Received: 9 June 2019 | Revised: 7 February 2020 | Accepted: 11 February 2020

DOI: 10.1002/mrm.28243

FULL PAPER

Magnetic Resonance in Medicine

# Scanner invariant representations for diffusion MRI harmonization

**Daniel Moyer**[1,2] | **Greg Ver Steeg**[2] | **Chantal M. W. Tax**[3] | **Paul M. Thompson**[1]

[1]Imaging Genetics Center, Mark and Mary Stevens Institute for Neuroimaging and Informatics, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

[2]Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA

[3]Cardiff University Brain Research Imaging Centre (CUBRIC), Cardiff University, Cardiff, United Kingdom

**Correspondence**
Paul M. Thompson, Imaging Genetics Center, Mark and Mary Stevens Institute for Neuroimaging and Informatics, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA.
Email: pthomp@usc.edu

**Funding information**
NIH (U.S. National Institutes of Health), Grant/Award Number: P41 EB015922, R01 MH116147, R56 AG058854, U01 AG024904, RF1 AG041915 and U54 EB020403; NSF Graduate Research Fellowship Program Grant, Grant/Award Number: DGE-1418060; DARPA, Grant/Award Number: W911NF-16-1-0575; NVidia; EPSRC, Grant/Award Number: EP/M029778/1; NWO, Grant/Award Number: 680-50-1527; Wellcome Trust Investigator, Grant/Award Number: 096646/Z/11/Z; Wellcome Trust Strategic, Grant/Award Number: 104943/Z/14/Z

**Purpose:** In the present work, we describe the correction of diffusion-weighted MRI for site and scanner biases using a novel method based on invariant representation.

**Theory and Methods:** Pooled imaging data from multiple sources are subject to variation between the sources. Correcting for these biases has become very important as imaging studies increase in size and multi-site cases become more common. We propose learning an intermediate representation invariant to site/protocol variables, a technique adapted from information theory-based algorithmic fairness; by leveraging the data processing inequality, such a representation can then be used to create an image reconstruction that is uninformative of its original source, yet still faithful to underlying structures. To implement this, we use a deep learning method based on variational auto-encoders (VAE) to construct scanner invariant encodings of the imaging data.

**Results:** To evaluate our method, we use training data from the 2018 MICCAI Computational Diffusion MRI (CDMRI) Challenge Harmonization dataset. Our proposed method shows improvements on independent test data relative to a recently published baseline method on each subtask, mapping data from three different scanning contexts to and from one separate target scanning context.

**Conclusions:** As imaging studies continue to grow, the use of pooled multi-site imaging will similarly increase. Invariant representation presents a strong candidate for the harmonization of these data.

**KEYWORDS**
diffusion MRI, harmonization, invariant representation

# 1 | INTRODUCTION

Observational conditions may vary strongly within a medical imaging study. Researchers are often aware of these conditions (eg, scanner, site, technician, facility) but are unable to modify the experimental design to compensate, due to cost or geographic necessity. In magnetic resonance imaging (MRI), variations in scanner characteristics such as the magnetic field strength, scanner vendor, receiver coil hardware, applied gradient fields, or primary image reconstruction methods may have strong effects on collected data[1-3]; multi-site studies in particular are subject to these effects.[4-7] Data harmonization is the process of removing or compensating for this unwanted variation through post hoc corrections. In the present work we focus on harmonization for diffusion MRI (dMRI), a modality known to have scanner/site biases[8-16] as well as several extra possible degrees of freedom with respect to protocol (eg, angular resolution, *b*-values, gradient waveform choice).

Several prior methods approach diffusion MRI harmonization as a regression problem. Supervised image-to-image transfer methods have been proposed,[17,18] while for the unsupervised case site effects are often modeled as covariate effects, either at a summary statistic level[2,7] or on the image directly.[19] All of these methods directly transform scans from one site/scanner context to another. Further, while all methods require paired scans to correctly validate their results (subjects or phantoms scanned on both target and reference scanners), supervised methods also require paired training data. The collection of such data is expensive and difficult to collect at a large scale.

In this paper, we instead frame the harmonization problem as an unsupervised image-to-image transfer problem, where harmonizing transformations may be learned without explicitly paired scans. We propose that a subset of harmonization solutions may be found by learning scanner invariant representations, that is, representations of the images that are uninformative of which scanner the images were collected on. These representations and the mappings between them may then be manipulated to provide image reconstructions that are minimally informative of their original collection site. We thus provide an encoder/decoder method for learning mappings to and from invariant representations computationally. This method has several advantages over regression-based methods, including a practical implementation that does not require paired data, that is, a traveling phantom as training input, and an extension to a multi-site case.

We demonstrate our proposed method on the MICCAI Computational Diffusion MRI challenge dataset,[20-22] showing substantial improvement compared to a recently published baseline method. We also introduce technical improvements to the training of neural architectures on diffusion-weighted data, and discuss the limitations and error modes of our proposed method.

## 1.1 | Relevant prior work

Harmonization has been an acknowledged problem in MR imaging and specifically diffusion imaging for some time.[22] Numerous studies have noted significant differences in diffusion summary measures (eg, fractional anisotropy; FA) between scanners and sites.[10,12,13] Further protocol differences arise between sites due to limitations of the available scanners, unavoidable changes or upgrades in scanners or protocols, or when combining data retrospectively from multiple studies; effects of variations in scanning protocols on derived measures include effects of voxel size,[11] *b*-values (the diffusion weightings used),[8,11] and angular resolution or *q*-space sampling[9,14-16] among other parameters. These problems were also examined by the MICCAI Computational Diffusion MRI 2017 and 2018 challenges,[20,21] which held an open comparison of methods for a supervised (paired) task.

Most previously proposed harmonization methods have relied on forms of regression. Harmonization of summary statistics (voxel-wise or region-wise) include random/mixed-effect models[7] as well as the scale-and-shift random effects regression of ComBat.[2,7] This latter method was adapted from the genomics literature,[23] and employs a variational Bayes scheme to learn model coefficients.

A more nuanced family of regression methods for diffusion imaging was recently introduced in a series of papers by Mirzaalian et al.[19,24,25] This was later analyzed empirically in Karayumak et al,[26] which compared it against ComBat[23] for summary statistics. This family of methods computes a power spectrum from a spherical harmonic (SH) basis, then generates a template from these images using multi-channel diffeomorphic mappings. The resulting template is used to compute spatial maps of average SH power spectra by scanner/protocol, which are then used in a scale regression on individual subjects. While these papers take a very different approach from our own, the resulting method has a very similar usage pattern and output. We compare our approach directly to this method.

In a supervised (paired) task, direct image-to-image transfer has been explored both in the harmonization context[20,27,28] as well as the similar super-resolution context.[17,18] This family of methods generally relies on high expressive capacity function fitting (eg, neural networks) to map directly between patches of pairs of images. This requires explicitly paired data, in that the same brains must be scanned at all sites. These methods perform well empirically, as tested by the CDMRI challenge,[21] but require paired data in the training set. Our proposed method does not require paired data to

train; however, in our opinion, best practice validation still requires paired data in the (holdout) test set.

## 2 | THEORY

Our goal is to map diffusion MRI scans from one scanner/site context to another, so that given an image from one site we could predict accurately what it would have looked like were it collected at another site. In order to do this, we construct an encoding function $q$ that takes each image $x$ to a corresponding vector $z$, and a conditional decoding function $p$ that takes each $z$ and a specified site $s$ back to an image $\hat{x}$ (the "reconstruction" of the original image).

We further wish to remove trends and biases in $x$ that are informative of $s$ from the reconstruction $\hat{x}$, so that all data remapped to a given site $s'$ have the same bias (this is the harmonization task). In order to do so, it would be sufficient to constrain $z$, the intermediate representation, to be independent of $s$, denoted $z \perp s$. This is a hard constraint, and direct optimization of $q$ and $p$ subject to that constraint would be non-trivially difficult.

Instead, we choose to relax the constraint $z \perp s$ to the mutual information $I(z, s)$. Mutual information, taken from information theory, quantifies the amount of information shared between two variables, for example, $z$ and $s$. $I(z, s) = 0$ if and only if $z \perp s$, and so its minimization is a relaxation of our desired constraint. For a comprehensive reference on information theory, we refer the reader to Chapters 2 and 8 of Cover and Thomas.[29]

After relaxing the independence constraint to mutual information, we would like to optimize $q$ and $p$ so that $q(z|x)$ has minimal scanner-specific information, and so that $p(x|z)$ has minimal differences from the original data. We demonstrate one solution for doing this using a variational bound on $I(z, s)$, parameterizing $p$ and $q$ using neural networks. The underlying theory is explored in Moyer et al,[30] where it is used in the context of algorithmic fairness. We reproduce it here for clarity, and further reinterpret their theoretical results in the imaging harmonization context, adding our own data processing inequality interpretation of test time remapping.

Learning the mapping $q$ does *not* require matching pairs of data $(x, x')$ from pairs of sites $(s, s')$. Best practices in validation and testing *do* require such data, but during training we can minimize $I(z, s)$ without having examples of the same subject collected on different scanners. This is due to our bound of $I(z, s)$ described in Equation 1, which is not reliant on inter-site correspondence.

At test time, we can manipulate this mapping to reconstruct images at a different site than they were originally collected at; we use this mapping as our harmonization tool. Again, by the data processing inequality, the amount of information these (new) reconstructed images contain about their original collection site is bounded by $I(z, s)$, which we explicitly minimize.

### 2.1 | Scanner invariant variational auto-encoders

We wish to learn a mapping $q$ from data $x$ (associated with scanner $s$) to some latent space $z$ such that $z \perp s$, yet also where $z$ is maximally relevant to $x$. We start by relaxing $z \perp s$ to $I(z, s)$, and then bounding $I(z, s)$ (detailed demonstration in Appendix A):

$$I(z,s) \leq \underbrace{-\mathbb{E}_{x,s,z\sim q}[\log p(x|z,s)]}_{\text{Conditional Reconstruction}} + \underbrace{\mathbb{E}_x[\, KL[\, q(z|x) \parallel q(z)\,]\,]}_{\text{Compression}} - \underbrace{H(x|s)}_{\text{Const}} \tag{1}$$

where $q(z)$ is the empirical marginal distribution of $z$ under $q(z|x)$, the specified encoding which we control, and $p(x|z, s)$ is a variational approximation to the conditional likelihood of $x$ given $z$ and $s$ again under $q(z|x)$. Here, $KL$ denotes the Kullback-Leibler divergence and $H$ denotes Shannon entropy.

