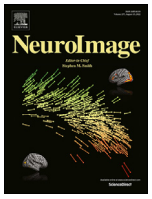
NeuroImage 257 (2022) 119297

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage



A deep learning-based multisite neuroimage harmonization framework established with a traveling-subject dataset

Dezheng Tian[a,b,c], Zilong Zeng[a,b,c], Xiaoyi Sun[a,b,c,d], Qiqi Tong[e], Huanjie Li[f], Hongjian He[g], Jia-Hong Gao[h,i,j], Yong He[a,b,c,k], Mingrui Xia[a,b,c,*]

[a] State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing 100875, China
[b] Beijing Key Laboratory of Brain Imaging and Connectomics, Beijing Normal University, Beijing 100875, China
[c] IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing 100875, China
[d] School of Systems Science, Beijing Normal University, Beijing 100875, China
[e] Research Center for Healthcare Data Science, Zhejiang Lab, Hangzhou 311121, China
[f] School of Biomedical Engineering, Dalian University of Technology, Dalian 116024, China
[g] Center for Brain Imaging Science and Technology, Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou 310027, China
[h] Center for MRI Research, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China
[i] Beijing City Key Laboratory for Medical Physics and Engineering, Institute of Heavy Ion Physics, School of Physics, Peking University, Beijing 100871, China
[j] IDG/McGovern Institute for Brain Research, Peking University, Beijing 100871, China
[k] Chinese Institute for Brain Research, Beijing 102206, China

ARTICLE INFO

Keywords:
Big data
Machine learning
Multicenter
Gray matter
Convolutional network
Site effect

ABSTRACT

The accumulation of multisite large-sample MRI datasets collected during large brain research projects in the last decade has provided critical resources for understanding the neurobiological mechanisms underlying cognitive functions and brain disorders. However, the significant site effects observed in imaging data and their derived structural and functional features have prevented the derivation of consistent findings across multiple studies. The development of harmonization methods that can effectively eliminate complex site effects while maintaining biological characteristics in neuroimaging data has become a vital and urgent requirement for multisite imaging studies. Here, we propose a deep learning-based framework to harmonize imaging data obtained from pairs of sites, in which site factors and brain features can be disentangled and encoded. We trained the proposed framework with a publicly available traveling subject dataset from the Strategic Research Program for Brain Sciences (SRPBS) and harmonized the gray matter volume maps derived from eight source sites to a target site. The proposed framework significantly eliminated intersite differences in gray matter volumes. The embedded encoders successfully captured both the abstract textures of site factors and the concrete brain features. Moreover, the proposed framework exhibited outstanding performance relative to conventional statistical harmonization methods in terms of site effect removal, data distribution homogenization, and intrasubject similarity improvement. Finally, the proposed harmonization network provided fixable expandability, through which new sites could be linked to the target site via indirect schema without retraining the whole model. Together, the proposed method offers a powerful and interpretable deep learning-based harmonization framework for multisite neuroimaging data that can enhance reliability and reproducibility in multisite studies regarding brain development and brain disorders.

1. Introduction

Advances in magnetic resonance imaging (MRI) in recent decades have provided powerful techniques for noninvasively exploring the structures and functions of the human brain in vivo, facilitating our understanding of neurobiological mechanisms underlying the development of complex cognitions and clinical impairments related to brain disorders (Cao et al., 2017a; Fornito et al., 2015; Park and Friston, 2013). Multisite MRI data acquisition in recently launched large brain research projects, such as the IMAGEN (Schumann et al., 2010) and ABCD (Casey et al., 2018) projects, has accumulated critical neuroimaging resources to facilitate brain investigations with impres-

* **Corresponding author at:** State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Key Laboratory of Brain Imaging and Connectomics, IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing 100875, China.
E-mail address: mxia@bnu.edu.cn (M. Xia).

https://doi.org/10.1016/j.neuroimage.2022.119297.
Received 18 December 2021; Received in revised form 31 March 2022; Accepted 9 May 2022
Available online 12 May 2022.

sive statistical power (Laird, 2021; Poldrack and Gorgolewski, 2014; Xia and He, 2017). However, considerable heterogeneity among imaging datasets collected from different sites, which results in differences known as site effects, has been widely documented in both raw structural and functional imaging data (Li et al., 2020; Radua et al., 2020) and image-derived brain characteristics, such as gray matter volume (GMV) (Melzer et al., 2020) and functional connectivity (Noble et al., 2017a; Yamashita et al., 2019). The site effect predominantly results from both the sampling of divergent populations and the different scan equipment across different sites and is a major source of the inconsistencies in the findings reported from different studies on the same topic. Therefore, developing methods for harmonizing imaging data across different scan sites has become a fundamental and urgent requirement for multisite imaging studies.

To correct for the site effect in multisite imaging data, several harmonization strategies have been proposed, which can be summarized into two major categories: conventional statistics-based harmonization methods and recently developed deep learning (DL)-based harmonization methods. Conventional statistical methods are usually applied via linear regression of univariate metrics, with sites indexed as a categorical covariate, for example, in the least squares-based general linear model (Rao et al., 2017) and Bayesian estimation-based ComBat (Fortin et al., 2018; Fortin et al., 2017). These methods have been utilized in multisite imaging studies and have shown a powerful capacity for removing linear site effects in brain metrics (Pomponio et al., 2020; Xia et al., 2019; Yu et al., 2018). However, noticeable limitations have been observed for this type of harmonization method. First, the site effect is mathematically assumed to be linear, while the actual effect may be much more complex. Second, brain characteristics are independently considered in these models, largely neglecting the spatial and topological relationships among brain regions. To overcome these defects, recently proposed DL-based harmonization methods, including the U-net autoencoder (Dewey et al., 2019) and cycle-generative adversarial networks (Chen et al., 2022c), allow for mapping the complex abstract representations of the nonlinear spatial patterns of site effects in MRI data. These models have been primarily applied to the harmonization of diffusion tensor images (Moyer et al., 2020; Tong et al., 2020), structural images (Zuo et al., 2021), and morphological measurements (Zhao et al., 2019), successfully eliminating site effects in such data containing complex spatial or topological information. However, the interpretability is relatively low for most of these established DL-based harmonization methods, for which high-dimensional representations are difficult to delineate. Additionally, the model training strategy of site pairing is a common approach for DL-based methods, and the fusion of data from multiple sites in a single model will greatly increase the model's complexity and require much more training data. Designing a harmonization framework with high expandability will facilitate the application of DL-based methods.

Another critical factor for establishing reliable multisite image harmonization models is the selection of training data. The core objectives of multisite harmonization are to eliminate nonbiological factors, such as MRI equipment and scan protocols, while simultaneously retaining the biological factors of participants across different sites. Therefore, the innovative traveling subject dataset, in which each participant is scanned at all sites, has become a valuable resource for the training of harmonization models, as it can minimize bias in population sampling across sites and ensure that established models only learn nonbiological factors (Noble et al., 2017b; Tong et al., 2019; Yamashita et al., 2019). Although existing multisite imaging studies have shown that harmonization models based on nontraveling subject datasets, be they conventional statistics- or DL-based models, can efficiently remove site effects (Garcia-Dias et al., 2020), whether intersite differences in biological factors are overeliminated is unknown. Benefiting from the publicly available traveling subject dataset, several recent studies have established harmonization methods that can separate and protect biological factors from complex site effects and have achieved outstanding performance with a small training sample (Yamashita et al., 2019). However, DL-based harmonization models for brain measurements using the traveling subject dataset are still lacking.

Here, we propose a DL-based harmonization framework that can disentangle both site factor and brain factor representations from site effects based on a publicly available traveling subject dataset. Taking the widely used GMV measurement as an illustration, we first examined whether this framework can significantly eliminate site effects in the GMV maps of nine scan sites. Then, we investigated whether the site factor and brain factor encoders embedded in the framework can capture intersite and intersubject variability, respectively. We further compared the proposed method with several conventional statistical harmonization methods in terms of site effect removal, data distribution homogenization, and intrasubject similarity improvement. Finally, we evaluated the expandability of the proposed framework for adding new sites to the target site via an indirect schema.

2. Methods

2.1. The deep learning-based representation disentanglement (DeRed) framework for multisite imaging data harmonization

We proposed a DL-based bidirectional framework (Fig. 1a) for neuroimaging data harmonization, which enables the transfer of imaging data from a given site to a target site and vice versa. Specifically, this framework contains four encoders for disentangling site factors and brain factors in imaging data from the source and target sites and two decoders for synthesizing harmonized images from the encoders. This design allows harmonized imaging data to contain both target site information and natural brain features. This framework was inspired by a disentangled unsupervised cycle-consistent adversarial network (DUNCAN) (Liu et al., 2021), which was developed to remove MRI artifacts based on representation disentanglement. As shown in Fig. 2a, the site factor encoder in DeRed contains three residual blocks, which can avoid the convergence performance degradation caused by structural redundancy (He et al., 2016a, b). Each residual block includes a set of two-dimensional (2D) convolution layers and leaky rectified linear unit (LeakyReLU) activation (Fig. 2b). When the feature maps pass through the residual block, the size is reduced by half, and the output of each residual block can be used as image features at different scales. Notably, each input slice of the site factor encoder must undergo an average pooling process before the residual blocks because the representation related to the scanning site or equipment should be abstract, regardless of anatomical details, and should not be extracted from the shallower layer. For the brain factor encoder, we replaced the average pooling layer with an additional residual block and added an instance normalization operation (Huang and Belongie, 2017) for each residual block to enable the capture of more detailed anatomical features and to improve the convergence speed.

