by image distortions or head motions. A detailed illustration of the spatial patterns for the algorithms evaluated for Task 2a and Task 2b is provided in Figs. S4 and S5 in the Supplementary Materials. Table 5 shows the mean APE values in GM and WM and the two ROIs with the highest or lowest APE values for Task 2a. The corresponding results for Task 2b are shown in Table S2 in the Supplementary Materials.

The estimation errors have similar spatial patterns as in Task 1. In particular, the errors in gray matter are significantly higher than the



**Fig. 5.** An illustration of the APE of the zeroth and second order RISH features, R0 and R2, of diffusion signals at two b-shells for the algorithms evaluated for Task 1. The reference (dashed lines) illustrates the difference between Prisma-ST and Connectom-ST scans with no harmonization.



**Fig. 6.** An illustration of the APE of MK and RTOP computed using multi-shell diffusion signals for algorithms evaluated for Task 1. The reference (dashed lines) illustrates the difference between Prisma-ST and Connectom-ST scans with no harmonization.

errors in white matter. The cortical regions in the superiorparietal, the entorhinal, the temporalpole and the frontalpole have the highest APE values. On the other hand, white matter in paracental areas consistently exhibits the lowest APE for most algorithms.

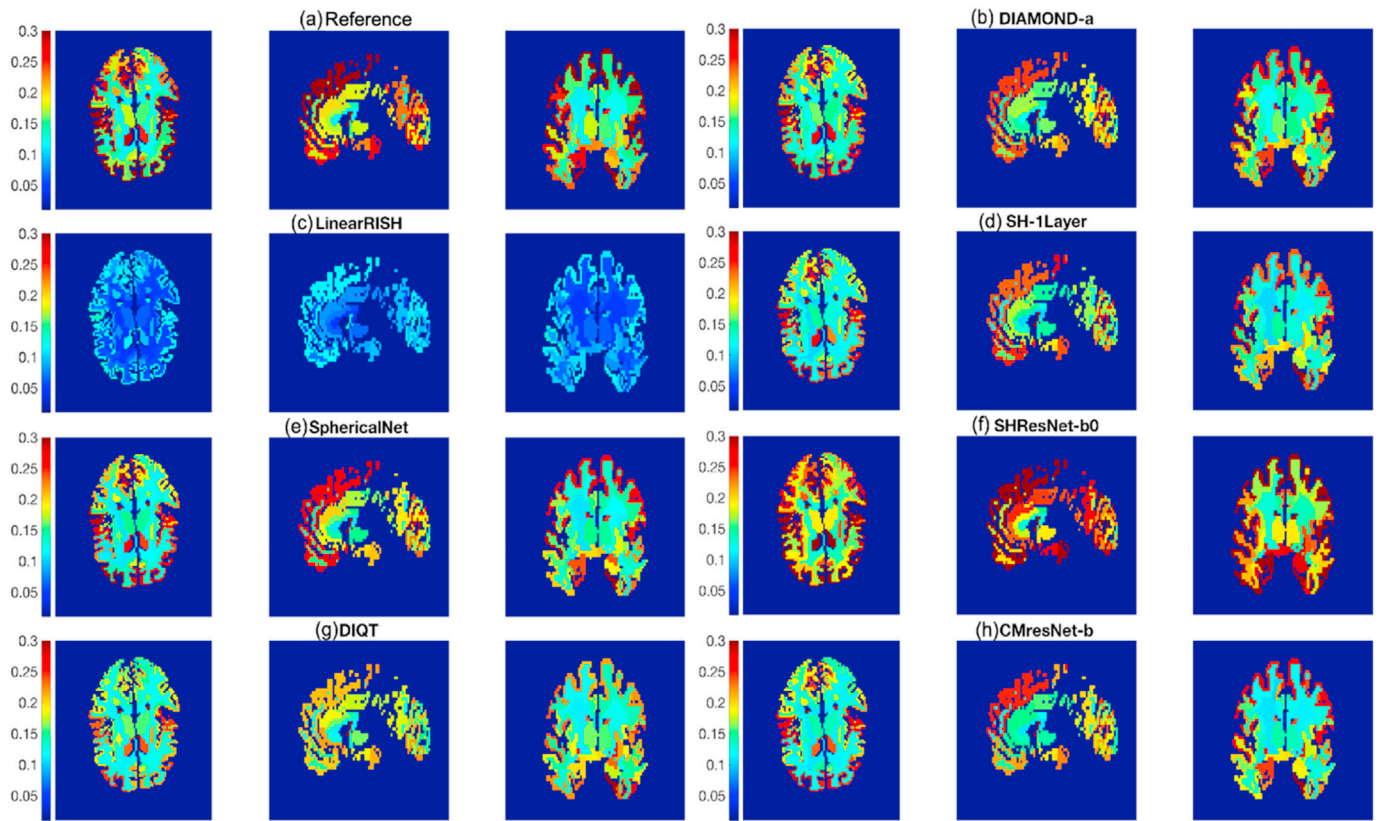
#### 4. Discussion

In this work, we have evaluated the performance of 19 algorithms that have participated in an open competition on harmonization of multi-shell dMRI scans from different scanners and protocols. Next, we discuss several factors that may limit the results of this study and possible solutions to further improve the performance in future work.

##### 4.1. Sample size of training data

The performance of harmonization algorithms may depend on the

number of subjects used in training (Karayumak et al., 2019). Ideally, the number of subjects should be large enough so that the training dataset contains sufficient information to learn the variability across datasets acquired using different scanners and/or protocols. If the harmonization algorithm is based on deep networks, the number of training examples should be large enough to ensure the convergence of the training algorithms and to avoid overfitting. But, in practice, travelling subjects are expensive and logistically difficult, making them infeasible for multi-center studies. Thus, many studies only use one subject, or a phantom, to test the stability of imaging measures (Teipel et al., 2011). The benchmark database evaluated in this study is one of the largest open datasets with scans from travelling subjects using different scanners and protocols. However, the analysis results show that the performance of several algorithms with complex network structure is worse than a simple one-layer network, i.e. Sh-1Layer, indicating the training data is still insufficient for complex algorithms. Thus, more effective network



**Fig. 7.** Illustration of the average regional APE values of all dMRI measures in different brain regions defined by the Desikan-Killiany atlas from FreeSurfer for several representative algorithms evaluated for Task 1 and the reference.

structures and training algorithms need to be explored in future multi-site harmonization studies.