The bound in Equation 1 has three components: a conditional reconstruction, a compressive divergence term, and a constant term denoting the conditional entropy of the scan given the scanner. We can interpret Equation 1 as stating that the information in $z$ about $s$ is bounded above by uncertainty of $x$ given $z$ and $s$, plus a penalty on the information in $z$ and a constant representing the information $s$ has about $x$ overall.

Intuitively, this breakdown makes sense: if we reconstruct given $s$, and are otherwise compressing $z$, the optimal compressive $z$ has no information about $s$ for reconstruction; $q(z|x)$ can always remove information about $s$ without penalty, because the reconstruction term is handed that information immediately. Further, if $x$ is highly correlated with $s$, that is, $H(x|s)$ is very low, then our bound will be worse.

We can now construct a variational encoding/conditional-decoding pair $q$ and $p$ which fits our variational bound of $I(z, s)$ nicely, and which also fits our overall goal of re-mapping $x$ accurately through $p(x|z, s)$. Following Kingma and Welling,[31] we use a generative log-likelihood as an objective:

$$\max \log \mathbb{E}_{(x,s)}[p(x|s)] \tag{2}$$

Here, however, we inject the conditional likelihood to match our bound for $I(z, s)$. This also fits our test time desired Markov chain (with condition $z \perp s$) where $\hat{x}$ is the harmonized reconstruction at new site $s'$:

$$s \rightarrow x \rightarrow z \rightarrow \hat{x} \leftarrow s'$$

Following the original VAE derivation (again in Kingma and Welling), we can derive a similar VAE with $s$-invariant encodings by introducing the encoder $q(z|x)$:

$$\log p(x|s) = \log \int \frac{p(x,z|s)}{q(z|x)} q(z|x) dz = \log \mathbb{E}_{z\sim q}\left[\frac{p(x,z|s)}{q(z|x)}\right] \tag{3}$$

$$\geq \mathbb{E}_{z\sim q}[\log p(x,z|s) - \log q(z|x)] \tag{4}$$

$$= \mathbb{E}_{z\sim q}[\log p(z|s) - \log q(z|x) + \log p(x|z,s)]. \tag{5}$$

we assume that the prior $p(z|s) = p(z)$, that is, that the conditional prior is equal to the marginal prior over $z$. In the generative context, this would be a strong model mis-specification: if we believe that there truly are generating latent factors, it is unlikely that those factors would be independent of $s$. However, we are not in such a generative frame, and instead would like to find a code $z$ that is invariant to $s$, so it is reasonable to use a prior that also has this property.

Taking this assumption, we have

$$\log p(x|s) \geq -KL[\, q(z|x) \parallel p(z)] + \mathbb{E}_{z\sim q}[\log p(x|z,s)]. \tag{6}$$

This is a conditional extension of the VAE objective from Kingma and Welling.[31] Putting this objective together with the penalty term in Equation 1, we have the following variational bound on the combined objective (up to a constant):

$$\begin{aligned} &\mathbb{E}_{(x,s)}[\log P(x|s)] - \lambda I(z,s) \geq \\ &\mathbb{E}_{(x,s)}[\underbrace{-KL[q(z|x)\|p(z)]}_{\text{Div. from Prior}} \underbrace{- \lambda KL[q(z|x)\|q(z)]}_{\text{Div. from Marg.}} \\ &\quad + (1+\lambda)\underbrace{\mathbb{E}_{z\sim q}[\log p(x|z,s)]}_{\text{Cond. Reconstruction}}]. \end{aligned} \tag{7}$$

We use the negation of Equation 7 as the main loss term for learning $q$ and $p$, where we want to minimize the negative of the bound. As described in Higgins et al,[32] an additional parameter $\alpha$ may be multiplied with the divergence from the prior (the first term of Equation 7) for further control over the VAE prior.

As we have it written in Equation 7, the site variable $s$ has ambiguous dimension. For applications with only two sites, $s$ might be binary, while in the multi-site case, $s$ might be a one-hot vector (For a categorical variable with value $k$ out of $K$ possible values, its corresponding one-hot vector is a $K$-dimensional vector with zeros in every entry except for the $k$th entry, which is one). We conduct experiments for both in Sections 3 and 4. More complex $s$ values are also possible, but we do not explore them in this paper.

## 2.2 | Diffusion-space Error Propagation from SH representations

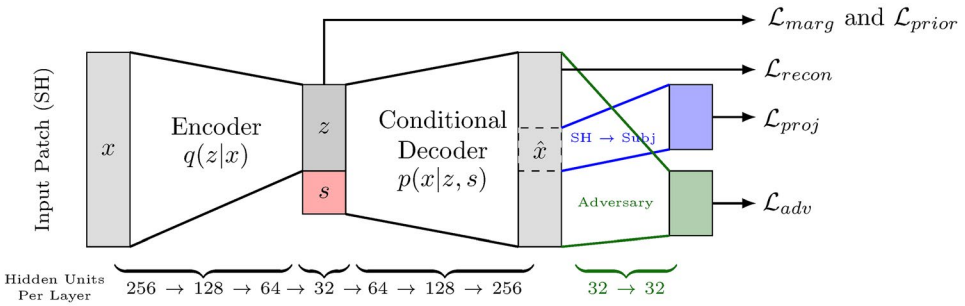
A convenient representation for diffusion-weighted MRI is the spherical harmonics (SH) basis.[24] These provide a countable set of basis functions from the sphere to and from which projection is easy and often performed (eg, in graphics). In this paper, our input data and the reconstruction error is computed with respect to the SH coefficients. However, for the eventual output, the data representation that we would like to use is *not* in this basis, but in the original image representation which is conditional on a set of gradient vectors (b-vectors). These vectors are in general different for each subject due to spatial positioning and motion, and often change in number between sites/protocols. Rigid transformation and alignment of scan data, used in many pre-processing steps, also change vector orientation. While the $\ell_2$ function norm is preserved under projection to the SH basis (ie, asymptotically SH projection is an isomorphism for $\ell_2$), this is not the case for a norm on general finite sets of vectors.

To correct for this, we construct a projection matrix from the shared continuous SH basis to the diffusion gradient directions. This projection can then be used in conjunction with decoder output $p(x|z, s)$ to map output SH coefficients to the original subject-specific $b$-vector representation. We allow each $b_0$ channel to "pass through" the projection (mapped as identity), as they are without orientation. While we use the SH representation for both input and reconstruction (to leverage our invariance results), we augment the loss function from Equation 1 with a "real-space" loss function, the reconstruction loss in each subject's original domain. This encourages the overall loss function to be faithful to our use-case in the original image space.

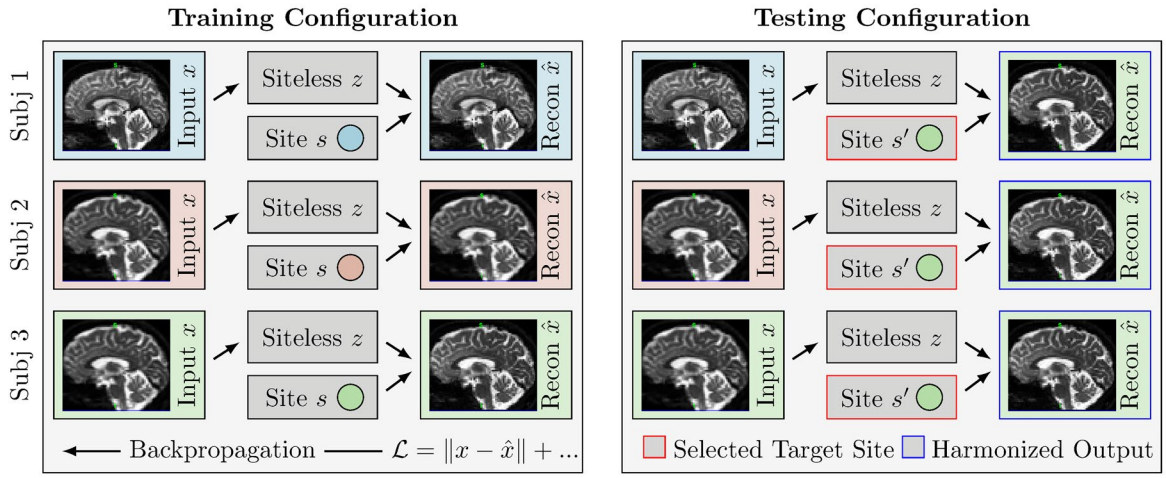
# 3 | METHODS

## 3.1 | Computational implementation

We parameterize $q$ and $p$ using neural networks, fitting their parameters by mini-batch gradient-based optimization. The loss in Equation 7 is defined generally, and invariant representations may be learned using many different function parameterizations. However, the flexibility of neural networks as function approximators make them ideal for this application. We apply these networks to small image patches, concatenating patch-wise outputs to create harmonized images. The overall architecture is shown in Figure 1A, and the training and testing configurations are diagrammed in Figure 1B, with exact parameters given in Section 3. We discuss the use of patches and its relative advantages and drawbacks in Section 5. As shown in




**(A)** Diagram of network configuration and losses




**(B)** Diagram of training and testing schema for the proposed method

**FIGURE 1** Diagrams describing network configuration and training/testing schema

Figure 1B, each sample consists of a single unpaired patch, and batches of data consist of patches and protocol identifiers (one-hot vectors). As diagrammed on the right-hand side of Figure 1B, protocol identifiers are manipulated at test time to produce harmonized reconstructions.

Our primary reconstruction loss is computed in the SH domain with respect to the entire patch. We then add a secondary loss function for the center voxel based on the SH-to-DWI projection, and an adversarial loss which attempts to predict which scanner/protocol each reconstructed patch is from (seen at the right of Figure 1A). We added this branch in order to provide additional information toward keeping remapped patches "reasonable" when remapping to new sites; this prediction can be performed without explicit pairing of patches. Our loss function is then, in abstract,

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \alpha\mathcal{L}_{\text{prior}} + \beta\mathcal{L}_{\text{proj}} - \gamma\mathcal{L}_{\text{adv}} - \lambda I(z,s) \qquad (8)$$

where $\mathcal{L}_{\text{recon}}$ is SH reconstruction loss (using MSE), $\mathcal{L}_{\text{proj}}$ is the DWI space loss, and $\mathcal{L}_{\text{adv}}$ is the adversarial loss on the SH reconstruction, with three hyper parameters controlling trade-offs between objectives. This loss function trivially extends from the single-site case (one target site to/from one base site) to a multi-site case, where $s$ is categorical.