The decoder (Fig. 2c) contains a two-step synthesis structure, integrating features extracted by the encoders. First, site factor features at different scales are mixed through a series of upsampling processes and residual blocks, but it should be noted that the size of each feature map is not reduced by half when passing through the residual block. Similarly, the mixing process for brain factor features also involves brain factor residual blocks and an upsampling process. After the first stage of the mixing process, the decoder produces two feature maps—one for the site factors and one for the brain factors—and the corresponding mean and maximum feature maps are also calculated. Second, these feature maps are concatenated and input into a brain factor residual block with a 2D

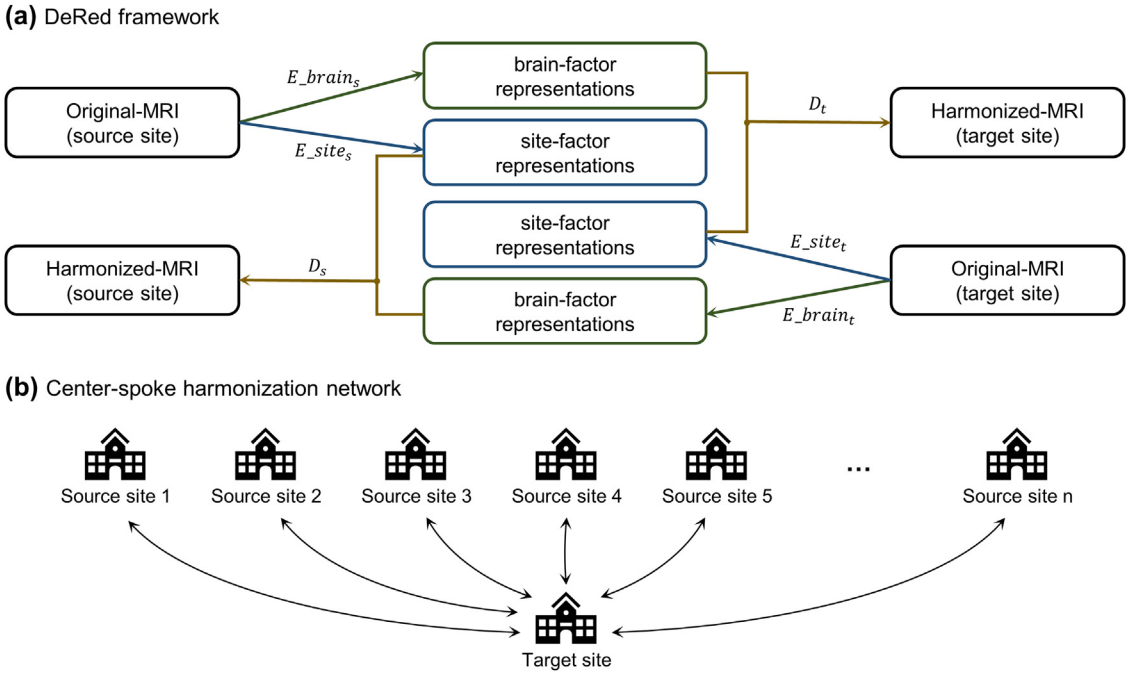


Fig. 1. Architecture of the DeRed framework and center-spoke harmonization network. (a) DL-based representation disentanglement framework. The site factor and brain factor features are extracted from the original MR images by the encoders, and the decoder synthesizes harmonized MR images by combining these two features. E_brain_s and E_site_s represent the brain factor and site factor encoders, respectively, for the source site. E_brain_t and E_site_t represent the brain factor and site factor encoders, respectively, for the target site. D_s and D_t represent the decoders for the source site and target site, respectively. **(b)** Center-spoke harmonization network with the target site located at the center. This harmonization network supports the bidirectional migration of MRI between the target site and source sites. Each two-way arrow represents an independent DeRed framework.

convolution operation. The data input into the site factor and brain factor encoders are 2D images obtained by slicing three-dimensional (3D) data along a certain direction, and the output results of the decoder maintain a shape consistent with that of the input data.

Based on the DeRed framework, we established a flexible harmonization network, as shown in Fig. 1b. The different sites can be understood as different nodes in this harmonization network, connected by edges played by DeRed. The harmonization network possesses a center-spoke topology with the target site, whose scanned images have the best data quality, at the center, and the data from the other sites are harmonized to this center site. Notably, scanning data from any site can be transferred to another site through the network edges. Furthermore, if a new site establishes a relationship with a site belonging to this harmonization network, it can also be transferred to any other site along the network edges.

2.2. Materials and T1 data processing

To minimize sampling bias across sites, we trained our harmonization framework using a traveling subject dataset from the DecNef Project Brain Data Repository (https://bicr-resource.atr.jp/srpbsts/), which was gathered by the Strategic Research Program for the Promotion of Brain Science (SRPBS) (Tanaka et al., 2021; Yamashita et al., 2019). This dataset included nine healthy participants (all male, ages 24-32 years), each of whom underwent T1-weighted MRI scans at 12 different centers. All of these sites used 3T scanners but with different manufacturers (Siemens, GE, and Philips), scanner types (Verio, Tim Trio, Spectra, Skyra, and Achieva), phase encoding directions (posterior to anterior and anterior to posterior), and numbers of channels per coil (8, 12, and 32). Data from three sites were excluded (ATT, UTO, and YC2) due to the inclusion of duplicate data. The detailed scanning parameters at each site are listed in Table 1. Written informed consent was obtained from the participants, and the data collection procedure was approved by the institutional review boards of each site.

In the current study, we selected the widely used GMV measurement (Grieve et al., 2013; Smallwood et al., 2013) derived from T1-weighted images as an example to examine the feasibility of the proposed harmonization method. The GMV was calculated by using Statistical Parametric Mapping (SPM12, https://www.fil.ion.ucl.ac.uk/spm/) (Ashburner, 2012) and the Computational Anatomy Toolbox (CAT12, http://dbm.neuro.uni-jena.de/cat12/) (Iglesias et al., 2015). Briefly, for each T1 scan, an N4 bias field inhomogeneity correction was first performed, and an adaptive maximum a posteriori (AMAP) approach was then used in tissue segmentation. Optimized shooting approach-based spatial registration was further performed to normalize all images to the standard Montreal Neurological Institute (MNI) space. Modulated normalization was then implemented to compensate for GMV changes caused by affine transformation and nonlinear warping. Finally, all GMV maps were smoothed with an 8-mm full-width at half-maximum (FWHM) Gaussian kernel.

2.3. Training and harmonization processes

ATV was selected as the target site (φ_t) in the harmonization process mainly for the following two reasons. First, the equipment manufacturer and number of channels per coil at ATV were the most frequently used among all the sites. Second, imaging data from ATV showed better quality with less noise than those from other sites during visual screening. Other sites were regarded as source sites (φ_s), resulting in 8 independent intersite harmonization pairings with ATV. For validation purposes, we also randomly chose another site as the target site (e.g., HUH) and retrained the DeRed harmonization framework based on the remaining eight source sites and the new target site.

Prior to the training process, we cropped all GMV maps from a matrix size of (181, 217, 181) to (176, 208, 176), which guaranteed that the sliced images could be restored to their original size after multiple-average pooling and upsampling operations. Moreover, to ensure the harmonization process within the gray matter regions and

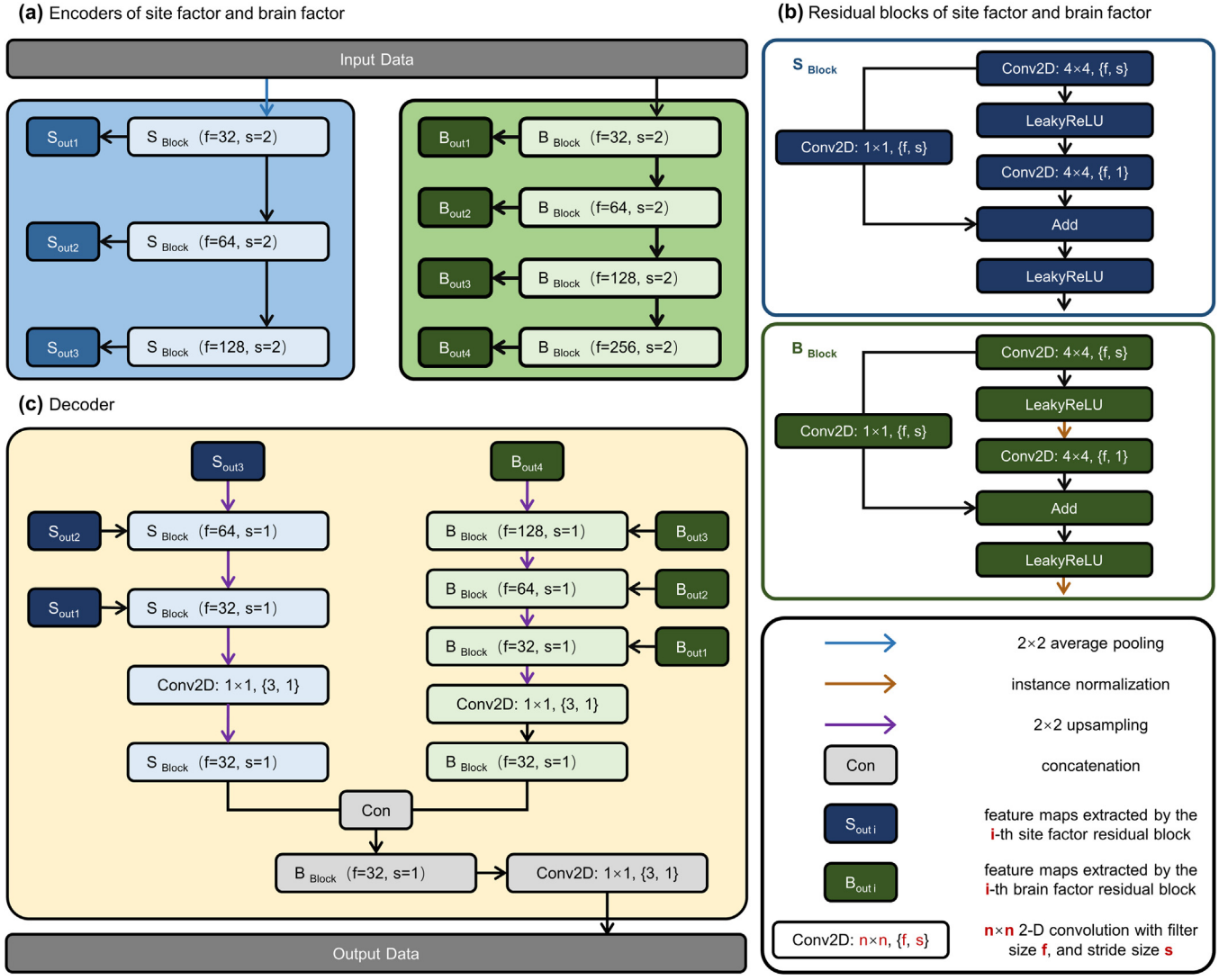


Fig. 2. Architecture of the encoders and decoder. (a) Architecture of the site factor encoder and brain factor encoder. The $S_{out\,i}$ and $B_{out\,i}$ models represent the feature maps extracted by the *i-th* site factor and brain factor residual blocks, respectively. **(b)** Architecture of the site factor residual block (S_{Block}) and brain factor residual block (B_{Block}). **(c)** Architecture of the decoder, which integrates the outputs from both the site factor and brain factor encoders.

reduce the computational burden, we constrained the data training process within a gray matter mask, which was determined by averaging the GMV maps of all scans and further applying a threshold of 0.2 mm^3.