#### 4.2. Inter-subject variability and adaptive prediction

In the evaluation results, the LinearRISH method has demonstrated better performance than other algorithms in most of the evaluation metrics, followed by the DIAMOND and SHORE-based approaches. Among other reasons, three possible factors that contribute to its superior performances are as follows: First, the simple linear-regression approach is able to avoid overfitting and provide robust data predictions. The voxel-wise linear models used in LinearRISH are able to reliably reduce space dependent variability in the datasets. Second, this approach adaptively selected different training subjects to predict each testing dataset based on the similarities in the corresponding Prisma-ST scans. Though it is expected that there exist subject-independent factors that are related to the cross-scanner and cross-protocol variability in the dataset, there may also exist other factors, such as head sizes and shapes, that lead to subject-specific variability across scans. The adaptive prediction approach used by LinearRISH was shown to be an effective solution to reduce subject-dependent variability in learning mappings between dMRI data from different scanners and protocols. However, the potential of adaptive prediction in clinical applications with pathological data need to be explored in future work. Third, LinearRISH is the only method that uses 6th order SH representations, while other methods use 4th order SH representation. The higher order SH representations may provide more detailed information on the data that is helpful to improve the performance of harmonization algorithms.

#### 4.3. On evaluation metrics

This study used several metrics, including DTI, DKI, MAP-MRI and

RISH features, to evaluate the performance of harmonization algorithms on a global and regional level. Note that the FA and MD measures were computed only using dMRI signals at the lower b-shell at  $b = 1200 \text{ s/mm}^2$ , since at higher b-values higher-order cumulants become more prominent. Moreover, the RISH features were computed separately for the two b-shells at  $b = 1200, 3000 \text{ s/mm}^2$ , respectively. For most algorithms, the prediction error of RISH features is higher for higher b-values. Moreover, the MK measure from DKI and the RTOP measure from MAP-MRI were computed by integrating dMRI signals at both b-shells. The two measures reflect different aspects of the signals. In particular, the MK measure was computed based on the cumulant expansion, which do not provide an accurate prediction of diffusion signals to high b-values. Moreover, the maximum b-value at  $b = 3000 \text{ s/mm}^2$  might be beyond the radius of convergence for DKI, jeopardizing accuracy but potentially improving precision (Chuhutin et al., 2017). On the other hand, the RTOP measure was computed based on an analytical expression of the ensemble average propagator estimated using diffusion signals. Though other approaches were proposed to compute the RTOP measure, these methods all rely on continuous extrapolation of diffusion signals. Thus, the RTOP measure is sensitive to the decay of diffusion signals and the estimation methods. Although at least two b-shells are needed to estimate the MK and RTOP for characterizing non-exponential signal decays, more b-values are expected to further improve the consistency of estimation results. In the evaluation results, the LinearRISH approach achieved the best performance in almost all metrics except RTOP which may be caused by errors in relative signal decays along the radial direction of b-values of the predicted signals. Caution should be placed in the results from the RISH (R0 and R2) metrics since these are based on the same underlying model as LinearRISH. However, the evaluation with different metrics should reduce potential bias in metrics whose underlying model is similar to the proposed algorithms. These diversified metrics could provide a guideline on the selection of harmonization algorithms corresponding to different metrics.

**Table 3**

Summary of the average percentage APE values in GM, WM and two ROIs from the Desikan-Killiany atlas from FreeSurfer with the highest and lowest APE for Task 1.

	APE (%)		ROIs with the highest APE	ROIs with the lowest APE
	GM	WM		
SHInterp	27.5	17.1	ctx-rh-rostralmiddlefrontal, ctx-rh-superiortemporal	wm-lh-bankssts, wm-rh-insula
DIAMOND-a	23.2	15.1	ctx-lh-entorhinal, ctx-rh-superiorparietal	wm-rh-paracentral, wm-lh-paracentral
DIAMOND-b	23.2	15.0	ctx-lh-entorhinal, ctx-rh-superiorparietal	wm-lh-paracentral, wm-rh-paracentral
SHOREInterp	27.9	17.1	ctx-rh-superiortemporal, ctx-rh-cuneus	wm-lh-bankssts, wm-lh-insula
LinearRISH	10.7	6.7	ctx-rh-superiortemporal, ctx-lh-parsorbitalis	wm-lh-paracentral, wm-lh-bankssts
TWRF-a	26.0	25.5	ctx-lh-entorhinal, wm-rh-parsorbitalis	ctx-lh-isthmuscingulate, ctx-lh-bankssts
TWRF-b	26.2	22.1	ctx-lh-entorhinal, ctx-lh-superiorfrontal	wm-lh-inferiorparietal, wm-lh-lateraloccipital
SH-1Layer	22.0	13.6	ctx-lh-entorhinal, ctx-rh-parsorbitalis	wm-lh-bankssts, wm-rh-paracentral
SHResNet-a	22.8	14.3	ctx-lh-entorhinal, ctx-rh-rostralmiddlefrontal	wm-lh-bankssts, wm-lh-paracentral
SHResNet-b0	30.8	18.6	ctx-rh-parsorbitalis, ctx-lh-parsorbitalis	wm-rh-paracentral, wm-lh-insula
SHResNet-b1	32.6	20.9	ctx-rh-inferiorparietal, ctx-rh-cuneus	wm-lh-bankssts, wm-rh-lateralorbitofrontal
SHResNet-b2	29.2	18.3	ctx-lh-entorhinal, ctx-rh-parsorbitalis	wm-lh-bankssts, wm-lh-insula
SHResNet-b3	31.0	20.4	ctx-lh-entorhinal, ctx-lh-parahippocampal	wm-rh-insula, wm-rh-lateralorbitofrontal
SHResNet-b4	24.0	16.1	ctx-lh-entorhinal, ctx-rh-pericalcarine	wm-lh-paracentral, wm-rh-paracentral
SphericalNet	23.0	14.1	ctx-lh-entorhinal, ctx-lh-parsorbitalis	wm-lh-paracentral, wm-lh-bankssts
TinyPatcher	22.8	15.6	ctx-lh-entorhinal, ctx-lh-postcentral	wm-lh-paracentral, wm-rh-lateralorbitofrontal
DIQT	20.9	13.7	ctx-lh-entorhinal, ctx-lh-parahippocampal	wm-lh-bankssts, wm-rh-precuneus
CMResNet-a	23.0	15.4	ctx-lh-entorhinal, ctx-lh-parsorbitalis	wm-rh-paracentral, wm-lh-paracentral
CMResNet-b	21	14	ctx-lh-entorhinal, ctx-rh-rostralmiddlefrontal	wm-lh-paracentral, wm-rh-paracentral

The ROI based analysis, e.g. Fig. 7, shows that the GM has much higher APE values than WM. A possible reason is that GM has more significant heterogeneity due to different layers and cortical folds, which make it more difficult to harmonize the dMRI data. The findings in the regional analysis suggest that caution should be taken when interpreting results on these high-APE regions. Future work is needed to develop more