We use a standard adversarial training scheme for defining and minimizing $\mathcal{L}_{\text{adv}}$ (see eg, Chapter 7.13 of Goodfellow, Bengio, and Courville[33]). The adversarial loss $\mathcal{L}_{\text{adv}}$ is the softmax cross-entropy loss of a secondary "adversary" network, shown in green in Figure 1A. We alternate between optimizing the primary network (minimizing Equation 8), and the adversary (minimizing $\mathcal{L}_{\text{adv}}$).

We optimize these networks by differentiating the loss functions (Equation 8 and $\mathcal{L}_{\text{adv}}$) with respect to the network weights (ie, backpropagation[34]) and then using the Adam optimizer,[35] which is a first order optimization method. Optimization is undertaken using mini-batches. To compute gradients of the divergences in Equation 7 efficiently, we use the re-parameterization trick of Kingma and Welling,[31] using both a diagonal Gaussian conditional $q(z|x)$ and a Gaussian prior $p(z)$. We also use the closed form bound for $KL[q(z|x)\|q(z)]$ from Moyer et al.[30]

## 3.2 | Data and pre-processing

To evaluate our method, we use the 15 subjects from the 2018 CDMRI Challenge Harmonization dataset.[21,36] These subjects were imaged on two different scanners: a 3 T GE

Excite-HD "Connectom" and a 3 T Siemens Prisma scanner. For each scanner, two separate protocols were collected, one of which matches between the scanners at a low resolution, and another which does not match at a high resolution. This results in four different "site" combinations, for which all subjects were scanned, resulting in forty different acquisitions (10 subjects, 2 scanners, 2 protocols each). We split this into 9 training subjects, 1 validation subject, and 5 held out-test subjects.

The low resolution matching protocol had an isotropic spatial resolution of 2.4 mm with 30 gradient directions (TE = 89 ms, TR = 7200 ms) at two shells $b = 1200, 3000\,\mathrm{s\,mm^{-2}}$, as well as a minimum of 4 $b_0$ acquisitions, at least one of these with reverse phase encoding. These volumes were then corrected for EPI distortions, subject motion, and eddy current distortions using FSL's TOPUP/eddy.[37,38] Subjects from the "Connectom" scanner were then registered to the "Prisma" scanner using a affine transformation, fit to a co-temporally acquired $\mathrm{T_1}$-weighted image volume (previously registered to each corresponding FA volume). The $b$-vectors were then appropriately rotated. In the case of the "Connectom" scanner, geometric distortions due to gradient non-linearities were corrected for using in-house software.[39,40] The high resolution protocols are identical in pre-processing to their low resolution counterparts, but have isotropic voxel sizes of 1.5 mm (TE = 80 ms, TR = 4500 ms) and 1.2 mm (TE = 68 ms, TR = 5400 ms) for "Prisma" and "Connectom" scanners respectively, each with 60 gradient directions per shell, same b-shell configurations ($b = 1200, 3000\,\mathrm{s\,mm^{-2}}$). We downsample the spatial resolution of the high resolution scans to 2.4 mm isotropic to test the multi-task method, but keep the angular resolution differences. To simplify notation, we refer to the four scanner/protocol combinations by their scanner make and number of gradient directions: Prisma 30, Prisma 60, Connectom 30, and Connectom 60.

All scans were masked for white matter tissue. This was done in order to focus our analysis on the tissue most commonly assessed using diffusion MRI (see eg, for a review[41]). We map each of these scans to an 8th-order SH representation for input into our method, but retain the original domain for training outputs. We use the minimal $\ell_2$ weighted solution in the case of under-determined projections, which corresponds with the SVD solution (using the pseudo-inverse). This is well-defined, unlike direct projection.

## 3.3 | Experimental protocol

The original CDMRI 2018 challenge[21] specified three supervised tasks, mapping between one base "site" (Prisma 30) and the three target "sites" (Prisma 60, Connectom 30, and Connectom 60). We modify this task, removing correspondence/pairing knowledge between sites (keeping this information for validation and testing), and including the inverse mapping task (target to base). This results in six tasks, two for each target site.

We train a "single-site" network for each of the six tasks, learning representations for Prisma 30 and a single target site, a multi-site variant across all six tasks. During training the method is not provided corresponding patches, and is *only* given individual patches. A single sample corresponds to one patch, not a pair of patches. Paired patches are only used to calculate error measures.

We measure the performance of each method on the hold-out set of subjects using the Root Mean Squared Error (RMSE) between each method's output and the ground truth target images in the original DWI basis (after pre-processing). For comparison we also include results from Mirzaalian et al,[19] which is the only other unsupervised method we are aware of in the literature.

We further assess the performance of each method by estimating the fiber orientation distributions (FODs) for each reconstruction using Multi-shell multi-tissue constrained spherical deconvolution (MSMT-CSD),[42] with response functions estimated using the method proposed in Dhollander et al.[43] For both of these steps we use the implementations from MRtrix3.[44] For each FOD we compute the maxima at each voxel and compare it to the closest maxima of the ground truth image to compute angular error.

In order to assess the fidelity of common local diffusion model summary measures before and after harmonization, we measure the Mean Average Percent Error (Mean APE) and the Coefficient of Variation (CV) between method-estimated and observed summary measures, reported in Table 1. We measured Mean APE and CV for Fractional Anisotropy, Mean Diffusivity, Mean Kurtosis,[45] and Return-to-Origin-Probability (RTOP).[46] This mirrors the analysis in Ning et al.[47]

In order to test the specific effects of our compressive regularizations, we conducted two ablation tests of our method, comparing it to the "regular networks" with parameters described in Section 3.4. We re-trained both single-site and multi-site methods with the invariance parameter $\lambda$ set to 0, but otherwise the same settings. We further trained two more networks for $\lambda$ and $\alpha$ set to 0. Effectively this ablates the added invariance-inducing compressive elements of the loss function. We then compared their performance by computing the voxelwise difference in RMSE in each heldout test subject.

For the proposed methods and corresponding ablated networks we assessed the amount to which we removed site information from of the learned representation $z$ by attempting predict $s$ from $z$. If there is no information in $z$ about $s$ then we would expect the optimal predictor to do no better than random. To this end we trained feed forward networks to predict

**TABLE 1** Here we report the mean absolute percent error (APE) and mean coefficient of variation (CV) per voxel for each of the methods for four common diffusion summary measures: Fractional anisotropy (FA), mean diffusivity (MD), mean kurtosis (MK),[45] and Return-to-Origin-Probability (RTOP)[46]

| | | | FA | | MD | | MK | | RTOP | |
|---|---|---|---|---|---|---|---|---|---|---|
| * | P30 | Method | APE | CV | APE | CV | APE | CV | APE | CV |
| Connectom 30 | to | Mirzaalian | 0.46 | 0.48 | 0.42 | 0.80 | 0.99 | 1.02 | 0.19 | 1.22 |
| | | Single-task | 0.25 | 0.26 | 0.12 | 0.17 | 3.37 | 0.30 | 0.11 | 0.28 |
| | | Multi-task | 0.28 | 0.30 | 0.12 | 0.16 | 3.15 | 0.28 | 0.11 | 0.16 |
| | from | Mirzaalian | 0.50 | 0.39 | 0.52 | 0.59 | 0.96 | 1.05 | 0.22 | 0.26 |
| | | Single-task | 0.29 | 0.24 | 0.22 | 0.18 | 3.72 | 0.30 | 0.13 | 0.18 |
| | | Multi-task | 0.30 | 0.27 | 0.21 | 0.17 | 3.90 | 0.31 | 0.12 | 0.17 |
| Prisma 60 | to | Mirzaalian | 0.52 | 0.55 | 0.60 | 0.67 | 0.99 | 1.05 | 0.29 | 0.41 |
| | | Single-task | 0.34 | 0.37 | 0.12 | 0.23 | 3.44 | 0.31 | 0.14 | 1.47 |
| | | Multi-task | 0.34 | 0.37 | 0.12 | 0.19 | 3.17 | 0.29 | 0.13 | 0.45 |
| | from | Mirzaalian | 0.64 | 0.45 | 0.48 | 0.44 | 0.96 | 0.98 | 0.22 | 0.28 |
| | | Single-task | 0.41 | 0.30 | 0.38 | 0.15 | 0.48 | 0.14 | 0.09 | 0.16 |
| | | Multi-task | 0.42 | 0.32 | 0.38 | 0.15 | 0.45 | 0.13 | 0.08 | 0.11 |
| Connectom 60 | to | Mirzaalian | 0.86 | 0.79 | 0.88 | 0.92 | 1.01 | 1.05 | 1.11 | 26.18 |
| | | Single-task | 0.35 | 0.36 | 0.26 | 0.81 | 3.22 | 0.35 | 0.16 | 0.50 |
| | | Multi-task | 0.35 | 0.36 | 0.14 | 0.23 | 3.31 | 0.36 | 0.14 | 0.37 |
| | from | Mirzaalian | 3.19 | 1.26 | 4.64 | 5.97 | 0.88 | 0.90 | 0.63 | 0.69 |
| | | Single-task | 1.86 | 0.40 | 2.93 | 0.31 | 4.44 | 0.27 | 0.10 | 0.17 |
| | | Multi-task | 1.77 | 0.38 | 2.84 | 0.29 | 4.56 | 0.27 | 0.10 | 0.32 |

*Notes*: The APE measure is the same as the error metric reported in Ning et al[47]; similar to Ning et al,[47] we report values as decimals (where 1.00 corresponds to 100%), and not actual percentages. It is well known that the APE measure is biased towards methods reporting smaller values[41,48] We therefore also report the Coefficient of Variation, computed by dividing the RMSE by the observed sample mean. This measure is also sometimes referred to as the Relative RMSE.

$s$ from $z$ ("post-hoc adversaries"). As shown in Moyer et al,[30] the cross-entropy error of these networks is a lower bound for the mutual information $I(s, z)$. The post-hoc adversaries had same configuration as the patch-adversaries (two 32-unit layers using tanh(·) activations and the softmax cross-entropy loss).