The inputs of the training model were obtained by slicing along a certain anatomical direction (coronal, sagittal, or transverse); slices that did not intersect with the gray matter mask were not included in the subsequent training process. A section position was then randomly determined during each epoch to ensure uncertainty during the training process, and slices of imaging data of all subjects at φ_t and φ_s were extracted at this position. Notably, we hold that spatially adjacent slices assist in capturing brain factor representative information; therefore, we set the spatial resolution of the training slices to (176, 208, 3) for the transverse orientation, (176, 176, 3) for the coronal orientation, and (208, 176, 3) for the sagittal orientation. Thus, the *i-th* individual slice can be predicted repetitively at different channels for the *(i-1)-th, i-th* and *(i+1)-th* slice inputs. The images produced by the three channels were averaged to obtain the final harmonized single slice.

Furthermore, if the harmonization process is simply based on a single slicing direction, it cannot fully summarize the global spatial information contained in 3D images. Therefore, we independently trained three models. The training set of each model was obtained by slicing the image data from different anatomical directions, and then the output was averaged as the final harmonization result of the 3D image.

We defined four convergence constraint losses for the harmonization procedure:

First, we expect the site factor encoder to extract the same representation at the same site across different subjects:

$$Loss_{SiteConsistency} = \mathbb{E}_{x_s\sim\varphi_s} E\_site_s^i(x_s) - E\_site_s^i(x_s)_{\mu_1} + \mathbb{E}_{x_t\sim\varphi_t} E\_site_t^i(x_t) - E\_site_t^i(x_t)_{\mu_1} \quad (1)$$

where x_s and x_t denote the images from φ_s and φ_t, respectively. $E\_site_s^i(.)$ and $E\_site_t^i(.)$ denote the *i-th* feature map outputs of the *i-th* residual block in the site factor encoders of φ_s and φ_t, respectively. $E\_site_s^i(x_s)_\mu = \frac{1}{n}\sum_{x_k\sim\varphi_s} E\_site_s^i(x_k)$ and $E\_site_t^i(x_t)_\mu = \frac{1}{n}\sum_{x_k\sim\varphi_t} E\_site_t^i(x_k)$ denote the average *i-th* site factor residual block outputs of n subjects from φ_s and φ_t, respectively.

Second, we expect the brain factor encoders of φ_s and φ_t to extract the same representation from imaging data acquired from the same subject at different sites:

Table 1
Details of the scanning parameters in the traveling subject dataset.

Site	ATV	COI	HKH	HUH	KPM	KUS	KUT	SWA	YCI
Manufacturer	Siemens	Siemens	Siemens	GE	Philips	Siemens	Siemens	Siemens	Philips
Platform	Verio	Verio	Spectra	Signa HDxt	Achieva	Skyra	Tim Trio	Verio	Achieva
Magnetic field strength (T)	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0
Number of channels per coil	12	12	12	8	8	32	32	12	8
Phase encoding	PA	AP	PA	PA	AP	AP	PA	PA	AP
Echo time (ms)	2.98	2.98	2.38	1.928	3.31	2.98	3.4	2.98	3.176
Repetition time (ms)	2300	2300	1900	6788	7.1	2300	2000	2300	6.99
Flip angle (°)	9	9	10	20	10	9	8	9	9
Image dimension	240 × 256 × 256	176 × 240 × 256	224 × 320 × 320	180 × 256 × 256	170 × 256 × 256	224 × 232 × 256	240 × 256 × 208	240 × 256 × 256	200 × 256 × 256
Pixel dimension	1 × 1 × 1	1 × 1 × 1	0.81 × 0.75 × 0.75	1 × 1 × 1	1 × 1 × 1	1 × 1 × 1	0.9375 × 0.9375 × 1	1 × 1 × 1	1 × 1 × 1

Abbreviations: ATV, Siemens Verio scanner at the Advanced Telecommunications Research Institute International; COI, Center of Innovation at Hiroshima University; HKH, Hiroshima Kajikawa Hospital; HUH, Hiroshima University Hospital; KPM, Kyoto Prefectural University of Medicine; KUS, Siemens Skyra scanner at Kyoto University; KUT, Siemens Tim Trio scanner at Kyoto University; SWA, Showa University; YC1, Yaesu Clinic scanner 1; PA, posterior to anterior; AP, anterior to posterior.

$$Loss_{BrainConsistency} = \mathbb{E}_{x_s\sim\varphi_s,\ x_t\sim\varphi_t} E\_brain_s^i(x_s) - E\_brain_t^i(x_t)_1 \quad (2)$$

where $E\_brain_s^i(.)$ and $E\_brain_t^i(.)$ denote the *i-th* feature map outputs of the *i-th* residual block in the brain factor encoders from φ_s and φ_t, respectively.

Third, we encourage the decoders to reconstruct the images by merging the site factor representation and the brain factor representation from their own sites. This self-reconstruction loss can be formulated as:

$$Loss_{Self\,Reconstruction} = \mathbb{E}_{x_s\sim\varphi_s} x_s - \hat{x}_{s2} + \mathbb{E}_{x_t\sim\varphi_t} x_t - \hat{x}_{t2} \quad (3)$$

Fourth, the site factor representation from φ_t is necessary for the decoder in φ_t to reconstruct the images, even if the brain factor representation belongs to φ_s. In the same way, the decoder of φ_s can reconstruct images according to the site factor representation from φ_s and the brain factor representation from φ_t. The cross-reconstruction loss can be formulated as:

$$Loss_{CrossReconstruction} = \mathbb{E}_{x_s\sim\varphi_s} x_s - \tilde{x}_{s2} + \mathbb{E}_{x_t\sim\varphi_t} x_t - \tilde{x}_{t2} \quad (4)$$

where $\hat{x}_s = D_s(E\_brain_s(x_s), E\_site_s(x_s)_\mu)$ and $\tilde{x}_s = D_s(E\_brain_t(x_t), E\_site_s(x_s)_\mu)$ denote the reconstructed images, both of which contain the site factor representation from φ_s; however, their brain factor representations come from φ_s and φ_t, respectively. In contrast, $\hat{x}_t = D_t(E\_brain_t(x_t), E\_site_t(x_t)_\mu)$ and $\tilde{x}_t = D_t(E\_brain_s(x_s), E\_site_t(x_t)_\mu)$ denote the reconstructed images, both of which contain site factor representations from φ_t, but they derive their brain factor representations from φ_t and φ_s, respectively.

In summary, the total loss function of DeRed can be formulated as:

$$\begin{aligned} Loss_{total} &= \lambda_{SC} Loss_{siteConsistency} + \lambda_{BC} Loss_{brainConsistency} \\ &+ \lambda_{SR} Loss_{self\,Reconstruction} + \lambda_{CR} Loss_{crossReconstruction} \end{aligned} \quad (5)$$

where λ_{SC} =5, λ_{BC} =10, λ_{SR} =10 and $\lambda_{CR} = 20$ are the loss weights used to balance the contributions of the different terms. All the parameters were tuned by grid searches.

In our implementation, the proposed method was implemented by using the Keras library with TensorFlow 2.0 on an NVIDIA GeForce RTX 2080Ti GPU, and we adopted the Adam optimizer with the learning rate set to $1e^{-4}$. Training one DeRed model between two sites took approximately 2.5 to 3 hours with slices taken from nine subjects in a certain orthogonal direction (i.e., sagittal, transverse, or coronal).

2.4. Evaluation of harmonization outcome

We trained the DeRed harmonization network with a total of 81 images from all subjects scanned across all sites and obtained the corresponding harmonization results, which were used to quantify the inter-site differences and explain the representation captured by DeRed.

2.4.1. Correction for site effects

We adopted two methods to examine whether the proposed framework can reduce the site effects on the GMV maps. First, we performed linear discriminant analysis (LDA), a classic dimensionality reduction technique, to project the GMV measurement into two coordinates with the scanning site as a prior classification label. LDA is commonly used to project features into a lower dimension space by maximizing the distance between classes and minimizing the variation within each class. In this study, the site effect was reflected by the clustering of data from the same site. Second, we used one-way analysis of variance (ANOVA) to quantitatively test for significant site differences in GMV. Significance in the voxelwise comparison was denoted by a voxel-level $p<0.001$ with a cluster-level Gaussian random field (GRF)-corrected $p<0.05$.

2.4.2. Interpretability of the encoders

To assess whether each kind of encoder (i.e., site factor and brain factor) captured the corresponding features, we examined the output images by blocking the opposite input of the decoder in turn. To interpret the site factor encoder, we set all values of the brain factor feature

maps to zero and fed them into the decoder. The image synthesized in this case could be understood to contain only the site factor representation (i.e., I_{site}). Assuming that each site factor encoder captures the characteristics of the scanner, the intersite variance of I_{site} should be spatially similar to the intersite variance in the original GMV images. Thus, we first calculated the variance of each voxel of I_{site} and averaged the GMV variance maps of each subject across sites. We then applied Spearman's correlation to examine the spatial correlation of these two variance maps.