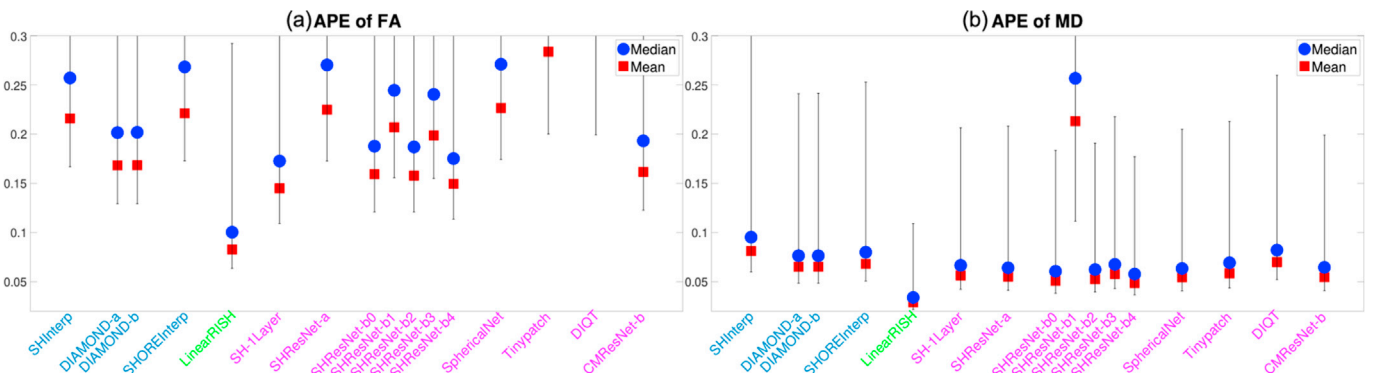
reliable algorithms to harmonize these ROIs. Finally, we mention that a limitation of the analysis is that only the APE is used to compare different dMRI measures. The APE metric may not reflect the impact of the differences in dMRI data on statistical analysis in clinical studies. Further evaluation metrics will be needed in future studies to highlight the differences between dMRI data for clinical studies from different angles.

#### 4.4. On signal representations and models

The harmonization algorithms used four types of approaches to represent or model diffusion signals including spherical harmonics (SH), Spherical Harmonics and a Radial Decomposition (SHARD), the DIAMOND model and the SHORE model. In particular, SH was the most commonly used approach that provides a complete basis for the representation of single-shell diffusion signals. Typically, fourth- or sixth-order SH representations were used to remove high-frequency noise. Moreover, the SH coefficients were usually separately estimated for different b-shells without modeling the radial signal decay. On the other hand, the novel SHARD representation and the representations based on the DIAMOND and SHORE models were computed using multi-shell diffusion signals. These approaches use either parsimonious/band-limited representations or biophysical models to remove noise in the signals. In the interpolation-based approaches, SH, DIAMOND and SHORE were all used to predict dMRI signals without removing cross-scanner and cross-protocol variability. It was interesting that all these approaches, especially the DIAMOND based methods, provided lower prediction error than the reference, indicating that these biophysical and statistical hybrid models accurately characterize the diffusion properties of the underlying tissues and reduce the variability across scanners, particularly variability caused by measurement noise.

## 5. Conclusion

In this study, we have presented a comprehensive evaluation of the performance of multi-shell harmonization algorithms for reducing cross-scanner and/or cross-protocol variability in dMRI data using a benchmark database. The results showed that some algorithms were able to significantly reduce the variability to a similar level as the scan-rescan variability between different scans of the same subjects using the same scanner and protocol. Moreover, the results from different types of algorithms indicate that inter-subject variability and measurement noise are important factors causing variability among dMRI scans. The performance of a large set of algorithms evaluated in this study could provide useful information to guide the selection of harmonization algorithms in future multi-site studies with travelling subjects. The benchmark database used in this study can be obtained on the following webpage: <https://www.cardiff.ac.uk/cardiff-university-brain-research-imaging-centre/research/projects/cross-scanner-and-cross-protocol-diffusion-MRI-data-harmonisation>. Evaluation pipelines and the results are available at: <https://projects>



**Fig. 8.** Global APE values for the FA and MD measures computed using diffusion signals at the b-shell with  $b = 1200 \text{ s/mm}^2$  for algorithms evaluated for Task 2a.