## 3.4 | Configuration and parameters

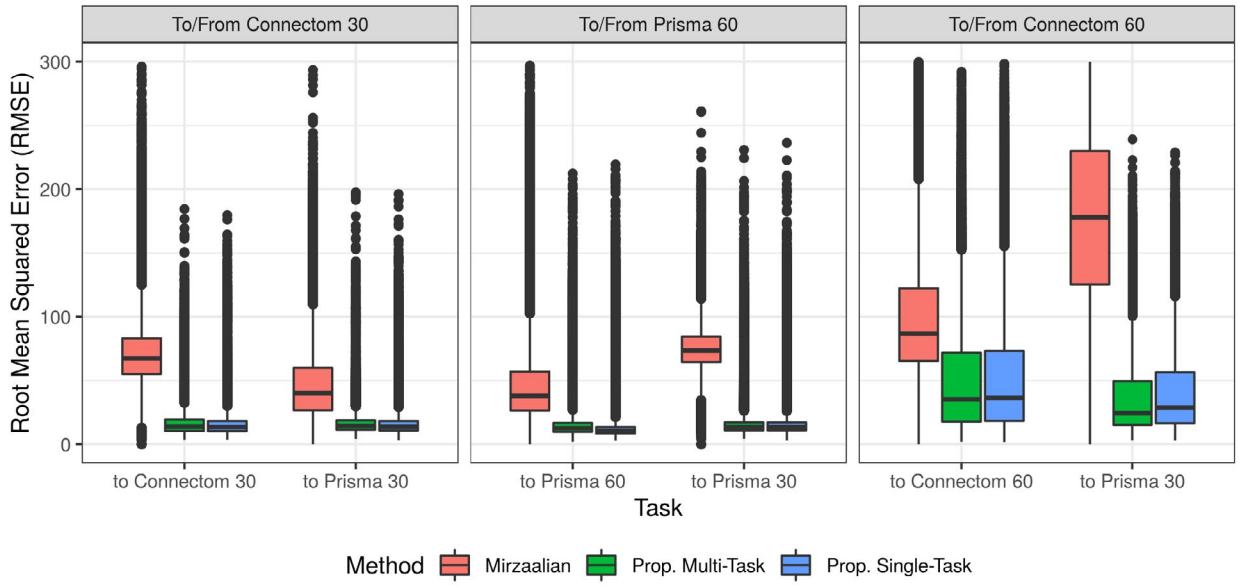
We implemented our method for image patches composed of a center voxel and each of its six immediate neighbors. Each of these voxels has two shells of DWI signal, which we mapped to the SH 8th order basis, plus one $b_0$ channel. Unravelling these patches and shells, the input is then a vector with $91 \times 7 = 637$ elements.

We use three-layer fully connected neural networks for encoder $q(z|x)$ and conditional decoder $p(x|z, s)$, with 256, 128, 64 hidden units respectively for the encoder, and the reverse (64, 128, then 256) for the decoder. The latent code $z$ is parameterized by a 32 unit Gaussian layer ($z$). This layer is then concatenated with the scanner/protocol one-hot representation $s$, and input into the decoder. We use tanh($x$) transformations at each hidden layer, with sigmoid output from the encoder for the variance of the Gaussian layer. The adversary is a fully connected two-layer network with 32 units at each layer, with tanh($x$) units again at each hidden node.

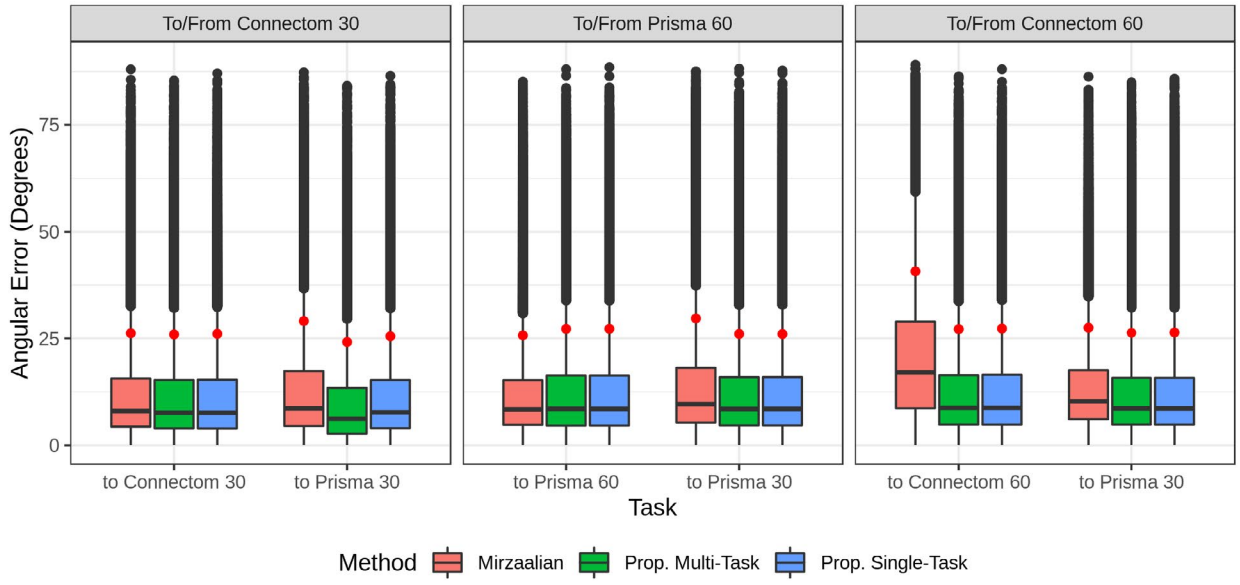
For each task we train our network for 1000 epochs, which took 19 hours to train in the pair-wise case on standard desktop equipped with an external Nvidia Titan-Xp with 12 GB of RAM using TensorFlow (32 GB of CPU RAM, 4 cores). We loosely tune the hyper parameters so losses are approximately on the same order of magnitude, with $\alpha = 1.0$, $\beta = 1.0$, $\gamma = 10.0$, and $\lambda = 0.01$. We use these same parameters for both the pair-wise tasks as well as the multi-task experiments. We use an Adam learning rate of 0.0001 and a batch size of 128. For each batch provided for primary network training we provide 10 epochs for training the adversary.

# 4 | RESULTS

Figure 2A plots the root mean squared error (RMSE) by voxel of the baseline, single-site proposed method, and multi-site proposed method, as evaluated on the holdout test subjects, in




**(A)** Voxel-wise RMSE of baseline and proposed methods. Lower is better




**(B)** Voxel-wise angular deflection of fiber orientation of baseline and proposed methods. Lower is better

**FIGURE 2** Here, we plot the voxel-wise performance as measured by RMSE (top) and angular deflection (bottom), that is, the angular difference between the global maxima of the Fiber Orientation Distributions recovered by MSMT-CSD[42] from the ground truth and reconstructed images, measured in degrees. This is shown for the Mirzaalian et al[19] method as well as our two proposed methods on each of the six harmonization tasks (Prisma 30 to and from each of the other scanner/site combinations). All RMSE values are calculated in the original signal representation. In both plots, lower is better. For the angular deflection error, the 90th percentile data point plotted in red

the original signal representation. Our proposed methods show improvement over the baseline method in each case. In the pair-wise task between similar protocols (mapping between Prisma 30 and Connectom 30), these improvements have non-overlapping inner quartile range. For dissimilar protocols, that is, mapping between Prisma 30 and Prisma 60 or Connectom 60, our proposed method shows improvements, though the difference is less pronounced. Surprisingly, for higher resolution target images the multi-site method performs as well or better than the pair-wise method and the baseline; this may be due to the multi-task method receiving many more volumes overall, allowing it to gather more information (albeit biased by other scanners) or preventing it from overfitting.

Figure 2B plots the voxel-wise angular deflection of each method, as measured by MSMT-CSD. For Connectom 30 and Prisma 60, both to and from Prisma 30, all three methods are comparable, with median errors well below 20°, and 90th percentile errors all slightly above 25°. For mappings to