To interpret the brain factor encoder, we fed brain factor feature maps and empty site factor feature maps (i.e., feature maps with 0 values) into the decoder, and the image synthesized in this case could be understood to contain only the brain factor representation (i.e., I_{brain}). To examine whether the brain factor encoder truly captured these individual heterogeneity-related representations, we first assessed Spearman's correlation of each voxel between the original GMV and age across subjects, preserving voxel groups S_α whose GMV was significantly correlated with age. Then, we also preserved those voxel groups S_β with a significant correlation between I_{brain} and age. The overlap between S_α and S_β was then calculated by $\frac{S_\alpha \cap S_\beta}{S_\alpha \cup S_\beta}$. Second, the similarity between the original GMV and I_{brain} was calculated for each subject according to Spearman's correlation coefficient.

We also evaluated whether the I_{brain} could capture sex differences in GMVs. Briefly, 47 male and 47 female participants were selected from the open MRI dataset collected in KUT (Tanaka et al., 2021), ensuring that the age distribution (19-66 years, 35.94±13.17 years) was strictly matched between the two groups. The MRI equipment and scan parameters of the T1-weighted images were consistent with those of the KUT site in the traveling subject dataset. We used the same processing procedure to obtain the GMV map for each subject. The DeRed harmonization model trained on the traveling-subjects dataset was used to capture these I_{brain} maps while blocking the site factor decoder. We then used a two-sample t test to determine the sex differences between the GMV maps of the original data and the I_{brain} maps. The similarity between these two sex difference patterns (t-maps) was calculated by using Spearman's correlation. The overlap of the significant sex differences between the original data and the I_{brain} maps was calculated as $\frac{S'_\alpha \cap S'_\beta}{S'_\alpha \cup S'_\beta}$, where S'_α and S'_β were voxels with significant between-group differences in the original data and I_{brain} maps, respectively.

2.5. Comparison between DeRed and other harmonization methods

Several harmonization methods have been proposed to remove site effect differences in recent multicenter studies, including general linear model harmonization (GLM), global scaling harmonization (GS), and ComBat harmonization (see the Supplementary Information for detailed descriptions of these methods). To examine the advantages of our proposed methods, we compared DeRed with these harmonization methods in terms of site effect removal, GMV distribution coherence, intrasubject similarity improvement and intersubject difference reservation. A leave-one-subject-out cross-validation strategy was utilized for each method. Briefly, we excluded the data of the *i-th* subject at all sites, trained the framework with the remaining 72 scanned images from the other subjects, and applied the trained model to harmonize the data from the *i-th* subject. This procedure was repeated nine times to select each subject as the test data in turn.

2.5.1. Site effect removal

To test whether site effects could be removed by all the methods, we used ANOVA on the harmonized GMV maps for each method. Significance in the voxelwise comparison was denoted by a voxel-level $p<0.001$ with a cluster-level GRF-corrected $p<0.05$. Furthermore, we used Wilcoxon signed-rank tests to compare F values between the original and harmonized data and between harmonization results from different methods. Additionally, we examined whether the proposed method could remove site effects impacting covariance (Chen et al., 2022a). Briefly, for the original and harmonized GMV maps, we first divided the whole brain into 116 regions by using the Anatomical Automatic Labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002) and calculated a covariance matrix among the regions for each site. We then used the Euclidean distance measure to estimate the intersite similarity of the covariance matrix. Finally, the Wilcoxon signed-rank test was used to test the Euclidean distance differences between the original and harmonized GMV maps.

2.5.2. GMV distribution consistency

For the original data and harmonized data of each method, we first calculated the average GMV map across subjects and estimated their probability distribution for each site. We then estimated the averaged bidirectional KL divergence between each pair of probability distributions for different sites. KL divergence was further compared between the original and harmonized data and between harmonization results from different methods with Wilcoxon signed-rank tests. The averaged bidirectional KL divergence was calculated as:

$$KL_{Averaged}(P(x), Q(x)) = [KL(P(x), Q(x)) + KL(Q(x), P(x))]/2 \quad (6)$$

$$KL(P(x), Q(x)) = \sum_{i \in X} P(i) * \left[log \frac{P(i)}{Q(i)} \right] \quad (7)$$

where $P(x)$ and $Q(x)$ represent the probability distributions of the GMV values at site p and site q, respectively. For validation purposes, we also evaluated the GMV distribution consistency by using the JS divergence measure as follows:

$$JS(P(x), Q(x)) = \{KL(P(x), [P(x) + Q(x)]/2) + KL(Q(x), [P(x) + Q(x)]/2)\}/2 \quad (8)$$

2.5.3. Intersubject difference reservation

The differences across subjects were calculated using the Euclidean distances in the original GMV maps within each site and further averaged across all sites to obtain a reference intersubject difference matrix. Then, for the harmonization results from each method, we calculated the intersubject difference matrix within each site. Spearman's correlation was further used to estimate the correlation between each matrix and the reference matrix. A significant correlation coefficient indicated the preservation of intersubject differences.

2.5.4. Intrasubject similarity improvement

For each subject, we calculated Spearman's correlation coefficient between the GMV map of any pair of sites among the nine sites as the intrasubject similarity. These correlation coefficients were then compared between the original and harmonized data and between harmonization results from different methods using Wilcoxon signed-rank tests.

2.6. Expandability of the DeRed harmonization network

We evaluated the expandability of the proposed harmonization framework through a simulation in which data from a new site were incorporated into the network via a previously defined source site. Therefore, we randomly selected one site (YC1) as a newly included site and incorporated it into the ATV-centered harmonization network via a *double-jump* scheme. The data from YC1 were first transferred to KPM (i.e., $DeRed_a$) and then to the ATV via the DeRed framework between KPM and the ATV (i.e., $DeRed_b$). For both the original and harmonized data, paired t tests were used to quantify the significant site effects between YC1 and KPM and between YC1 and ATV, and a one-way ANOVA was used to assess the significant site effects across nine sites.

2.7. Performance with a smaller training sample

We tested the performance of the proposed method when training it with an even smaller sample. Four subjects were randomly selected to train the multisite center-spoke harmonization network, which was then applied to harmonize the GMV maps of the remaining five subjects. For the harmonized data, a one-way ANOVA was used to examine the site effects as previously described, and the root mean square error (RMSE) was calculated to estimate the intrasubject differences. This procedure was repeated ten times to avoid sampling bias. The F values derived from the ANOVA and RMSE were compared with those derived from the original data and the harmonized data obtained from the model trained with eight subjects.

3. Results

3.1. Site effect removal with DeRed

We first visualized the heterogeneity in the original and harmonized GMV maps across nine sites by projecting their dominant features into a 2D space using LDA decomposition. The site-clustered distribution of the LDA features indicated noticeable intersite heterogeneity in the original GMV maps (Fig. 3a). Specifically, data from HUH and HKH were the most distant from other datasets, which might essentially be due to their unique scanner models (GE Signa HDxt for HUH and Siemens Sepctra for HKH). However, the harmonized data showed a relatively homogeneous distribution, implying the effective removal of the site effect (Fig. 3b). Subsequent statistical analysis confirmed the finding from one-way ANOVA, which revealed significant site effects across the nine sites in the original GMV maps, primarily in the medial temporal and occipital cortices, insula, and cerebellum (Fig. 3c, voxel-level $p<0.001$, GRF-corrected $p<0.05$). In contrast, no significant site effect was observed in the harmonized GMV maps derived from our proposed DeRed framework (Fig. 3d, voxel-level $p<0.001$, GRF-corrected $p<0.05$). To further illustrate the order in which scan properties (e.g., MRI manufacturer, scanner type, and phase coding) contributed to the site effects, we performed hierarchical clustering of regions that showed significant site effects across the nine sites. We found that the scanner manufacturer was the factor with the greatest contribution to the site effect (Fig. S1). Moreover, validation analysis using site HUH as the target site also showed that the harmonized data exhibited a homogeneous distribution with site effects removed (Fig. S2), indicating the high robustness of the proposed framework.

3.2. Interpretability of the encoders

We examined the representations of the site factor and brain factor encoders by blocking their respective opposite outputs. As illustrated by randomly chosen data (e.g., sub-01 at YC1) in Fig. 4a, the outputs from the site factor encoders were decoded into a field map with abstract boundaries of the brain and blurry texture on the background. In contrast, images decoded from the brain factor encoders showed the detailed structure of the gray matter anatomy, which was highly similar to that of the original GMV maps. Further quantitative analysis showed that the intersite variance of I_{site} was significantly spatially correlated with the intersite variance of the original images in log-log coordinates (Fig. 4b, Spearman's correlation, $\rho=0.42$, $p<0.001$), suggesting that the site factor encoder captured the variance of actual physical factors across the scanner. For I_{brain}, we first found that these values were significantly spatially correlated with the original GMV maps for each individual at each site (Spearman's correlation, $\rho=0.993 \pm 0.002$, all $p<0.001$). We then examined the overlap of clusters showing significant correlations with those in the original data. In the original data, we found that the GMV was significantly positively correlated with age in the right precuneus, inferior frontal gyrus, and left parahippocampus and negatively correlated with age mainly in the dorsolateral prefrontal, visual, and lateral temporal cortices (voxel-level $p<0.001$, GRF-corrected $p<0.05$). I_{brain} showed similar brain-age correlation distributions at all nine sites (Fig. 4c, overlap ratio of the significant voxels: 75.54% ± 2.43%). Moreover, we examined the sex differences between the original data and the I_{brain} from KUT (voxel-level $p<0.001$; GRF-corrected $p<0.05$; Fig. 5a and Fig. 5b). The between-group t-maps of the original data and I_{brain} were significantly spatially similar (Spearman's correlation, $\rho=0.99$; $p<0.001$; Fig. 5c). Males exhibited significantly higher GMVs in the sensorimotor, visual, and orbitofrontal cortices, as well as in the hippocampus, than females. These between-group differences were highly overlapping between the original data and I_{brain} (97.65%; Fig. 5d). Together, these results suggest that the brain factor encoders successfully captured the biological details of the individual GMV maps.