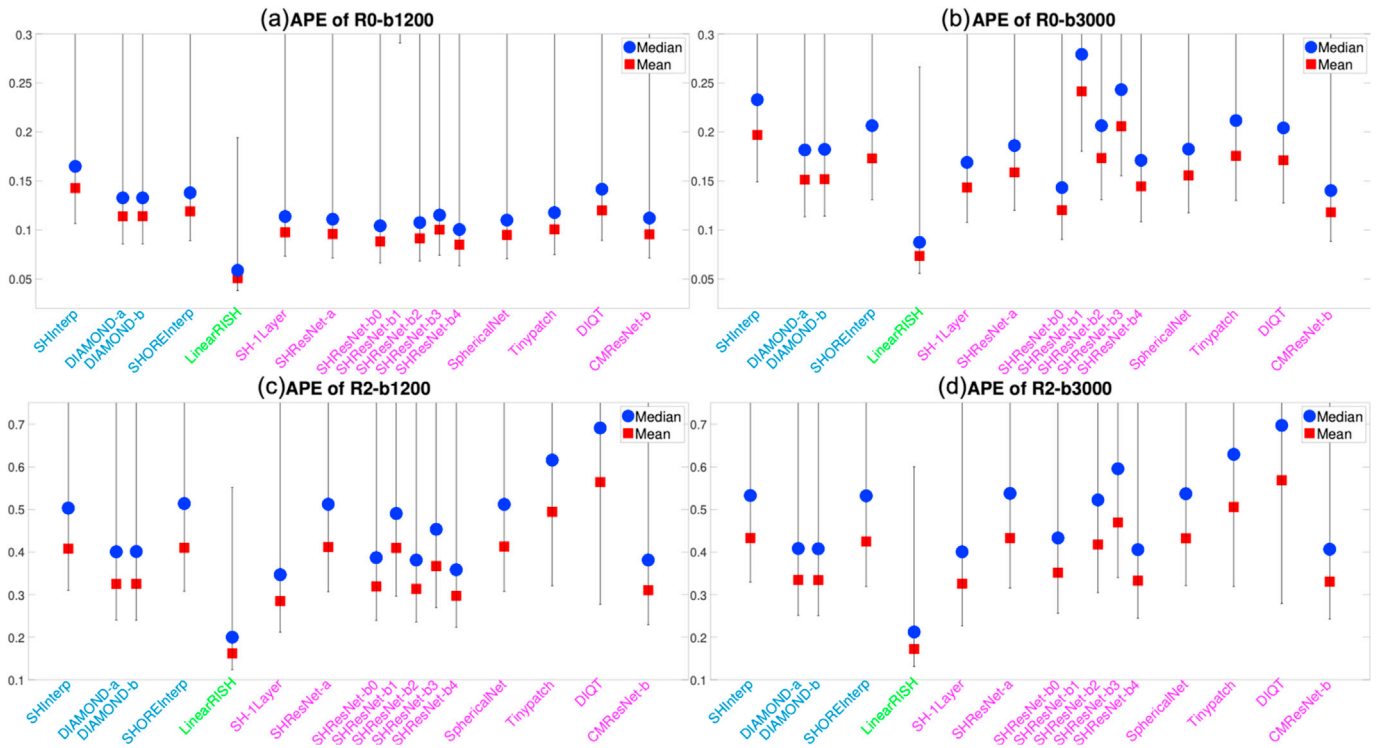


Fig. 9. Global APE values for the zeroth and second order RISH features, R0 and R2, of diffusion signals at two b-shells for the algorithms evaluated for Task 2a.

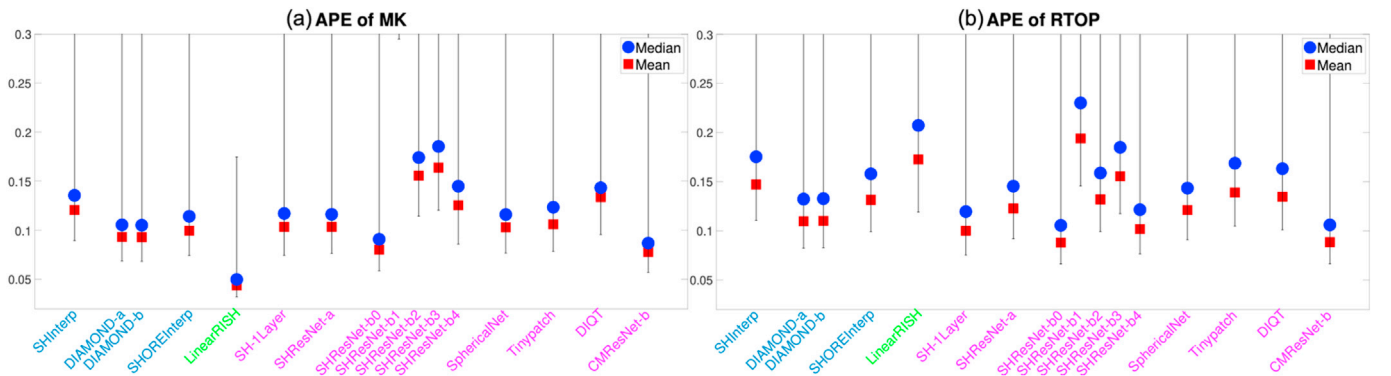


Fig. 10. Global APE values for the MK and RTOP measures computed using multi-shell diffusion signals for algorithms evaluated for Task 2a.

Table 4

Summary of the percentage APE values, i.e. APE<sub>x100%</sub>, of reported measures in global evaluation for Task 2a.

	APE (%): Mean (Standard deviation)							
	FA	MD	R0 (b = 1200)	R2 (b = 1200)	R0 (b = 3000)	R2 (b = 3000)	MK	RTOP
SHInterp	21.6 (14.5)	8.1 (5.4)	14.3 (10.0)	40.8 (24.8)	19.7 (13.2)	43.3 (26.5)	12.1 (9.0)	14.7 (9.6)
DIAMOND-a	16.8 (11.1)	6.5 (4.4)	11.4 (7.9)	32.6 (19.2)	15.2 (9.7)	33.4 (20.6)	9.3 (6.8)	11.0 (7.0)
DIAMOND-b	16.9 (11.1)	6.5 (4.4)	11.3 (7.8)	32.5 (19.2)	15.2 (9.8)	33.6 (20.8)	9.3 (6.8)	11.0 (7.0)
SHOREInterp	22.1 (14.4)	6.8 (4.6)	11.9 (8.3)	41.0 (23.7)	17.3 (11.3)	42.5 (24.5)	10.0 (7.1)	13.2 (8.5)
LinearRISH	8.3 (5.3)	2.9 (2.0)	5.1 (3.6)	16.2 (9.9)	7.3 (4.9)	17.3 (10.6)	4.4 (3.2)	17.3 (10.1)
SH-1Layer	14.5 (9.5)	5.6 (3.8)	9.8 (6.7)	28.5 (17.5)	14.3 (9.6)	32.6 (18.2)	10.4 (7.4)	10.0 (6.5)
SHResNet-a	22.5 (14.7)	5.5 (3.8)	9.6 (6.7)	41.2 (24.0)	15.9 (10.9)	43.3 (24.7)	10.4 (7.7)	12.3 (8.2)
SHResNet-b0	15.9 (10.8)	5.1 (3.3)	8.8 (5.9)	31.9 (20.0)	12.0 (7.9)	35.1 (20.4)	8.0 (5.8)	8.8 (5.7)
SHResNet-b1	20.7 (13.9)	21.3 (9.5)	47.1 (24.9)	41.0 (25.7)	24.1 (16.9)	288.2 (284.8)	46.0 (29.3)	19.4 (12.8)
SHResNet-b2	15.8 (10.7)	5.3 (3.5)	9.1 (6.1)	31.4 (19.5)	17.3 (11.4)	41.8 (23.5)	15.6 (11.6)	13.2 (8.5)
SHResNet-b3	19.9 (13.0)	5.8 (3.9)	10.0 (7.1)	36.7 (21.7)	20.6 (13.8)	47.0 (25.3)	16.4 (11.9)	15.6 (10.3)
SHResNet-b4	15.0 (10.3)	4.9 (3.2)	8.5 (5.6)	29.7 (18.9)	14.5 (9.6)	33.3 (20.1)	12.6 (8.2)	10.2 (6.6)
SphericalNet	22.7 (15.0)	5.4 (3.7)	9.5 (6.6)	41.3 (24.3)	15.6 (10.6)	43.2 (25.2)	10.3 (7.7)	12.1 (8.1)
TinyPatch	28.4 (16.7)	5.9 (3.9)	10.1 (6.8)	49.4 (24.9)	17.6 (11.0)	50.6 (24.8)	10.6 (7.3)	13.9 (8.7)
DIQT	34.3 (16.2)	7.0 (4.7)	12.0 (8.0)	56.4 (22.9)	17.1 (11.1)	56.8 (23.0)	13.4 (10.7)	13.5 (8.4)
CMResNet-b	16.2 (10.6)	5.5 (3.6)	9.5 (6.4)	31.0 (18.5)	11.8 (7.8)	33.1 (19.4)	7.8 (5.9)	8.9 (5.7)

**Table 5**

Summary of the average percentage of APE values and two ROIs from the Desikan-Killiany atlas from FreeSurfer with the highest and lowest APE values for algorithms evaluated for Task 2a.