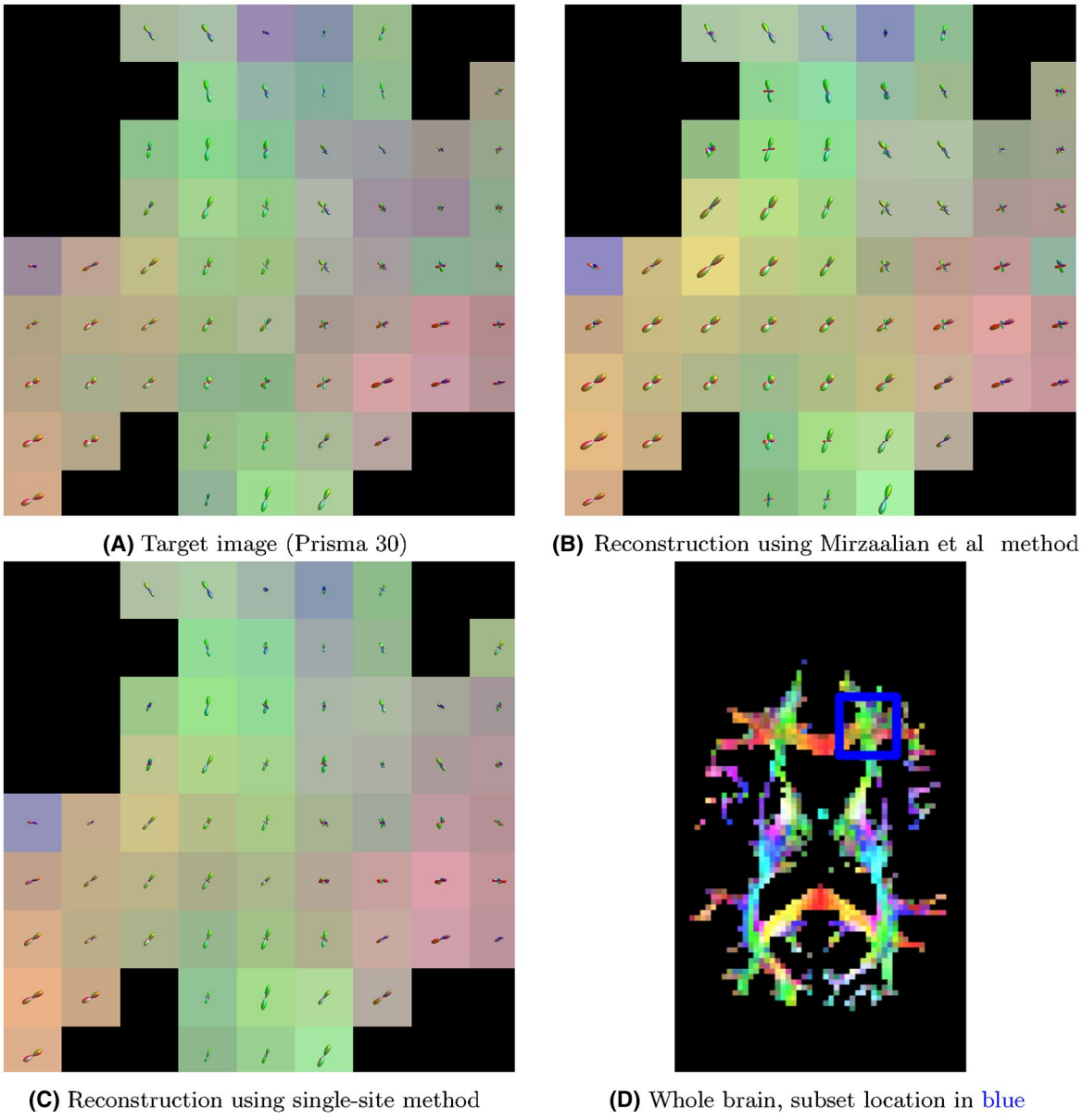



**FIGURE 3** Here, we plot exemplar FOD glyphs (estimated via MSMT-CSD) from the actual data, a reconstruction using the Mirzaalian et al method, and a reconstruction using the proposed single-site method. Inputs to each reconstruction were the data from the Prisma 60 protocol/site. The background colors represent the direction of the FOD maxima

the Connectom 60 protocol, the Mirzaalian et al method has generally higher error, though the inner quartile ranges still overlap for all methods. We plot a subset of the FODs from the original image and two of the reconstructions in Figure 3.

Figures 4-6 show the spatial distribution of the error for each tested method on a single test subject, for mappings between Prisma 30 and Connectom 30, Prisma 60, and Connectom 60 respectively. For the Prisma 30 to Connectom 30 mapping, overall the Mirzaalian baseline[19] has higher error than the other methods as shown by the overall coloring. The Mirzaalian baseline[19] and the multi-site proposed method show significant white matter patterning (though in varying degree); optimally we would like to see uncorrelated residuals, like those shown in the single-site method.

The Connectom 60 error plots (Figure 6) have a strong spatial patterns at both the occipital and frontal poles, shown in all methods. This wide-scale effect is somewhat mitigated by the proposed methods, but is still present in all error distributions.

Table 1 reports the Absolute Percent Error (APE) and the estimated Coefficient of Variation (CV) for each method voxel-wise for four commonly used diffusion summary measures: Fractional Anisotropy (FA), Mean Diffusivity (MD), Mean Kurtosis (MK),[45] and Return-to-Origin-Probability (RTOP).[46] It is well known that APE is biased towards methods reporting smaller values, and becomes inaccurate and inflated as actual observed values approach zero.[48,49] In our context this means that for FA and MD, more spherical tensors are weighted strongly, while more anisotropic tensors are

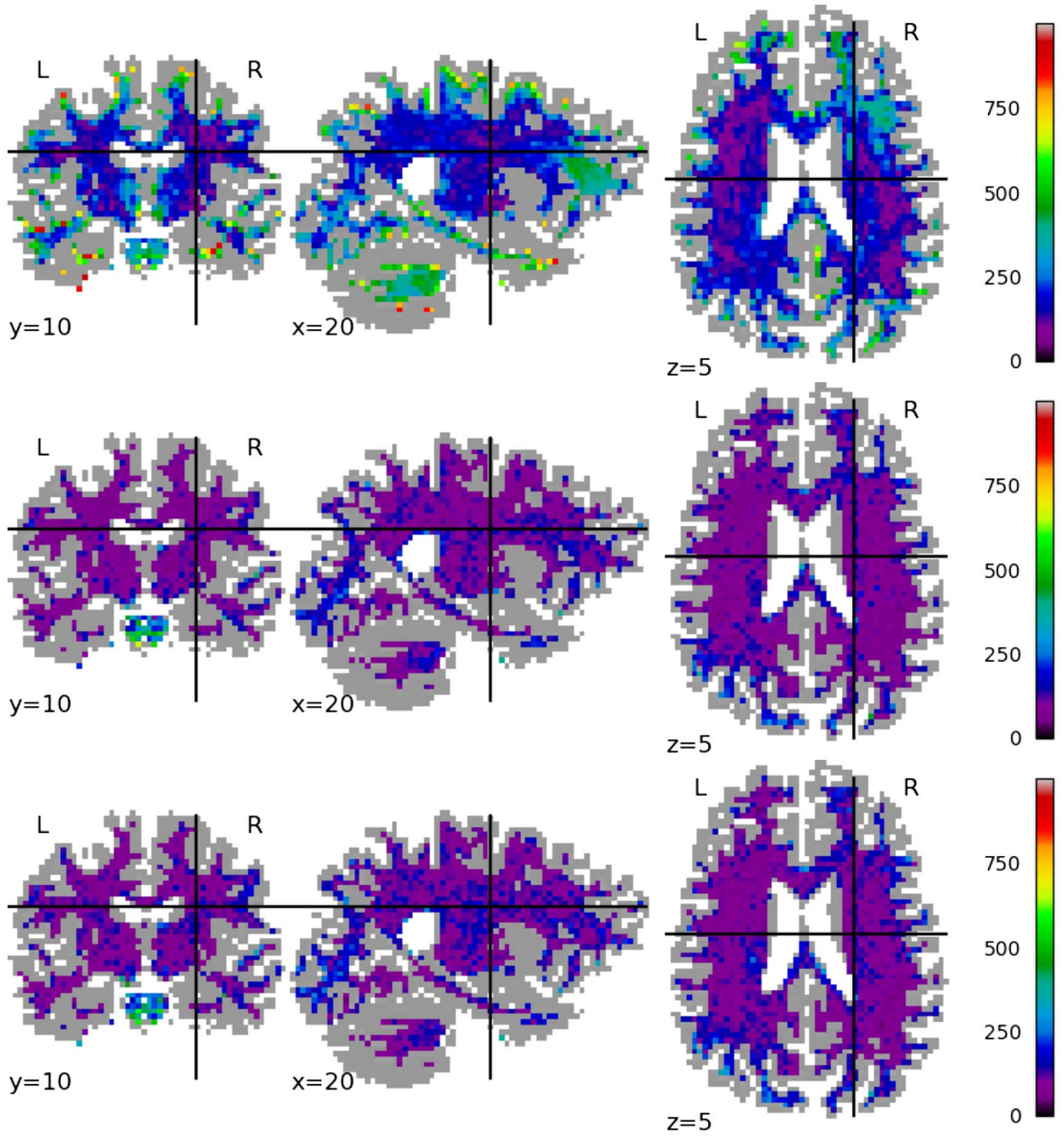



**FIGURE 4** We plot the spatial distribution of RMSE per voxel, displayed in slices centered at $(x, y, z) = (10, 22, 35)$ for mappings from the Prisma 30 protocol to the Connectom 30 protocol, for (top row) the Mirzaalian[19] baseline, (center row) the single-site proposed method, and (bottom row) the multi-site proposed method. The color scale is the same between the rows, as well as between this figure, Figures 5 and 6. All RMSE values are calculated in the original signal representation

weighted less. Due to this bias, we also report the estimated Coefficient of Variation (CV),[50] which is the RMSE divided by the observed sample mean (For unbiased estimates the RMSE divided by sample mean should further be multiplied by a factor of $\left(\sqrt{\frac{N}{N-1}}\right)\left(1-\frac{1}{4N}\right)$, where $N$ is the number of tested voxels. However, this number is very close to 1, and the resulting change is negligible). CV has also been used to assess summary statistic variation between scanners,[12,51] and is sometimes referred to as Relative RMSE.

For all reported summary measures except MK, the proposed methods map to and from Connectom 30 perform well under both error measures. Mapping to both Connectom 60 and Prisma 60 from Prisma 30 has higher error than the converse (Prisma 30 to Connectom/Prisma 60); this fits our intuitions about upsampling, as both "60" protocols have higher angular resolution.

For Mean Kurtosis in remapped scans to/from Connectom 30, the APE is very high while the CV is surprisingly low. This pattern is also seen in FA, MD, and MK for scans mapped to Connectom 60, and for MK in scans mapped from Prisma 60. Because the APE error is above 100% (but CV is small), we believe that the methods are overestimating small actual values, since underestimation error is bounded at 100% for non-negative measures. In order to further verify