3.3. DeRed showed better harmonization performance than conventional methods

We compared the performance of the proposed DeRed harmonization framework with those of several conventional methods, including GS, GLM, and ComBat. First, we found that significant site effects in the original data could be entirely eliminated by DeRed and ComBat but were partly retained in data processed with GS and GLM (Fig. 6a, ANOVA, voxel-level $p<0.001$, GRF-corrected $p<0.05$). Further between-method comparisons showed that the site effect (F value estimated in ANOVA) was significantly lower in the harmonized data from DeRed than in those from other methods (Fig. 6b, Wilcoxon signed-rank tests, $p<0.001$, Bonferroni-corrected). Moreover, the Euclidean distances among the covariance matrices of different sites were significantly shorter in the harmonized data derived from DeRed than in the original data (Wilcoxon signed-rank tests; $p<0.01$; Fig. S3), suggesting that DeRed could also significantly reduce site effects impacting covariance.

Second, we found that the probability distributions of the averaged GMV maps were divergent across the nine sites, and the distributions of the harmonized data tended to be more consistent (Fig. 7a). Quantitatively, the harmonized data derived from all methods showed significantly lower KL divergence than the original data, and data derived from DeRed exhibited the lowest KL divergence among the harmonization methods (Fig. 7b; Wilcoxon signed-rank tests; $p<0.001$; Bonferroni-corrected). The results of utilizing the JS divergence measure to estimate distribution consistency remained identical to the findings yielded with the KL divergence measure (Fig. S4).

Finally, the intersubject distance matrix for each site derived from the harmonized data produced by each method was significantly correlated with the original averaged matrix (Fig. 8a and Fig. S5, Spearman's correlation, $\rho=0.90 \pm 0.04$, all $p<0.001$), indicating that all harmonization methods maintained the intersubject differences in the GMV. Moreover, we found that the intrasubject similarity in the GMV was significantly increased by all harmonization methods (Wilcoxon signed-rank tests, $p<0.001$, Bonferroni-corrected). Importantly, DeRed demonstrated the statistically highest intrasubject similarity among all harmonization methods (Fig. 8b, Wilcoxon signed-rank tests, $p<0.001$, Bonferroni-corrected), indicating that the proposed framework has the greatest ability to increase intrasubject consistency across sites.

3.4. DeRed was expandable for adding new sites into the harmonization network

We added YC1 into the ATV-centered harmonization network again as a new site via a *double-jump* schema (newly trained YC1→KPM and previously trained KPM→ATV) (Fig. 9a). Significant site effects were observed in the original GMV maps between YC1 and KPM, between YC1 and ATV, and across all nine sites (Fig. 9b-9d, voxel-level $p<0.001$, GRF-corrected $p<0.05$). These site effects were removed by using the *double-jump* harmonization schema (Fig. 9b-9d, voxel-level $p<0.001$,

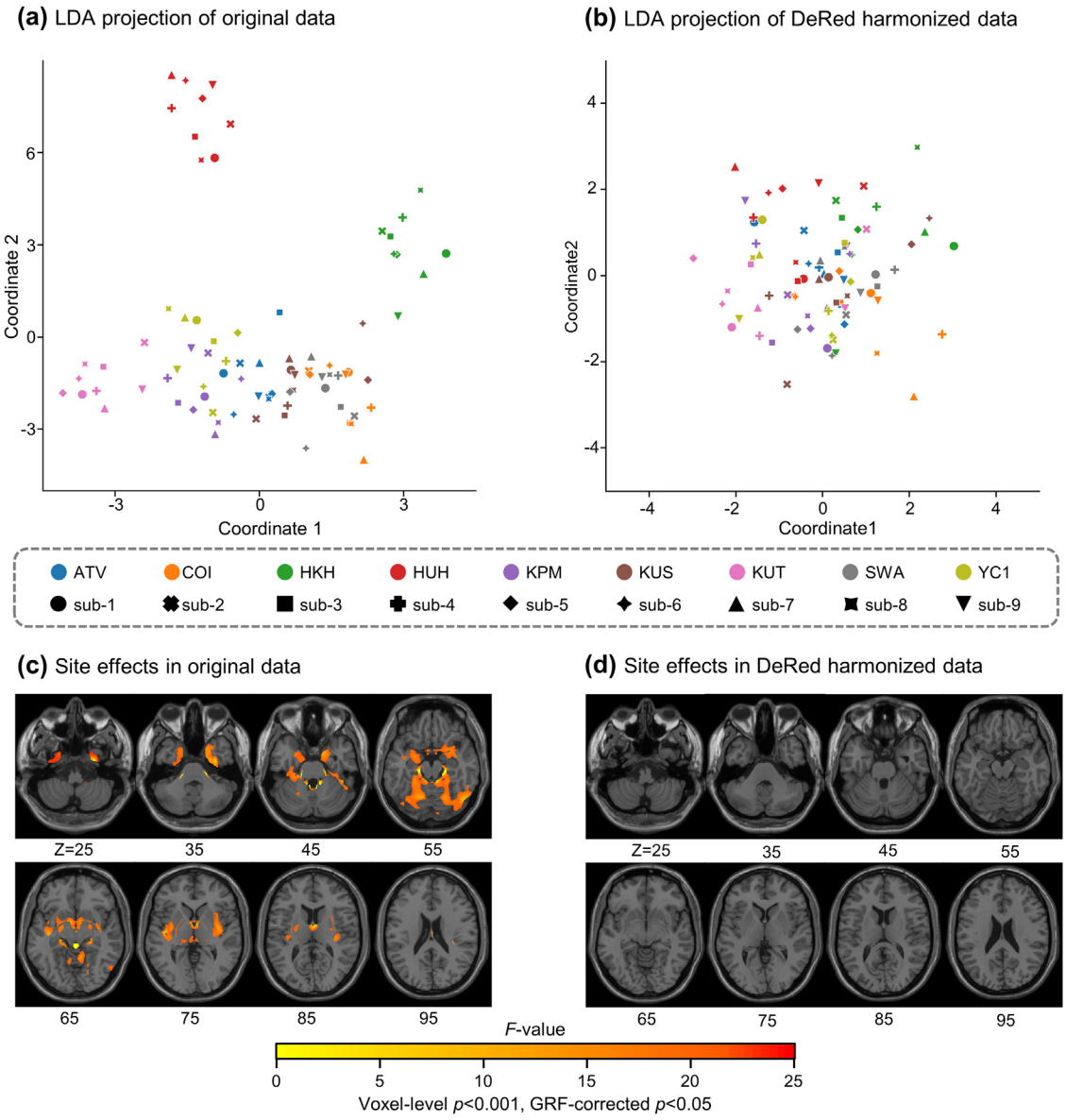


Fig. 3. Site effects in data before and after harmonization. (a) and **(b)** illustrate the LDA projection of the GMV before and after harmonization. A datapoint represents a projected GMV measurement from a subject; its color represents the site from which it originates, and its shape represents the subject to which it belongs. **(c)** and **(d)** illustrate the site effects identified by one-way ANOVA in the original and harmonized GMV sliced along the transverse anatomical orientation. There were no significant differences across sites for all voxels after DeRed harmonization.

GRF-corrected $p<0.05$), suggesting excellent expandability of the proposed harmonization network.

3.5. *Training with a smaller sample size can reduce site effects*

When training our DeRed model with only four subjects, we found that this 4-subject model could significantly remove site effects in 5 of 10 repeated experiments, while ≤2 small clusters with site effects were identified in the remaining 5 cases (voxel-level $p<0.001$; GRF-corrected $p<0.05$). Moreover, although the site effects (F-values) and intrasubject differences (RMSEs) in the harmonized data derived from the 4-subject model were higher than those derived from the 8-subject model in the main results, they were significantly lower than those in the original data (Fig. S6; Wilcoxon signed-rank tests; $p<0.001$). These results suggest that the proposed DeRed model could reduce site effects even with a small training sample; however, a larger training sample (e.g., eight subjects) could achieve better performance.

4. Discussion

In this paper, we proposed a DL-based harmonization framework for multisite MRI data named DeRed, which was further trained with a traveling subject dataset. Taking the commonly used GMV metric as an example, the proposed framework showed good performance in eliminating the divergence in the GMV across different sites. Notably, the encoders embedded in the framework successfully captured both the abstract textures of site factors and the concrete biologically related brain features. Moreover, the proposed framework exhibited outstanding performance relative to conventional harmonization methods in site effect removal, data distribution homogenization, and intrasubject similarity improvement. Together, the proposed method offers a powerful and extendable DL-based harmonization framework for multisite neuroimaging data with high interpretability, facilitating the improvement of the reliability and reproducibility of multisite studies for brain development and brain disorders.