	APE (%)		ROIs with the highest APE	ROIs with the lowest APE
	GM	WM		
SHInterp	30.8	19.0	ctx-rh-postcentral, ctx-rh-superiorparietal	wm-rh-paracentral, wm-rh-parstriangularis
DIAMOND-a	22.8	15.3	ctx-rh-frontalpole, ctx-rh-superiorparietal	wm-lh-paracentral, wm-rh-paracentral
DIAMOND-b	22.8	15.2	ctx-rh-frontalpole, ctx-rh-superiorparietal	wm-lh-paracentral, wm-rh-paracentral
SHOREInterp	26.3	17.5	ctx-rh-superiorparietal, ctx-rh-postcentral	wm-rh-paracentral, wm-lh-paracentral
LinearRISH	12.7	8.6	ctx-rh-postcentral, ctx-rh-superiorparietal	wm-rh-paracentral, wm-lh-paracentral
SH-1Layer	22.9	13.6	ctx-lh-frontalpole, ctx-rh-inferiortemporal	wm-rh-paracentral, wm-lh-precentral
SHResNet-a	25.1	15.9	ctx-rh-superiorparietal, ctx-rh-postcentral	wm-rh-paracentral, wm-lh-paracentral
SHResNet-b0	18.2	13.0	ctx-lh-entorhinal, ctx-rh-parsorbitalis	wm-rh-paracentral, wm-rh-parsopercularis
SHResNet-b1	43.8	35.7	ctx-lh-inferiortemporal, ctx-rh-parsorbitalis	ctx-rh-caudalanteriorcing., ctx-rh-transversetemporal
SHResNet-b2	28.5	17.0	ctx-rh-postcentral, ctx-rh-transversetemporal	wm-rh-paracentral, wm-rh-insula
SHResNet-b3	32.2	20.0	ctx-rh-precentral, ctx-rh-postcentral	wm-rh-paracentral, wm-rh-rostralanteriorcing.
SHResNet-b4	23.1	14.6	ctx-rh-temporalpole, ctx-rh-unknown	wm-rh-paracentral, wm-rh-precentral
SphericalNet	24.9	15.8	ctx-lh-parsorbitalis, ctx-rh-superiorparietal	wm-rh-paracentral, wm-lh-paracentral
Tinypatch	24.2	20.1	wm-rh-entorhinal, wm-rh-entorhinal	wm-lh-transversetemporal, wm-lh-fusiform
DIQT	33.5	18.4	ctx-lh-parsorbitalis, ctx-rh-parsorbitalis	wm-rh-parsopercularis, wm-rh-supramarginal
CMResNet-b	19	12	ctx-rh-superiorparietal, ctx-rh-postcentral	wm-rh-paracentral, wm-lh-precentral

[iq.harvard.edu/files/cdmri2018/files/mushac\\_evaluation.zip](https://iq.harvard.edu/files/cdmri2018/files/mushac_evaluation.zip). The organizing team of MUSHAC can be contacted by email at [cdmri18@cs.ucl.ac.uk](mailto:cdmri18@cs.ucl.ac.uk).

### CRedit authorship contribution statement

**Lipeng Ning:** Conceptualization, Formal analysis, Writing - original draft, Writing - review & editing. **Elisenda Bonet-Carne:** Conceptualization, Writing - review & editing. **Francesco Grussu:** Conceptualization, Writing - review & editing. **Farshid Sepehrband:** Conceptualization, Writing - review & editing. **Enrico Kaden:** Conceptualization. **Jelle Veraart:** Conceptualization. **Stefano B. Blumberg:** Methodology. **Can Son Khoo:** Methodology. **Marco Palombo:** Methodology. **Iasonas Kokkinos:** Methodology. **Daniel C. Alexander:** Methodology. **Jaume Coll-Font:** Methodology. **Benoit Scherrer:** Methodology. **Simon K. Warfield:** Methodology. **Suheyyla Cetin Karayumak:** Methodology. **Yogesh Rathi:** Methodology. **Simon Koppers:** Methodology. **Leon Weninger:** Methodology. **Julia Ebert:** Methodology. **Dorit Merhof:** Methodology. **Daniel Moyer:** Methodology. **Maximilian Pietsch:** Methodology. **Daan Christiaens:** Methodology. **Rui Azeredo Gomes Teixeira:** Methodology. **Jacques-Donald Tournier:** Methodology. **Kurt G. Schilling:** Methodology. **Yuankai Huo:** Methodology. **Vishwesh Nath:** Methodology. **Colin Hansen:** Methodology. **Justin Blaber:** Methodology. **Bennett A. Landman:** Methodology. **Andrey Zhylka:** Methodology. **Josien P.W. Pluim:**

**Methodology.** **Greg Parker:** Resources. **Umesh Rudrapatna:** Resources. **John Evans:** Resources. **Cyril Charron:** Resources. **Derek K. Jones:** Resources. **Chantal M.W. Tax:** Conceptualization, Data curation, Investigation, Writing - review & editing, Resources.

### Acknowledgements

CMWT is supported by a Rubicon grant (680-50-1527) from the Netherlands Organisation for Scientific Research (NWO) and Wellcome Trust grant (096646/Z/11/Z). This project has received funding under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 634541 and from Engineering and Physical Sciences Research Council (EPSRC EP/R006032/1, M020533/1), funding FG. DKJ were supported by MRC grant MR/K004360/1. Scan costs were supported by the National Centre for Mental Health (NCMH) with funds from Health and Care Support Wales and by the Wellcome Trust. JV is a Postdoctoral Fellow of the Research Foundation - Flanders (FWO; grant number 12S1615N). LN is supported in part by NIH grants R21MH115280, R21MH116352 and K01MH117346. SCK and YR are supported in part by NIH grant R01MH119222. AZ and JP have received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 765148. KS is supported in part by NIH grants R01EB017230, and T32EB001628. FS is supported in part by USC ADRG 5P50AG005142 and R01NS100973. EK acknowledges support from UK EPSRC EP/M020533/1, EP/N018702/1 and EU H2020 634541. DC is supported by the Flemish Research Foundation (FWO; fellowship number 12ZV420N). MP is funded by UKRI Future Leaders Fellowship MR/T020296/1. The data were acquired at the UK National Facility for In Vivo MR Imaging of Human Tissue Microstructure located in CUBRIC funded by the EPSRC (grant EP/M029778/1), and The Wolfson Foundation. Acquisition and processing of the data was supported by a Rubicon grant from the NWO (680-50-1527), a Wellcome Trust Investigator Award (096646/Z/11/Z), and a Wellcome Trust Strategic Award (104943/Z/14/Z).

### Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.neuroimage.2020.117128>.

### References

- Alexander, Daniel C., Zikic, Darko, Ghosh, Aurobrata, Tanno, Ryutaro, Wotschel, Viktor, Zhang, Jiaying, Kaden, Enrico, et al., 2017. Image quality transfer and applications in diffusion MRI. *Neuroimage* 152 (May), 283–298.
- Andersson, Jesper L.R., Sotiropoulos, Stamatios N., 2016. An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *Neuroimage* 125 (January), 1063–1078.
- Andersson, Jesper L.R., Skare, Stefan, Ashburner, John, 2003. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *Neuroimage* 20 (2), 870–888.
- Asman, A.J., Landman, B.A., 2014. Hierarchical performance estimation in the statistical label fusion framework. *Med. Image Anal.* 18 (7), 1070–1081.
- Avants, Brian B., Tustison, Nicholas J., Song, Gang, Cook, Philip A., Klein, Arno, Gee, James C., 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage* 54 (3), 2033–2044.
- Bammer, R., Markl, M., Barnett, A., Acar, B., Alley, M.T., Pelc, N.J., Glover, G.H., Moseley, M.E., 2003. Analysis and generalized correction of the effect of spatial gradient field distortions in diffusion-weighted imaging. *Magnetic Resonance in Medicine*. <https://doi.org/10.1002/mrm.10545>.
- Blumberg, Stefano B., Tanno, Ryutaro, Kokkinos, Iasonas, Alexander, Daniel C., 2018. Deeper image quality transfer: training low-memory neural networks for 3D images. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Springer International Publishing, 118–25.
- Blumberg, S.B., Palombo, M., Khoo, C.S., Tax, C.M.W., Tanno, R., Alexander, D.C., 2019. Multi-stage prediction networks for data harmonisation. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2019*, pp. 411–419.
- Boeckel, W., Forstmann, B.U., Keuken, M.C., 2017. A test-retest reliability analysis of diffusion measures of white matter tracts relevant for cognitive control. *Psychophysiology*. <https://doi.org/10.1111/psyp.12769>.
- Breiman, L., 2001. Random forests. *Mach. Learn.* 45 (1), 5–32.
- Cardoso, M. Jorge, Modat, Marc, Wolz, Robin, Melbourne, Andrew, Cash, David, Rueckert, Daniel, Ourselin, Sebastien, 2015. Geodesic information flows: spatially-