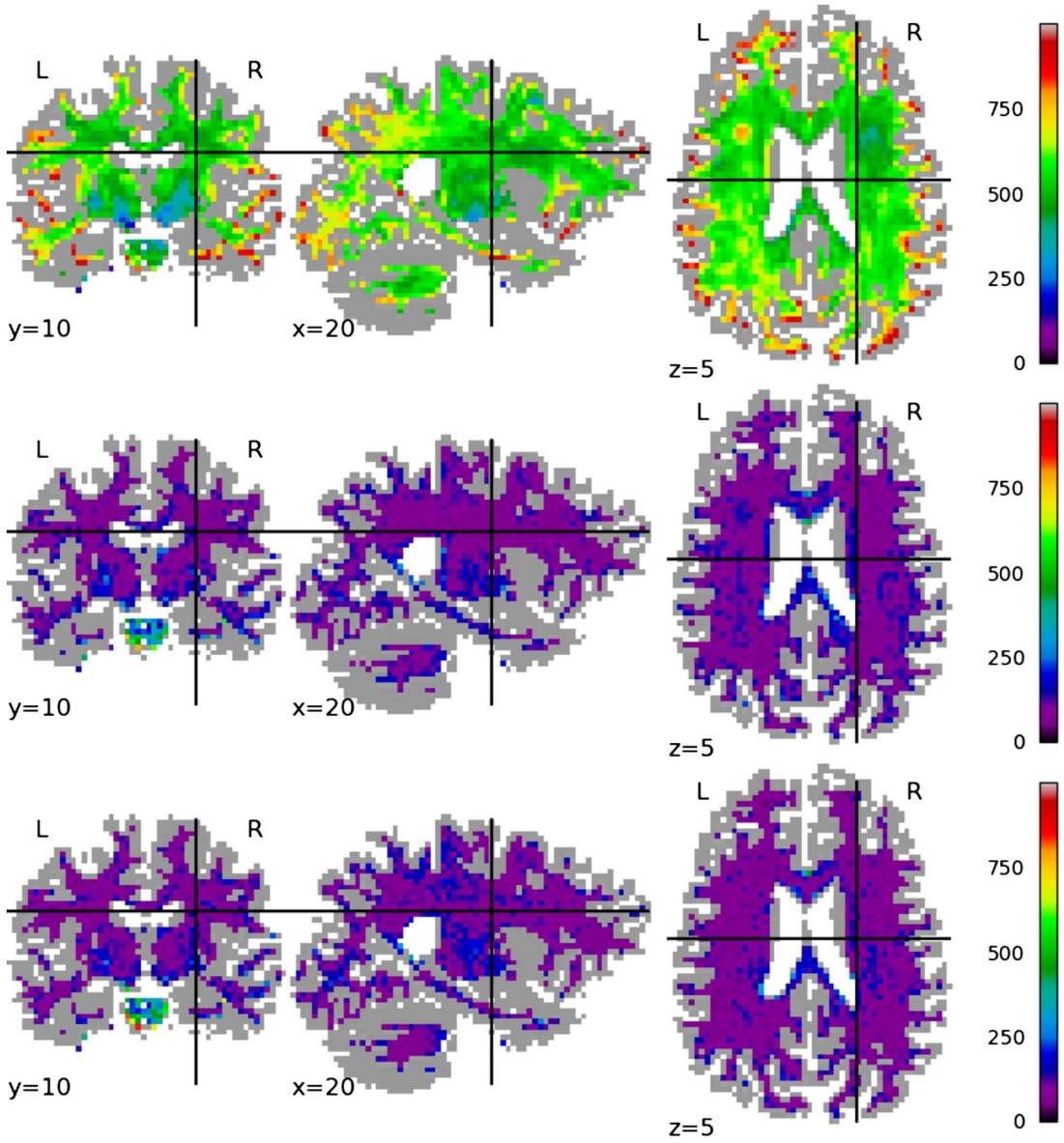



**FIGURE 5** We plot the spatial distribution of RMSE per voxel, displayed in slices centered at $(x, y, z) = (10, 22, 35)$ for mappings from the Prisma 30 protocol to the Prisma 60 protocol, for (top row) the Mirzaalian[19] baseline, (center row) the single-site proposed method, and (bottom row) the multi-site proposed method. The color scale is the same between the rows, as well as between this figure, Figures 5 and 6. All RMSE values are calculated in the original signal representation

this, we computed the Percent Error (without absolute values) shown in Table 2, indicating the average bias above or below the actual observed value. Since the MK PE for both proposed methods is very close to the APE, this indicates that on average small values are being overestimated. Discussion continues in Section 5.

## 4.1 | Ablation test results and post-hoc adversarial accuracies

Table 3 shows the results of the ablation tests, where we set either $\lambda$ to zero or both $\lambda$ and $\alpha$ to zero, effectively removing invariance and prior terms respectively from the primary loss function. We computed the difference in the RMSE between the regular model and the ablated models on the hold-out test dataset. For Connectom 30 both to and from Prisma 30, both the invariance term and the prior term hinder reconstruction performance. When only the invariance term is removed this effect is slight, but when both are removed the effect is much stronger in the multi-task setting.

For Prisma 60 mappings differences without the invariance term for the single task method are relatively small, while removing both the invariance and prior terms leads to large increases in RMSE. For mappings to Connectom 60, differences in RMSE follow a similar pattern to Prisma 60,

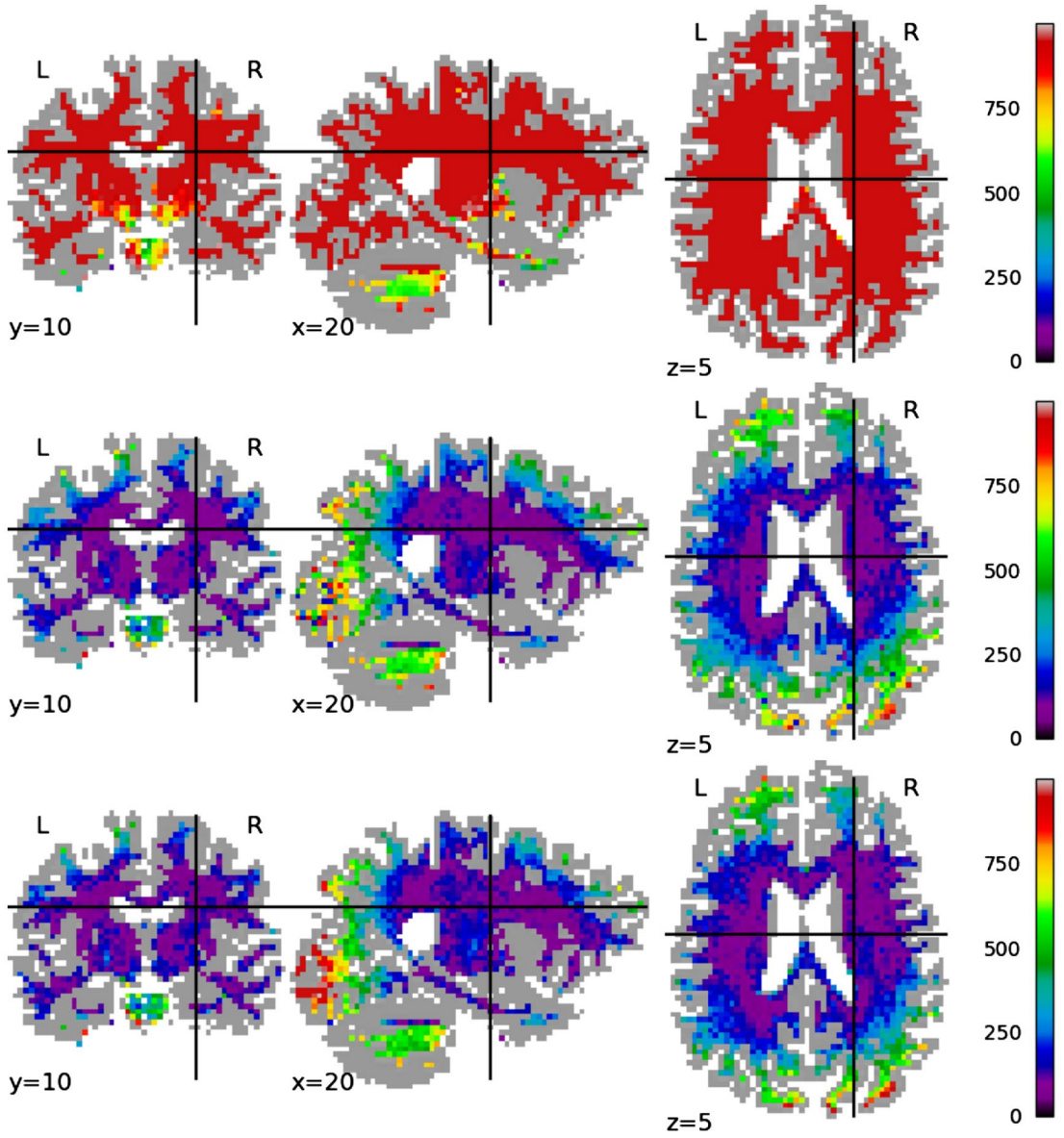



**FIGURE 6** We plot the spatial distribution of RMSE per voxel, displayed in slices centered at $(x, y, z) = (10, 22, 35)$ for mappings from the Prisma 30 protocol to the Connectom 60 protocol, for (top row) the Mirzaalian[19] baseline, (center row) the single-site proposed method, and (bottom row) the multi-site proposed method. The color scale is the same between the rows, as well as between this figure, Figures 4 and 5. All RMSE values are calculated in the original signal representation

with performance decreasing with both invariance and prior terms ablated. For the mappings from Connectom 60, performance strongly drops without the invariance term and then further drops without the prior term.

Table 4 shows the results of the post-hoc adversarial predictions. Setting $\lambda = 0$ uniformly increases post-hoc adversarial accuracy, and setting $\alpha = 0$ increases accuracy further in both Prisma 60 and Connectom 60 cases. For the multi-site model the prediction task is considerably harder, yet setting both $\lambda$ and $\alpha$ to zero induces relatively high adversarial accuracy in the multi-task setting (~60%).

It is unsurprising that the invariance term does not aid in reconstruction for more similar protocols/scanners. Inclusion of this term should lead to more compression, and thus less information in $z$ relevant to $x$, which in turn should lead to worse reconstruction. Further, this intuition also extends to the VAE prior term, which is a sufficient condition for the compressive portion of the invariance term (if $\mathcal{L}_{\text{prior}} = 0$ then $KL[q(z|x)\|q(z)] = 0$). It is interesting, however, that these terms lead to *increased* performance for dissimilar protocols/scanners, that is, Connectom 60 and Prisma 60. This indicates that these two loss terms are helpful for generalization.