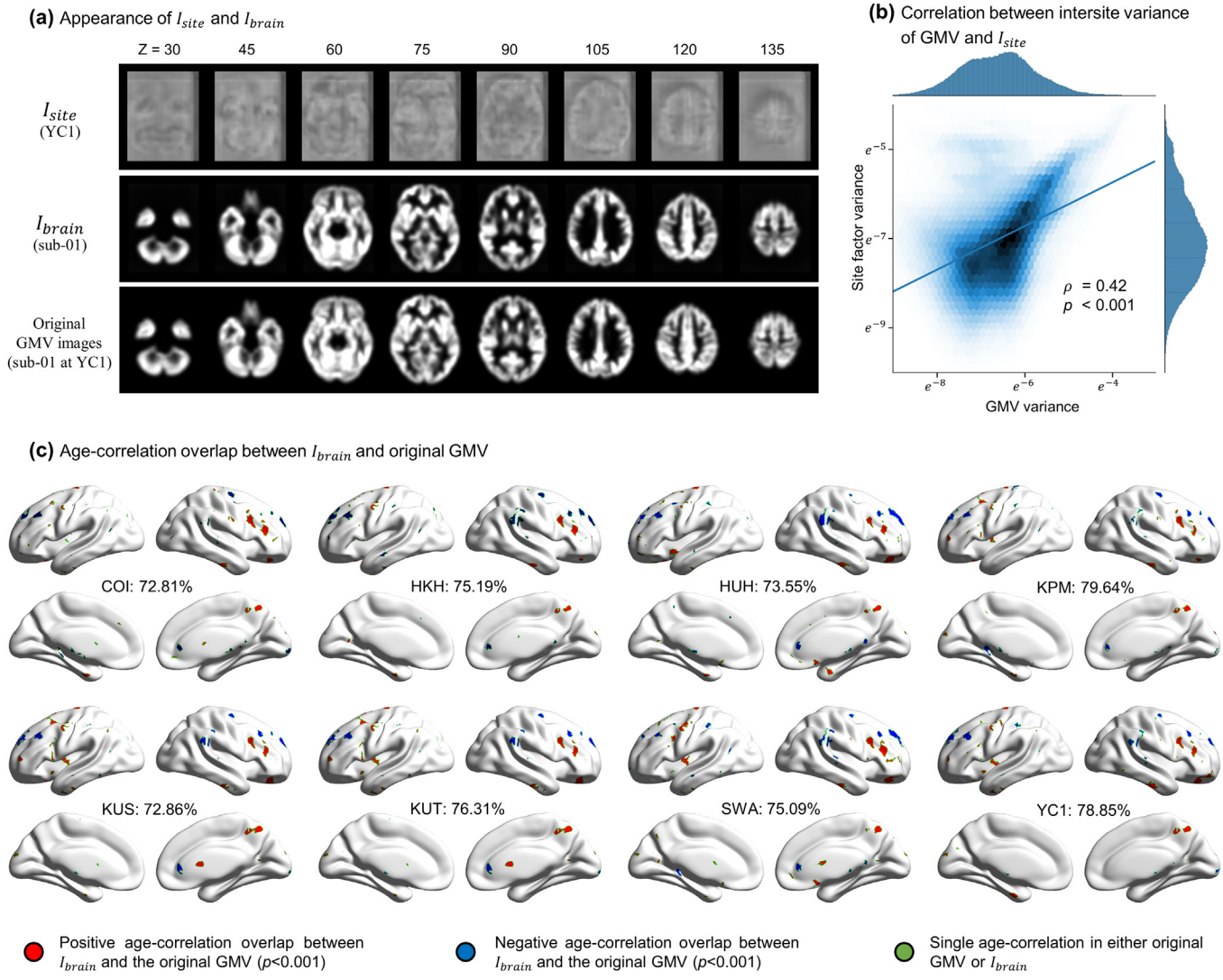


Fig. 4. Interpretability of site factor and brain factor encoders. (a) Appearance of the site factor (I_{site}) and brain factor (I_{brain}) feature maps. The first row represents the decoder outputs containing only the site factor representations of YC1. The second row represents the decoder outputs containing only the brain factor representations of subject-01. The last row represents the original GMV map of subject-01 from YC1. **(b)** Log-log correlation plot between the original GMV variance and the variance of the site factor feature maps. Each variance measurement is transformed by natural logarithm conversion. The color depth reflects the dot density within a single hexagon (Spearman's correlation ρ =0.42; p<0.001). **(c)** Age-correlation overlap clusters between brain factor mapping and the original GMV. The voxels in the red (blue) regions indicate positive (negative) age correlations (p<0.001) in both the original GMV and the brain factor mapping. Voxels colored green indicate a single age correlation in either the original GMV or brain factor mapping.

Compared with conventional statistics-based harmonization methods, the advantages of the proposed DL-based framework can be formulated from several perspectives. First, instead of taking a single metric as an independent variable, the DL model comprehensively extracts the global and local imaging information by integrating information from spatially neighboring units (e.g., voxels in a brain map) through a series of convolution and pooling operations (Bau et al., 2020). Many studies have suggested that adjacent voxels reflect closer correlations both in the anatomical structure and in the physiological mechanism (Cao et al., 2017b; Cigdem et al., 2019). These individual-specific anatomical details embodied within the MR images are repeatable across multisite measurements and should not be ignored during the harmonization process. Second, both DL-based methods and statistics-based methods attempt to explore the mapping relationship during the harmonization process. However, harmonization processes guided by statistical strategies, such as GLM and GS, seem to be limited in the ability to map linear polynomial functions. In our work, we employed a residual block inside the proposed framework, which has been shown to be especially important for fitting a more accurate function map, including a variety of high-dimensional and nonlinear characteristics between MR images and site effect representations (Lusch et al., 2018). Third, statistics-based harmonization frameworks scrupulously rely on the prior assumption. For example, ComBat describes the site effect of each voxel via additive and multiplicative factors, which are assumed to follow a normal distribution and inverse gamma distribution, respectively (Johnson et al., 2007). Nevertheless, the site effect reflected within the MR images can be understood as a heterogeneous mixture caused by the action of an asymmetrical magnetic field and complex neurophysiological activity (Vovk et al., 2007), which is difficult to generalize adequately with simple probability distributions. Compared with statistics-based methods, the proposed harmonization framework driven by the pixel-to-pixel loss function is not limited to the prior distribution assumption, allowing the harmonization results of DeRed to demonstrate better probability consistency across sites. Fourth, benefiting from the bidirectional design of the DeRed framework and the center-spoke design of the harmonization network, newly incorporated sites could be easily harmonized to the tar-

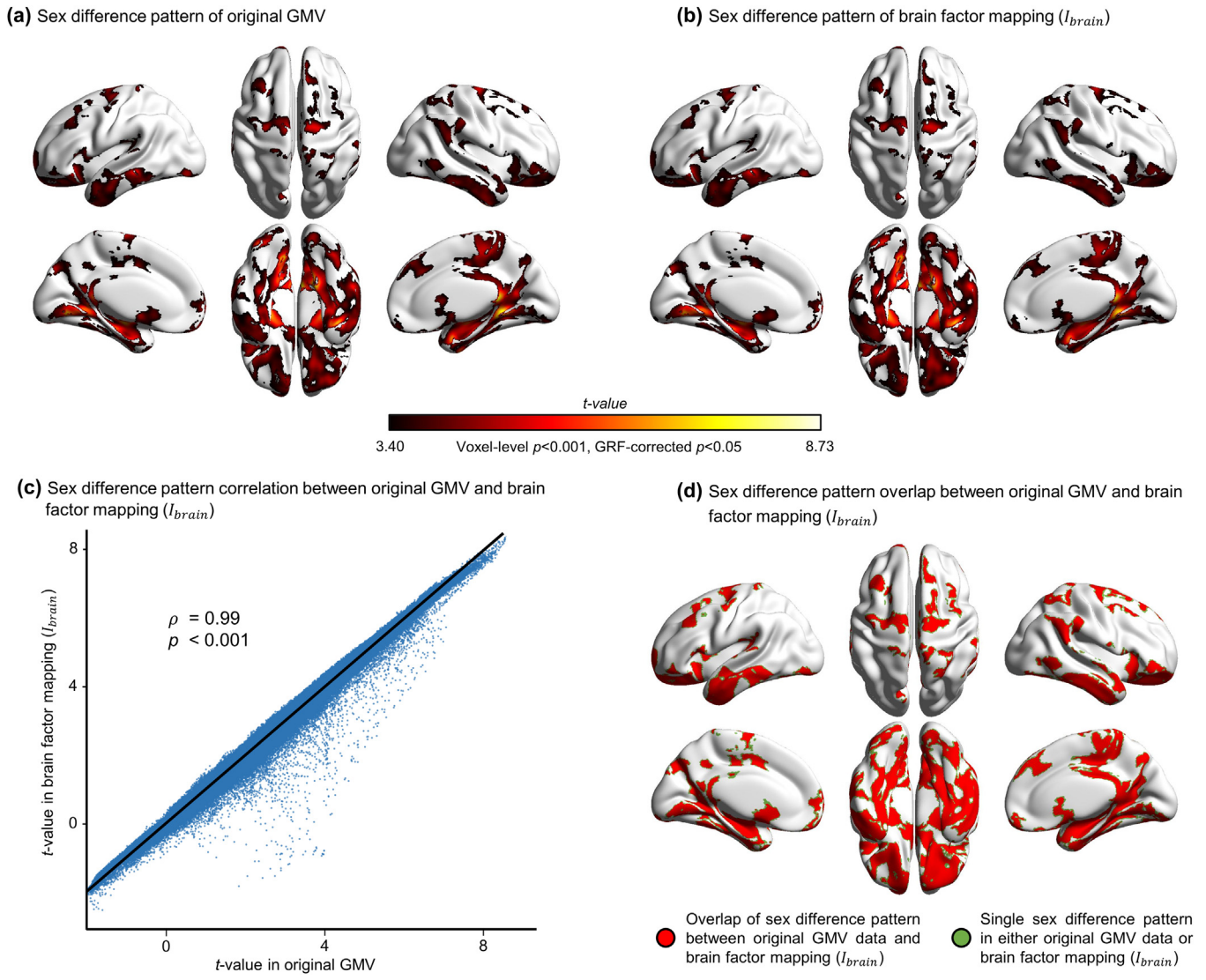


Fig. 5. Sex difference patterns of the original GMV data and the brain factor mapping. **(a)** and **(b)** illustrate the sex difference patterns of the original GMV data and I_{brain}, respectively (two-sample t test; voxel-level p<0.001; GRF-corrected p<0.05). **(c)** Correlation between the sex difference patterns of the original GMV data and I_{brain} (Spearman's correlation ρ>0.997; p<0.001). **(d)** Overlap between the sex difference patterns of the original GMV data and the brain factor representation. The voxels in red regions indicate significant sex-related differences in both the original GMV data and I_{brain}. Voxels colored green indicate single sex-related differences in either the original GMV data or I_{brain}. Voxels representing sex-related differences were highly overlapping (97.65%) between the original GMV data and I_{brain}.

get site via indirect schema without retraining the whole original harmonization network. In contrast, most statistics-based harmonization methods are sensitive to the included sites and require retraining upon incorporating a new site or removing an existing site. Therefore, the proposed framework exhibited more practical expandability than conventional statistics-based harmonization methods. Finally, a previous study demonstrated that the Bayesian-based ComBat harmonization method has the advantage of being able to estimate and remove site effects with a small sample size of twenty subjects (Fortin et al., 2017). Although a smaller training set of eight subjects was used in the current study, we further showed that the proposed framework could reduce site effects and increase intrasubject consistency when trained with only four subjects. Such an advantage provided by this DL-based model might be due to its use of a different training scheme from those of conventional statistical models. The GMV map of each subject was sliced into 560 individual images at all three orientations and then fed into the training procedure, which increased the utilization efficiency of the training sample.