- variant graphs and their application to segmentation and fusion. *IEEE Trans. Med. Imag.* 34 (9), 1976–1988.
- Cheng, Jian, Jiang, Tianzi, Deriche, Rachid, 2011. Theoretical analysis and practical insights on eap estimation via a unified hardi framework. In: *MICCAI Workshop on Computational Diffusion MRI (CDMRI)*. <https://hal.inria.fr/inria-00615430/>.
- Christiansen, Daan, Cordero-Grande, Lucilio, Hutter, Jana, Price, Anthony N., Deprez, Maria, Hajnal, Joseph V., Donald Tournier, J., 2018. Learning compact Q-space representations for multi-shell diffusion-weighted MRI. In: *IEEE Transactions on Medical Imaging*, October. <https://doi.org/10.1109/TMI.2018.2873736>.
- Chuhutin, Andrey, Hanson, Brian, Jespersen, Sune N., 2017. Precision and accuracy of diffusion kurtosis estimation and the influence of b-value selection. *NMR in Biomedicine* 30 (11), e3777.
- Duan, Fei, Zhao, Tengda, He, Yong, Ni, Shu, 2015. Test-retest reliability of diffusion measures in cerebral white matter: a multiband diffusion MRI study. *J. Magn. Reson. Imag.* <https://doi.org/10.1002/jmri.24859>.
- Fischl, Bruce, 2012. FreeSurfer. *Neuroimage* 62 (2), 774–781.
- Fortin, Jean-Philippe, Parker, Drew, Tunc, Birkhan, Watanabe, Takanori, Elliott, Mark A., Ruparel, Kosha, Roalf, David R., et al., 2017. Harmonization of multi-site diffusion tensor imaging data. *Neuroimage* 161, 149–170. November.
- Garyfallidis, Eleftherios, Brett, Matthew, Amirbekian, Bagrat, Rokem, Ariel, van der Walt, Stefan, Descoteaux, Maxime, Nimmo-Smith, Ian, Contributors, Dipy, 2014. Dipy, a library for the analysis of diffusion MRI data. *Front. Neuroinf.* 8 (February), 8.
- Glasser, Matthew F., Sotiropoulos, Stamatis N., Wilson, J. Anthony, Coalson, Timothy S., Fischl, Bruce, Andersson, Jesper L., Xu, Junqian, et al., 2013. The minimal preprocessing pipelines for the human connectome project. *Neuroimage* 80 (October), 105–124.
- Grech-Sollars, Matthew, Hales, Patrick W., Miyazaki, Keiko, Raschke, Felix, Rodriguez, Daniel, Wilson, Martin, Gill, Simrandip K., et al., 2015. Multi-centre reproducibility of diffusion MRI parameters for clinical sequences in the brain. *NMR Biomed.* 28 (4), 468–485.
- He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, Sun, Jian, 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778.
- Hu, Jie, Shen, Li, Albanie, Samuel, Sun, Gang, Wu, Enhua, 2017. Squeeze-and-Excitation networks. *arXiv [cs.CV]*. [arXiv. http://arxiv.org/abs/1709.01507](http://arxiv.org/abs/1709.01507).
- Huo, Yuankai, Xu, Zhoubing, Aboud, Katherine, Parvathaneni, Prasanna, Bao, Shunxing, Bermudez, Camilo, Resnick, Susan M., Cutting, Laurie E., Landman, Bennett A., 2018. Spatially localized atlas network tiles enables 3D whole brain segmentation from limited data. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Springer International Publishing, pp. 698–705.
- Jenkinson, M., M Smith, S., 2001. A global optimisation method for robust affine registration of brain images. *Med. Image Anal.* 5 (2), 143–156.
- Jenkinson, Mark, Pechaud, Mickael, Smith, Stephen, Others, 2005. BET2: MR-based estimation of brain, skull and scalp surfaces. In: *Eleventh Annual Meeting of the Organization for Human Brain Mapping*, vol. 17, p. 167 (Toronto).
- Jenkinson, Mark, Beckmann, Christian F., Timothy, E., Behrens, J., Woolrich, Mark W., Smith, Stephen M., 2012. “FSL.” *NeuroImage*. <https://doi.org/10.1016/j.neuroimage.2011.09.015>.
- Jensen, Jens H., Helpert, Joseph A., Ramani, Anita, Lu, Hanzhang, Kaczynski, Kyle, 2005. Diffusional kurtosis imaging: the quantification of non-Gaussian water diffusion by means of magnetic resonance imaging. *Magn. Reson. Med.*: Official Journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine 53 (6), 1432–1440.
- Jones, D.K., Alexander, D.C., Bowtell, R., Cercignani, M., Dell’Acqua, F., McHugh, D.J., Miller, K.L., Palombo, M., M Parker, G.J., Rudrapatna, U.S., Tax, C.M.W., 2018. Microstructural imaging of the human brain with a ‘super-scanner’: 10 key advantages of ultra-strong gradients for diffusion MRI. *Neuroimage* 182, 8–38.
- Karayumak, Cetin, Suheyla, Sylvain Bouix, Ning, Lipeng, James, Anthony, Tim Crow, Shenton, Martha, Kubicki, Marek, Rath, Yogesh, 2019. Retrospective harmonization of multi-site diffusion MRI data acquired with different acquisition parameters. *Neuroimage* 184 (January), 180–200.
- Kellner, Elias, Dhital, Bibek, Kiselev, Valerij G., Reisert, Marco, 2016. Gibbs-ringing artifact removal based on local subvoxel-shifts. *Magn. Reson. Med.* 76 (5), 1574–1581.
- Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.C., Christensen, G.E., Collins, D.L., Gee, J., Hellier, P., Song, J.H., Jenkinson, M., Lepage, C., Rueckert, D., Thompson, P., Vercauteren, T., Woods, R.P., Mann, J.J., Parsey, R.V., 2009. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *Neuroimage* 46 (3), 786–802.
- Koppers, Simon, Merhof, Dorit, 2018. DELIMIT PyTorch - an Extension for Deep Learning in Diffusion Imaging *arXiv [cs.LG]*. [arXiv. http://arxiv.org/abs/1808.01517](http://arxiv.org/abs/1808.01517).
- Koppers, Simon, Bloy, Luke, Berman, Jeffrey I., Chantal, M.W., Christopher Edgar, Tax, J., Merhof, Dorit, 2018. Spherical Harmonic Residual Network for Diffusion Signal Harmonization *arXiv [cs.CV]*. [arXiv. http://arxiv.org/abs/1808.01595](http://arxiv.org/abs/1808.01595).
- Kubicki, Marek, McCarley, Robert, Fredrik Westin, Carl, Park, Hae-Jeong, Maier, Stephan, Kikinis, Ron, Jolesz, Ferenc A., Shenton, Martha E., 2007. A review of diffusion tensor imaging studies in schizophrenia. *J. Psychiatr. Res.* 41 (1–2), 15–30.
- Landman, Bennett A., Huang, Alan J., Gifford, Aliya, Deepti, S., Vikram, Lim, Issel Anne L., Farrell, Jonathan A.D., Bogovic, John A., et al., 2011. Multi-parametric neuroimaging reproducibility: a 3-T resource study. *Neuroimage* 54 (4), 2854–2866.
- Liu, Ming-Yu, Breuel, Thomas, Kautz, Jan, 2017. Unsupervised image-to-image translation networks. *arXiv [cs.CV]*. [arXiv. http://arxiv.org/abs/1703.00848](http://arxiv.org/abs/1703.00848).
- Meijering, Erik H.W., Wiro, J., Niessen, Josien, Pluim, P.W., Viergever, Max A., 1999. Quantitative comparison of sinc-approximating kernels for medical image interpolation. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI’99*, 210–217. Springer Berlin Heidelberg.
- Merlet, Sylvain L., Deriche, Rachid, 2013. Continuous diffusion signal, EAP and ODF estimation via compressive sensing in diffusion MRI. *Med. Image Anal.* 17 (5), 556–572.
- Mirzaalian, H., Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., Grant, G., Marx, C.E., Morey, R.A., Flashman, L.A., George, M.S., McAllister, T.W., Andaluz, N., Shutter, L., Coimbra, R., Zafonte, R.D., Coleman, M.J., Kubicki, M., Westin, C.-F., Stein, M.B., Shenton, M.E., Rath, Y., 2016. Inter-site and inter-scanner diffusion MRI data harmonization. *Neuroimage* 135 (July), 311–323.
- Mirzaalian, H., Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., Karmacharya, S., Grant, G., Marx, C.E., Morey, R.A., Flashman, L.A., George, M.S., McAllister, T.W., Andaluz, N., Shutter, L., Coimbra, R., Zafonte, R.D., Coleman, M.J., Kubicki, M., Westin, C.-F., Stein, M.B., Shenton, M.E., Rath, Y., 2018. Multi-site harmonization of diffusion MRI data in a registration framework. *Brain Imaging and Behavior* 12 (1), 284–295.
- Moyer, D., Ver Steeg, G., Tax, C.M.W., Thompson, P.M., 2020. Scanner invariant representations for diffusion MRI harmonization. *Magn. Reson. Med.* 84, 2174–2189, 4.
- Mueller, Susanne G., Weiner, Michael W., Thal, Leon J., Petersen, Ronald C., Jack, Clifford R., Jagust, William, Trojanowski, John Q., Toga, Arthur W., Beckett, Laurel, 2005. “Ways toward an early diagnosis in alzheimer’s disease: the alzheimer’s disease neuroimaging initiative (ADNI).” *alzheimer’s & dementia*. <http://doi.org/10.1016/j.jalz.2005.06.003>.
- Nath, W., Lyu, I., Schilling, K.G., Parvathaneni, P., Hansen, C.B., Tang, Y., Huo, Y., Janve, V.A., Gao, Yurui, Stepniowska, I., Anderson, A.W., Landman, B.A., 2019. Enabling multi-shell b-value generalizability of data-driven diffusion models with deep SHORE. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI’19*.
- Ning, Lipeng, Carl-Fredrik, Westin, Yogesh, Rath, 2015. Estimating diffusion propagator and its moments using directional radial basis functions. *IEEE Trans. Med. Imag.* 34 (10), 2058–2078.
- Ning, Lipeng, Bonet-Carne, Elisenda, Grussu, Francesco, Seppebrand, Farshid, Kaden, Enrico, Veraart, Jelle, Blumberg, Stefano B., et al., 2019. Multi-shell diffusion MRI harmonization and enhancement challenge (MUSHAC): progress and results. In: *In Computational Diffusion MRI*, vols. 217–244. Springer International Publishing.
- Özarslan, Evren, Guan, Cheng, Timothy, Koay, Shepherd, M., Komolosh, Michael E., Okan Yrfanoğlu, M., Pierpaoli, Carlo, Peter, J., Basser, 2013. Mean apparent propagator (MAP) MRI: a novel diffusion imaging method for mapping tissue microstructure. *Neuroimage* 78 (September), 16–32.
- Pasternak, Ofer, Kelly, Sinead, Sydnor, Valerie J., Shenton, Martha E., 2018. Advances in microstructural diffusion neuroimaging for psychiatric disorders. *Neuroimage* 182 (November), 259–282.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.
- Pohl, Kilian M., Sullivan, Edith V., Rohlfing, Torsten, Chu, Weiwei, Kwon, Dongjin, Nolan Nichols, B., Zhang, Yong, et al., 2016. Harmonizing DTI measurements across scanners to examine the development of white matter microstructure in 803 adolescents of the NCANDA study. *Neuroimage* 130 (April), 194–213.
- Raffelt, David, Donald Tournier, J., Fripp, Jurgen, Crozier, Stuart, Connelly, Alan, Salvado, Olivier, 2011. Symmetric diffeomorphic registration of fibre orientation distributions. *Neuroimage* 56 (3), 1171–1180.
- Rudrapatna, S.U., Parker, G.D., Roberts, J., Jones, D.K., 2018. Can we correct for interactions between subject motion and gradient-nonlinearity in diffusion MRI. *Proceedings of the International Society for Magnetic Resonance in Medicine ... Scientific Meeting and Exhibition. International Society for Magnetic Resonance in Medicine. Scientific Meeting and Exhibition*.
- Scherrer, Benoit, Schwartzman, Armin, Taquet, Maxime, Sahin, Mustafa, Prabhu, Sanjay P., Warfield, Simon K., 2016. Characterizing brain tissue by assessment of the distribution of anisotropic microstructural environments in diffusion-compartment imaging (DIAMOND). *Magn. Reson. Med.* <https://doi.org/10.1002/mrm.26372>.
- Scherrer, Benoit, Taquet, Maxime, Schwartzman, Armin, St-Onge, Etienne, Rensonnet, Gaetan, Prabhu, Sanjay P., Warfield, Simon K., 2017. Decoupling axial and radial tissue heterogeneity in diffusion compartment imaging. In: *In Information Processing in Medical Imaging*, vols. 440–52. Springer International Publishing.
- Setsonpopp, K., Kimmlingen, R., Eberlein, E., Witzel, T., Cohen-Adad, J., McNab, J.A., Keil, B., Tisdall, M.D., Hoeft, P., Dietz, P., Cauley, S.F., Tountcheva, V., Matschl, V., Lenz, V.H., Heberlein, K., Potthast, A., Thein, H., Van Horn, J., Toga, A., Schmitt, F., Lehne, D., Rosen, B.R., Wedeen, V., Wald, L.L., 2013. Pushing the limits of in vivo diffusion MRI for the human connectome project. *Neuroimage* 80, 220–233.
- Shahim, Pashtun, Holleran, Laurena, Kim, Joong H., Brody, David L., 2017. Test-retest reliability of high spatial resolution diffusion tensor and diffusion kurtosis imaging. *Sci. Rep.* 7 (1), 11141.
- Sotiropoulos, Stamatis N., Jbabdi, Saad, Xu, Junqian, Andersson, Jesper L., Moeller, Steen, Auerbach, Edward J., Glasser, Matthew F., et al., 2013. Advances in diffusion MRI acquisition and processing in the human connectome project. *Neuroimage* 80 (October), 125–143.
- Tax, Chantal M.W., Grussu, Francesco, Kaden, Enrico, Ning, Lipeng, Rudrapatna, Umesh, John Evans, C., St-Jean, Samuel, et al., 2019. Cross-scanner and cross-protocol diffusion MRI data harmonisation: a benchmark database and evaluation of algorithms. *Neuroimage* 195 (July), 285–299.
- Teipel, Stefan J., Reuter, Sigrid, Stieltjes, Bram, Acosta-Cabrero, Julio, Ernemann, Ulrike, Fellgiebel, Andreas, Filippi, Massimo, Frisoni, Giovanni, Hentschel, Frank, Jessen, Frank, Klöppel, Stefan, Meindl, Thomas, Pouwels, Petra