**TABLE 2** Here, we report the mean Percent Error (PE) per voxel for each of the methods for four common diffusion summary measures

| * | P30 | Method | FA PE | MD PE | MK PE | RTOP PE |
|---|---|---|---|---|---|---|
| Connectom 30 | to | Mirzaalian | −0.27 | −0.26 | −0.87 | −0.09 |
| | | Single-task | −0.05 | 0.02 | 3.31 | −0.01 |
| | | Multi-task | −0.11 | 0.03 | 3.04 | −0.03 |
| | from | Mirzaalian | 0.22 | −0.45 | −0.95 | 0.20 |
| | | Single-task | 0.10 | 0.18 | 3.57 | −0.07 |
| | | Multi-task | 0.04 | 0.15 | 3.79 | −0.04 |
| Prisma 60 | to | Mirzaalian | −0.15 | −0.58 | −0.93 | −0.19 |
| | | Single-task | −0.15 | −0.00 | 3.38 | 0.00 |
| | | Multi-task | −0.14 | 0.05 | 3.03 | −0.05 |
| | from | Mirzaalian | 0.48 | −0.22 | −0.94 | 0.20 |
| | | Single-task | 0.18 | 0.29 | 0.44 | 0.03 |
| | | Multi-task | 0.15 | 0.32 | 0.33 | −0.01 |
| Connectom 60 | to | Mirzaalian | 0.23 | −0.88 | −0.91 | 0.36 |
| | | Single-task | 0.02 | 0.16 | 3.06 | −0.04 |
| | | Multi-task | −0.02 | 0.01 | 3.16 | −0.00 |
| | from | Mirzaalian | 2.32 | 3.65 | −0.73 | −0.62 |
| | | Single-task | 1.70 | 2.88 | 4.33 | −0.02 |
| | | Multi-task | 1.56 | 2.77 | 4.48 | 0.01 |
| | | Negative PE implies Actual Value > Estimated Value | | | | |

*Note*: Negative PE implies that the value from the real data was greater than the value from the harmonization method.

**TABLE 3** Here we report the mean per-voxel test set RMSE change between the regular model and two ablated models, where (top) the invariance term $\lambda$ was set to zero, and (bottom) the invariance term $\lambda$ and the VAE prior term $\alpha$ were set to zero

| | **$\Delta$RMSE for $\lambda = 0$** | | | | | |
|---|---|---|---|---|---|---|
| | **$\Delta\text{RMSE} = \text{RMSE}_{\text{reg.}} - \text{RMSE}_{\lambda=0}$** | | | | | |
| | **Connectom 30** | | **Prisma 60** | | **Connectom 60** | |
| | **To** | **From** | **To** | **From** | **To** | **From** |
| Proposed Single-task | 1.2 | 2.6 | −1.4 | 7.6 | −4.0 | −46.8 |
| Proposed Multi-task | 6.3 | 2.6 | 16.7 | 12.25 | −4.1 | −65.7 |
| | **$\Delta$RMSE for $\lambda = 0, \alpha = 0$** | | | | | |
| | $\Delta\text{RMSE} = \text{RMSE}_{\text{reg.}} - \text{RMSE}_{\lambda=0,\alpha=0}$ | | | | | |
| | **Connectom 30** | | **Prisma 60** | | **Connectom 60** | |
| | **To** | **From** | **To** | **From** | **To** | **From** |
| Proposed Single-task | 6.1 | 1.9 | −6.4 | −1.5 | −17.0 | −55.7 |
| Proposed Multi-task | 94.3 | 89.4 | −324.8 | −354.8 | −217.0 | −90.0 |

*Note*: Negative values indicate that the regular model has better performance than the ablated models.

**TABLE 4** Here, we report the mean per-patch test set classification accuracy for an adversary trained post hoc to predict the site variable $s$ from the latent representation $z$, for $z$ taken from the full model (center left column), the $\lambda$ ablated model (center right column), and the $\lambda$ and $\alpha$ ablated model (right column)

| | **Post-hoc Adversarial Accuracy, Predicting $s$ from $z$** | | | |
|---|---|---|---|---|
| | **Best Possible** | **Full Model** | **$\lambda = 0$** | **$\lambda = 0, \alpha = 0$** |
| Proposed Single-task, Connectom 30 | 0.5 | 0.61 | 0.63 | 0.63 |
| Proposed Single-task, Prisma 60 | 0.5 | 0.5 | 0.51 | 0.54 |
| Proposed Single-task, Connectom 60 | 0.5 | 0.63 | 0.68 | 0.85 |
| Proposed Multi-task | 0.25 | 0.41 | 0.41 | 0.62 |

*Notes*: The far left column shows the best possible performance. The architecture and training protocol for the adversaries is described in Section 3.3. Here, lower is better ("closer to invariant").

## 5 | DISCUSSION

Our proposed harmonization method is unsupervised in that we *do not* require multiple images from the same subject or phantom from separate sites (ie, paired data) in order to train our method. It is advisable to validate using such data, but due to the expense of collecting images from the same subject at varying sites it is advantageous to limit reliance on these data.

We believe it is important to understand the trade-off between reconstructive error and adversarial accuracy (eg, between performance in Figures 2A and 4). It is obviously desirable to have high reconstructive accuracy, yet any attempt to induce invariance necessarily removes information (ie, site information), which reduces this accuracy. At the other end of the spectrum, there is always a family of perfectly invariant solutions (constant images), but these also have no information about the subjects, and subsequently very high reconstructive error. It is thus important to consider both in selecting a remapping method.

Because of the VAE prior's sufficiency for compressing $z$, empirically we can create an acceptable method without the invariance term (ie, with $\lambda = 0$). This agrees with our intuition about Equation 1, where compression plus conditional reconstruction is a proxy for invariance. It appears that the exact form of compression is less impactful. However, best performance is achieved by including an invariance term.

It is tempting to attempt to interpret the encodings $z$, but these efforts should not be undertaken lightly. The encoding and decoding functions are designed to be non-linear, and individual components of $z$ may have interaction effects with other components. Further, the encodings $z$ are not images or patches, lacking a spatial domain. With careful construction analysis may be possible, but it is almost certainly non-trivial to do in the encoding domain.

In the current method we reconstruct images for a specific target site $s'$. We might instead look for a site agnostic image. This is philosophically challenging: images are by nature collected at sites, and there are no site-less images. While we can manipulate our method to produce an $s^*$ average site, the output image may not be representative of any of the images. It may be that all images must have site information, and that the quotient representation is not an image at all. On the other hand, for other tasks $y$ that are not images, for example, prediction of pathology or prognosis, we can use $z$ to make unbiased (scanner-agnostic) predictions of $y$. In cases where the actual goal is not in the image domain (for which the harmonization task is a pre-processing step), such a formulation may be beneficial, and could be built from our proposed method.

### 5.1 | Limitations

This method cannot remove long-range scanner-biases; this is due to the patch-based architecture. In theory, with larger patches, we could avoid this limitation; current hardware, in particular GPU memory and bus speeds, limit our computation to small patches for dMRI. Specific work in this domain has been done to reduce memory load,[17] but it is by no means solved, especially for high angular resolution data such as the HCP dataset.[52] We hypothesize that a similar architecture with larger patches or whole images could rectify this particular problem—architectures that may become accessible with increased hardware capabilities—or better model compression/computational reduction techniques.

In the present work, the proposed method was only evaluated on white matter, and not in grey matter (neither cortex nor subcortical structures). White matter analyses generally focus on models of restricted axonal compartments (fibers), with derived measures such as fiber orientation distribution functions (FODs) and voxel-wise data with generally high anisotropy. Grey matter analyses in contrast may focus more on signal from isotropic compartments and/or dendritic arbors,[53] and notably their models may be robust or vulnerable to site-bias in different ways. We have not considered grey matter signal or model summary statistics in this analysis, and thus we advise caution when applying this method to identified grey matter voxels. Further, as Tables 1 and 2 show, for low values of Mean Kurtosis the proposed method is inaccurate and has a positive bias in reconstruction. We advise caution when using this method where the accuracy of these measures for low relative values is critical.

## 6 | CONCLUSION

In the present work we have constructed a method for learning scanner-invariant representations. These representations can then be used to reconstruct images under a variety of different scanner conditions, and due to the data processing inequality the reconstruction's mutual information with the original scanner will be low. This we demonstrate to be useful for the unsupervised case of data harmonization in diffusion MRI; critically, we can harmonize data without explicit pairing between images, reducing the need for. Surprisingly in some cases the multi-task method outperforms a pairwise method with similar architecture. This may hint at further benefits for learning shared representations.

### ACKNOWLEDGMENTS

This work was supported by NIH (U.S. National Institutes of Health) grants P41 EB015922, R01 MH116147, R56 AG058854, RF1 AG041915, U01AG024904, and U54 EB020403, DARPA grant W911NF-16-1-0575, as well as the NSF Graduate Research Fellowship Program Grant Number DGE-1418060, and a GPU grant from NVidia. 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. Prior consent was obtained from all patients before each scanning session, along with the approval of the Cardiff University School of Psychology ethics committee. 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). We acknowledge the 2017 and 2018 MICCAI Computational Diffusion MRI committees (Francesco Grussu, Enrico Kaden, Lipeng Ning, Jelle Veraart, Elisenda Bonet-Carne, and Farshid Sepehrband) and CUBRIC, Cardiff University (Derek Jones, Umesh Rudrapatna, John Evans, Greg Parker, Slawomir Kusmia, Cyril Charron, and David Linden).