The disentangled representations of site and brain factors during the encoding stage followed by a combined decoding procedure enabled the conversion of GMV maps from the source to target sites. Although we found that the GMV maps directly decoded from the brain factor encoder could capture biological information, the decoding procedure for combining brain factors and site factors was not a simple addition but could involve more complex interactions. Thus, the GMV maps reconstructed from the framework where the site factor decoder was simply blocked may not be suitable for consideration as final harmonized images. To test this speculation, we further examined the within-subject differences by using the RMSEs between the images in the original data, the reconstructed images without the site factor (I_{brain}), and the features extracted in the brain factor encoder (F_{brain}). We found that the RMSEs were significantly lower in I_{brain} and F_{brain} than in the original data (Wilcoxon signed-rank tests; p<0.001; Fig. S7), suggesting that the site factor was indeed extracted by the site factor encoder. Furthermore, F_{brain} yielded a significantly lower RMSE than I_{brain} (Wilcoxon signed-rank tests; p<0.001; Fig. S7), indicating that the features extracted in

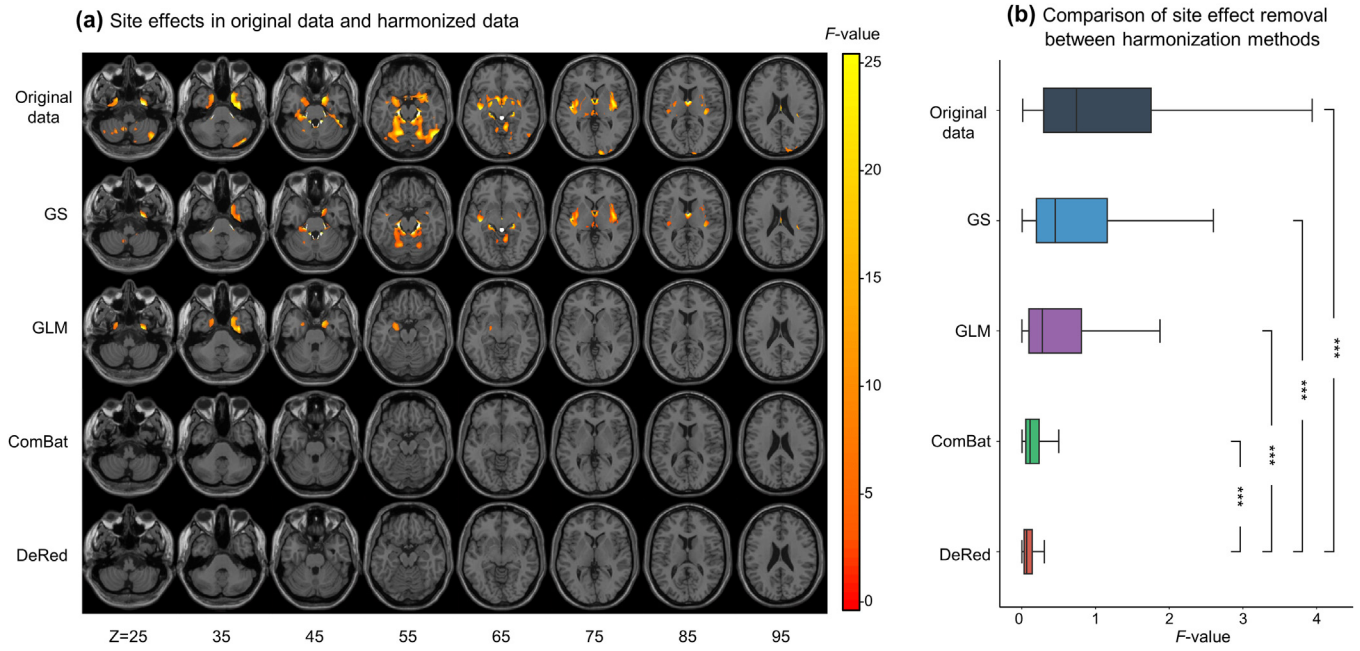


Fig. 6. Site effects removal of different harmonization methods. (a) Site effects identified by ANOVA in data before and after harmonization (voxel-level $p<0.001$, GRF-corrected $p<0.05$). **(b)** Comparison of site effects (F value) in data before and after harmonization by different methods. All harmonization results exhibit lower F values than those of the original state (Wilcoxon signed-rank tests; Bonferroni-corrected), and the F values of the DeRed harmonized data are significantly lower than those of other methods (Wilcoxon signed-rank tests; Bonferroni-corrected). ***, $p<0.001$.

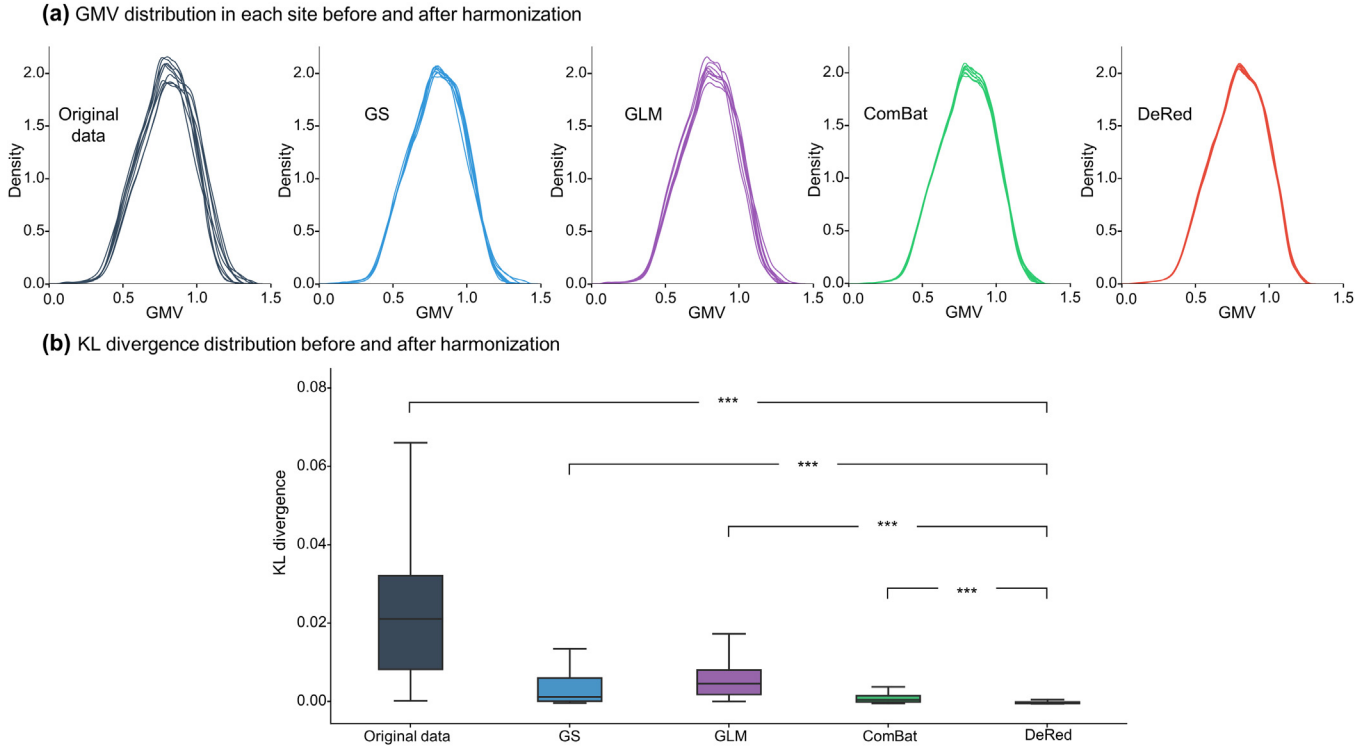


Fig. 7. Divergence in the GMV distribution across different sites before and after harmonization. (a) GMV distribution in different sites. Each curve represents the probability distribution of the GMV measurement for all voxels averaged across subjects in a site. **(b)** Boxplots of KL divergence across sites before and after harmonization by different methods. All harmonization data yielded lower KL divergence measures than that of the original data (Wilcoxon signed-rank tests; Bonferroni-corrected). DeRed demonstrated significantly lower KL divergences than the compared methods (Wilcoxon signed-rank tests; Bonferroni-corrected). ***, $p<0.001$.