- J.W., Hauenstein, Karl-Heinz, Hampel, Harald, 2011. Multicenter stability of diffusion tensor imaging measures: a European clinical and physical phantom study. *Psychiatr. Res. Neuroimaging* 194 (30), 363–371.
- Tournier, J-Donald, Calamante, Fernando, Gadian, David G., Alan, Connelly, 2004. Direct estimation of the fiber orientation density function from diffusion-weighted MRI data using spherical deconvolution. *Neuroimage* 23 (3), 1176–1185.
- Tustison, Nicholas J., Avants, Brian B., Cook, Philip A., Zheng, Yuanjie, Alexander, Egan, Yushkevich, Paul A., Gee, James C., 2010. N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imag.* 29 (6), 1310–1320.
- Veenith, Tonny V., Carter, Eleanor, Grossac, Julia, Newcombe, Virginia F.J., Outtrim, Joanne G., Lupson, Victoria, Williams, Guy B., Menon, David K., Coles, Jonathan P., 2013. Inter subject variability and reproducibility of diffusion tensor imaging within and between different imaging sessions. *PLoS One* 8 (6), e65941.
- Veraart, Jelle, Fieremans, Els, Novikov, Dmitry S., 2016. Diffusion MRI noise mapping using random matrix theory. *Magn. Reson. Med.: Official Journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine* 76 (5), 1582–1593.
- Vollmar, Christian, O’Muircheartaigh, Jonathan, Barker, Gareth J., Symms, Mark R., Thompson, Pamela, Kumari, Veena, Duncan, John S., Richardson, Mark P., Koepp, Matthias J., 2010. Identical, but not the same: intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0T scanners. *Neuroimage* 51 (4), 1384–1394.
- Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imag.* 20 (1), 45–57.
- Zhou, Tinghui, Krähenbühl, Philipp, Aubry, Mathieu, Huang, Qixing, Alexei, A., Efros, 2016. Learning dense correspondence via 3D-guided cycle consistency. *arXiv [cs.CV]*. [arXiv. http://arxiv.org/abs/1604.05383](http://arxiv.org/abs/1604.05383).
- Zhu, Tong, Hu, Rui, Qiu, Xing, Taylor, Michael, Tso, Yuen, Yiannoutsos, Constantin, Navia, Bradford, et al., 2011. Quantification of accuracy and precision of multi-center DTI measurements: a diffusion phantom and human brain study. *Neuroimage* 56 (3), 1398–1411.