## ORCID

*Daniel Moyer* http://orcid.org/0000-0003-4428-5012

## TWITTER

*Daniel Moyer* @PTenigma

## REFERENCES

1. Chen J, Liu J, Calhoun VD, et al. Exploration of scanning effects in multi-site structural MRI studies. *J Neurosci Methods*. 2014;230:37–50.
2. Fortin JP, Parker D, Tunc B, et al. Harmonization of multi-site diffusion tensor imaging data. *Neuroimage*. 2017;161:149–170.
3. Jovicich J, Czanner S, Greve D, et al. Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. *Neuroimage*. 2006;30:436–443.
4. Hawco C, Viviano JD, Chavez S, et al. A longitudinal human phantom reliability study of multi-center $T_1$-weighted, DTI, and resting state fMRI data. *Psychiatry Res-Neuroim*. 2018;282:134–142.
5. Kelly S, Jahanshad N, Zalesky A, et al. Widespread white matter microstructural differences in schizophrenia across 4322 individuals: results from the ENIGMA schizophrenia DTI working group. *Molecular Psychiatry*. 2018;23.5:1261–1269.
6. Magnotta VA, Matsui JT, Liu D, et al. Multicenter reliability of diffusion tensor imaging. *Brain Connectivity*. 2012;2:345–355.
7. Zavaliangos-Petropulu A, Nir TM, Thomopoulos SI, et al. Diffusion MRI indices and their relation to cognitive impairment in brain aging: the updated multi-protocol approach in ADNI3. *Front Neuroinform*. 2019;13:2.
8. Correia MM, Carpenter TA, Williams GB. Looking for the optimal DTI acquisition scheme given a maximum scan time: are more b-values a waste of time? *Magn Reson Imaging*. 2009;27:163–175.
9. Giannelli M, Cosottini M, Michelassi MC, et al. Dependence of brain DTI maps of fractional anisotropy and mean diffusivity on the number of diffusion weighting directions. *J Appl Clin Med Phys*. 2010;11:176–190.
10. Pagani E, Hirsch JG, Pouwels PJ, et al. Intercenter differences in diffusion tensor MRI acquisition. *J Magn Reson Imaging*. 2010;31:1458–1468.
11. Papinutto ND, Maule F, Jovicich J. Reproducibility and biases in high field brain diffusion MRI: an evaluation of acquisition and analysis variables. *Magn Reson Imaging*. 2013;31:827–839.
12. Vollmar C, O'Muircheartaigh J, Barker GJ, et al. Identical, but not the same: intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0 T scanners. *Neuroimage*. 2010;51:1384–1394.
13. White T, Magnotta VA, Bockholt HJ, et al. Global white matter abnormalities in schizophrenia: a multisite diffusion tensor imaging study. *Schizophrenia Bull*. 2009;37:222–232.
14. Zhan L, Leow AD, Jahanshad N, et al. How does angular resolution affect diffusion imaging measures? *Neuroimage*. 2010;49:1357–1371.
15. Zhan L, Mueller BA, Jahanshad N, et al. Magnetic resonance field strength effects on diffusion measures and brain connectivity networks. *Brain Connectivity*. 2013;3:72–86.
16. Zhan L, Franc D, Patel V, et al. How do spatial and angular resolution affect brain connectivity maps from diffusion MRI? In: *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*. Barcelona, Spain: IEEE; 2012:1–4.
17. Blumberg SB, Tanno R, Kokkinos I, Alexander DC. Deeper image quality transfer: training low-memory neural networks for 3D images. In: *MICCAI*. Granada, Spain: Springer; 2018118–125.
18. Tanno R, Worrall DE, Ghosh A, et al. Bayesian image quality transfer with CNNs: exploring uncertainty in dMRI super-resolution. In: *MICCAI*. Quebec City, Canada: Springer; 2017;611–619.
19. Mirzaalian H, Ning L, Savadjiev P. Multi-site harmonization of diffusion MRI data in a registration framework. *Brain Imaging Behavior*. 2018;12:284–295.
20. Ning L, Bonet-Carne E, Grussu F, et al. Multi-shell diffusion MRI harmonisation and enhancement challenge (MUSHAC): progress and results. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Granada, Spain: Springer; 2018;217–224.
21. Tax CM, Grussu F, Kaden E, et al. Cross-vendor and cross-protocol harmonisation of diffusion MRI data: a comparative study. In: Proceedings of the International Society for Magnetic Resonance in Medicine. Paris. 2018;471.
22. Zhu AH, Moyer DC, Nir TM, Thompson PM, Jahanshad N. Challenges and opportunities in dMRI data harmonization. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Granada, Spain: Springer; 2018:157–172.
23. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8:118–127.
24. Mirzaalian H, de Pierrefeu A, Savadjiev P, et al. Harmonizing diffusion MRI data across multiple sites and scanners. In: *MICCAI*. Munich, Germany: Springer; 2015;12–19.
25. Mirzaalian H, Ning L, Savadjiev P. Inter-site and inter-scanner diffusion MRI data harmonization. *Neuroimage*. 2016;135:311–323.
26. Karayumak SC, Bouix S, Ning L, et al. Retrospective harmonization of multi-site diffusion MRI data acquired with different acquisition parameters. *Neuroimage*. 2019;184:180–200.
27. Karayumak SC, Kubicki M, Rathi Y. Harmonizing diffusion MRI data across magnetic field strengths. In: I*nternational Conference on Medical Image Computing and Computer-Assisted Intervention*. Granada, Spain: Springer; 2018;116–124.
28. Koppers S, Bloy L, Berman JI, Tax CM, Edgar JC, Merhof D. Spherical harmonic residual network for diffusion signal harmonization. 2018; arXiv preprint arXiv:1808.01595
29. Cover TM, Thomas JA. *Elements of Information Theory*. Hoboken, NJ: John Wiley & Sons; 2012.

30. Moyer D, Gao S, Brekelmans R, Galstyan A, Ver Steeg G. Invariant representations without adversarial training. In: *Advances in Neural Information Processing Systems*. Montreal, Canada: Curran Associates, Inc.; 2018;31:9102–9111.
31. Kingma DP, Welling M. Auto-encoding variational Bayes. 2013; arXiv preprint arXiv:1312.6114.
32. Higgins I, Matthey L, Pal A, et al. beta-VAE: learning basic visual concepts with a constrained variational framework. *ICLR*. 2017;2:6.
33. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge, MA: MIT Press; 2016.
34. Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation. Technical Report, California Univ San Diego La Jolla Inst for Cognitive Science; 1985.
35. Kingma, DP, Ba J. Adam: a method for stochastic optimization. 2014; arXiv preprint arXiv:1412.6980.
36. Tax CM, Grussu F, Kaden E, et al. Cross-scanner and cross-protocol diffusion mri data harmonisation: a benchmark database and evaluation of algorithms. *Neuroimage*. 2019;195:285–299.
37. Andersson JL, Skare S, Ashburner J. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *Neuroimage*. 2003;20:870–888.
38. Andersson JL, Sotiropoulos SN. An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *Neuroimage*. 2016;125:1063–1078.
39. Glasser MF, Sotiropoulos SN, Wilson JA, et al. The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage*. 2013;80:105–124.
40. Rudrapatna S, Parker G, Roberts J, Jones D. Can we correct for interactions between subject motion and gradient-nonlinearity in diffusion MRI. In: Proceedings of the International Society for Magnetic Resonance in Medicine, Paris, France; 2018;1206.
41. Assaf Y, Pasternak O. Diffusion tensor imaging (DTI)-based white matter mapping in brain research: a review. *J Mol Neurosci*. 2008;34:51–61.
42. Jeurissen B, Tournier JD, Dhollander T, Connelly A, Sijbers J. Multi-tissue constrained spherical deconvolution for improved analysis of multi-shell diffusion MRI data. *NeuroImage*. 2014;103:411–426.
43. Dhollander T, Raffelt D, Connelly A. Unsupervised 3-tissue response function estimation from single-shell or multi-shell diffusion MR data without a co-registered $T_1$ image. In: ISMRM Workshop on Breaking the Barriers of Diffusion MRI, Singapore; 2016:5.
44. Tournier JD, Smith R, Raffelt D, et al. Mrtrix3: a fast, flexible and open software framework for medical image processing and visualisation. *NeuroImage*. 2019:116137.
45. Jensen JH, Helpern JA, Ramani A, Lu H, Kaczynski K. Diffusional kurtosis imaging: the quantification of non-Gaussian water diffusion by means of magnetic resonance imaging. *Magn Reson Med*. 2005;53:1432–1440.
46. Özarslan E, Koay CG, Shepherd TM, et al. Mean apparent propagator (MAP) MRI: a novel diffusion imaging method for mapping tissue microstructure. *NeuroImage*. 2013;78:16–32.
47. Ning L, Bonet-Carne E, Grussu F, et al. Cross-scanner and cross-protocol harmonisation of multi-shell diffusion MRI data: open challenge and evaluation results. In: Proceedings of the International Society for Magnetic Resonance in Medicine, Montreal, Canada; 2019.
48. Armstrong JS, Collopy F. Error measures for generalizing about forecasting methods: empirical comparisons. *Int J Forecasting*. 1992;8:69–80.
49. Makridakis S. Accuracy measures: theoretical and practical concerns. *Int J Forecasting*. 1993;9:527–529.
50. Abdi H. Coefficient of variation. *Encyclopedia Res Design*. 2010;1:169–171.
51. Cercignani M, Bammer R, Sormani MP, Fazekas F, Filippi M. Inter-sequence and inter-imaging unit variability of diffusion tensor MR imaging histogram-derived metrics of the brain in healthy volunteers. *Am J Neuroradiol*. 2003;24:638–643.
52. Sotiropoulos SN, Jbabdi S, Xu J, et al. Advances in diffusion MRI acquisition and processing in the human connectome project. *Neuroimage*. 2013;80:125–143.
53. Jespersen SN, Kroenke CD, Østergaard L, Ackerman JJ, Yablonskiy DA. Modeling dendrite density from magnetic resonance diffusion measurements. *Neuroimage*. 2007;34:1473–1486.

**How to cite this article:** Moyer D, Ver Steeg G, Tax CMW, Thompson PM. Scanner invariant representations for diffusion MRI harmonization. *Magn Reson Med*. 2020;84:2174–2189. https://doi.org/10.1002/mrm.28243

## APPENDIX A

### Derivation of the Bound in Equation 1

This bound is also found in Moyer et al,[30] where it is used in the context of Fair Representations. Again, we reproduce it here for clarity, but the demonstration remain unchanged. All entropic quantities are with respect to $q$ the empirical encoding distribution unless otherwise stated.

From the tri-variate identities of mutual information, we have that $I(z, s) = I(z, x) - I(z, x|s) + I(z, s|x)$. However, the distribution of $z$ is exactly given by $\int q(z|x)dx$ by construction, and thus the distribution of $z$ solely depends on $x$. Thus,

$$I(z,s|x) = H(z|x) - H(z|x,s) = H(z|x) - H(z|x) = 0. \tag{A1}$$

we can then write the following:

$$I(z,s) = I(z,x) - I(z,x|s) \tag{A2}$$

$$= I(z,x) - H(x|s) + H(x|z,s) \tag{A3}$$

$$\leq I(z,x) - H(x|s) - \mathbb{E}_{x,s,z\sim q}[\log p(x|z,s)] \tag{A4}$$

$$= \mathbb{E}_{z,x}[\log q(z|x) - \log q(z)] - H(x|s) - \mathbb{E}_{x,s,z\sim q}[\log p(x|z,s)] \tag{A5}$$

$$= \mathbb{E}_{x}[KL[q(z|x)\|q(z)]] - H(x|s) - \mathbb{E}_{x,s,z\sim q}[\log p(x|z,s)]. \tag{A6}$$

This inequality is tight if and only if the variational approximation $p(x|z, s)$ is correct; interpreted in an imaging context, if we cannot perform conditional reconstruction correctly this bound will not be tight.