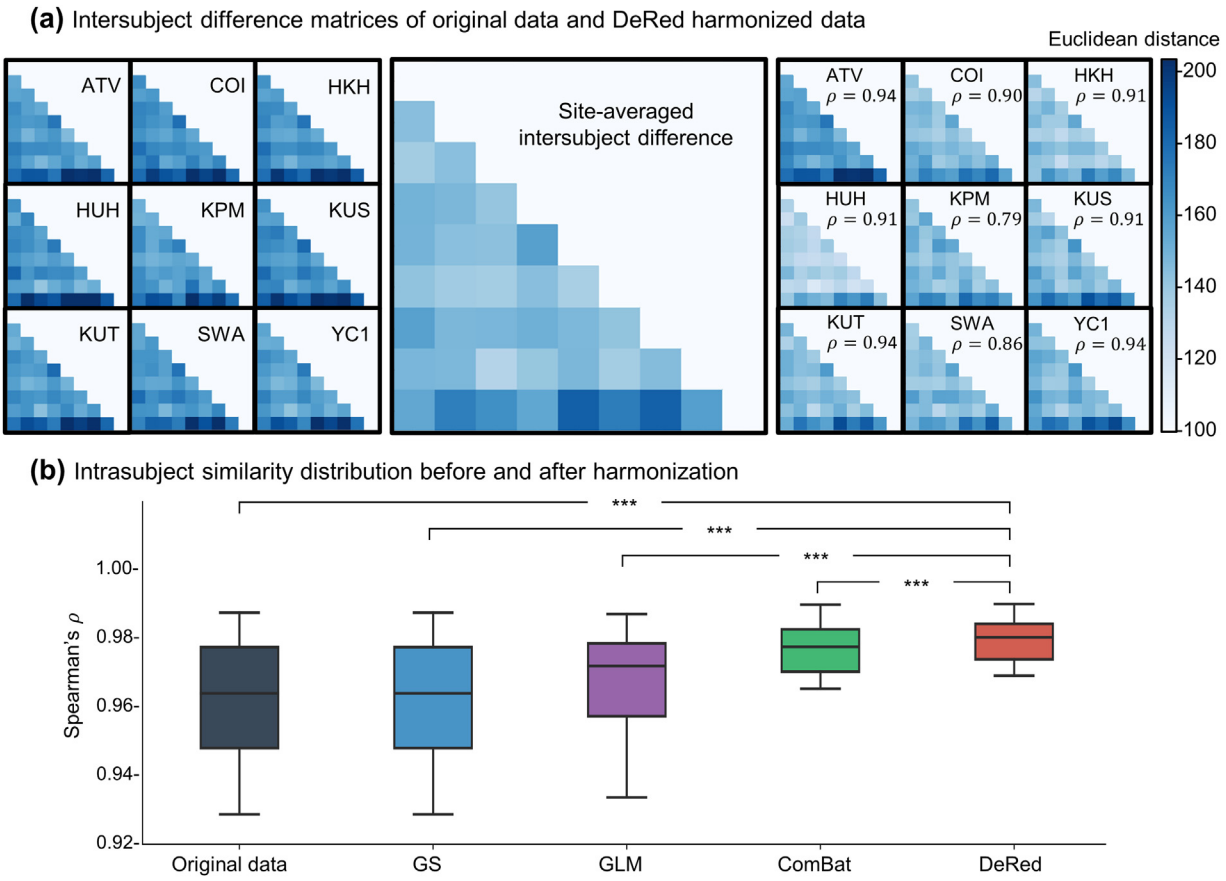


Fig. 8. Intersubject difference maintenance and intrasubject similarity improvement before and after harmonization. (a) Intersubject difference matrix before and after harmonization at each site. The difference matrices were averaged across sites before harmonization. The color depth of the *i-th* row and *j-th* column grid in each matrix represents the Euclidean distance between the *i-th* and *j-th* subjects. Spearman's correlation coefficients are shown as ρ (p<0.001). **(b)** Boxplots of the cross-site intrasubject similarities before and after harmonization with different methods. DeRed demonstrated significantly improved intrasubject similarity (Wilcoxon signed-rank tests). ***, p<0.001.

the brain factor encoder captured purer biological characteristics and that the decoding procedure might have added complex site features to the reconstructed images.

In the current study, we mainly employed the voxel-based GMV, which is a commonly used structural brain measurement, to validate our proposed framework. It should be noted that the proposed DL-based representation disentanglement and reconstruction strategy can be referenced to other multisite harmonization processes for structural and functional brain metrics in diverse formats. Regarding the volumetric images, our proposed framework can provide a robust contribution by fine tuning its network architecture. Designs similar to those of the embedded encoders and decoders in DeRed were adopted to extract the latent representations and remove artifacts in T1-w and T2-w MR images in a previous study (Liu et al., 2021). Of note, we set the input image size as 176 × 208 × 176 to ensure that each dimension was divisible by 2^4 so that it could pass the pooling and upsampling layers while minimally cropping the image. For images with different resolutions (e.g., 3-mm resolutions for fMRI data and 2-mm resolutions for DTI data), two methods could be utilized: resampling these images to 1-mm resolutions and cropping their borders to fit the input image size or adjusting the size of each residual block to adapt to different image sizes. For those data in a network format, such as the structural and functional connectivity matrices, the graph convolutional network (GCN) can be integrated into the proposed framework. Many studies have applied the GCN to reveal functional brain network similarity with comprehensive consideration of topological properties (Ktena et al., 2018) and to more efficiently predict the longitudinal development of cognitive performance (e.g., motor and cognitive scores) in preterm infants by identifying the local and global topology patterns of their structural brain networks (Kawahara et al., 2017). Illuminated by existing studies, the GCN can be used to depict complex topological mechanisms and identify abstract high-dimensional information, indicating that the application of the GCN may help to capture the site-specific topological effect, from which multisite structural and functional brain network harmonization can be reasonably performed.

Several issues and future directions should be further considered. First, to validate the harmonization effect of the proposed framework, we trained our model with a traveling subject dataset, which minimized the sampling bias across the scan sites. Although a recent study showed that a traveling subject dataset could improve the site effect estimation and removal abilities of conventional statistics-based harmonization methods (Maikusa et al., 2021), our results indicate that the proposed DL-based framework achieved better performance in terms of site effect removal, data distribution homogenization, and intrasubject similarity improvement than the statistical methods on the same traveling subject dataset. However, traveling subject MRI data collection designs are generally lacking in many multisite databases; thus, further training strategies (e.g., random bootstrap sampling) should be developed for unpaired intersite datasets. Second, the traveling subject dataset used in this work was acquired from a group of healthy participants aged from 24 to 32 years, and the biological validation was limited due to the lack of cognitive or clinical evaluations; thus, the generalizability of DeRed to MRI data acquired from special populations (e.g., children and adolescents or patients with brain disorders) needs to be further validated. Studies

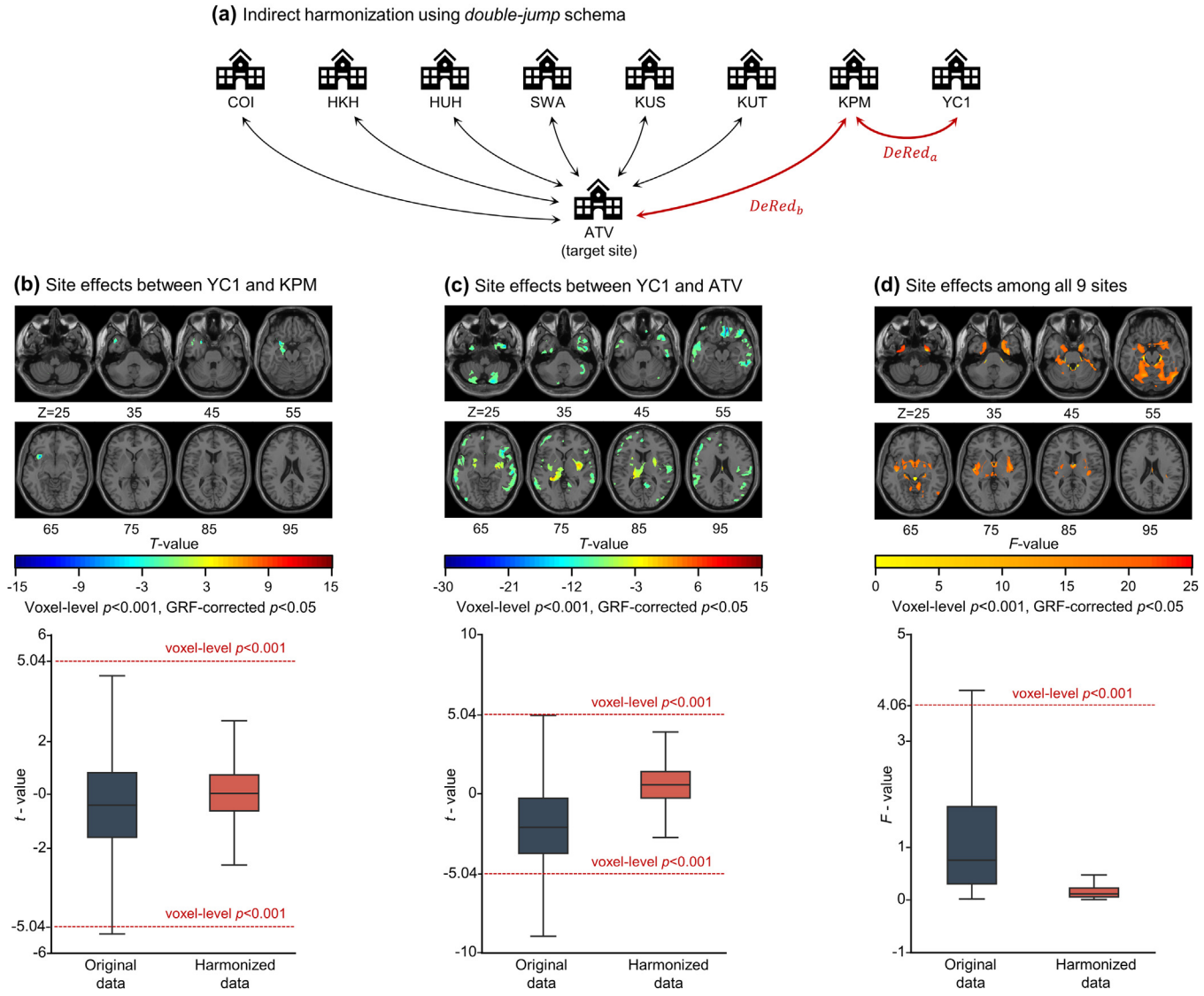


Fig. 9. Expandability of the center-spoke harmonization network. (a) Indirect harmonization using *double-jump* schema. Data from YC1 were first harmonized to KPM using $DeRed_a$ and then indirectly harmonized to ATV using $DeRed_b$. **(b)** Paired *t* test results of GMV between YC1 and KPM before and after first-jump harmonization using $DeRed_a$. **(c)** Paired *t* test results of GMV between YC1 and ATV before and after second-jump harmonization using $DeRed_a$ and $DeRed_b$. **(d)** ANOVA results among all sites before and after harmonization. The black boxplots represent the distributions of *F values* for the original GMV maps. The red boxplots represent the *F* value distributions of the harmonized data, where data from YC1 were indirectly transferred to ATV data, while data from other sites were directly transferred to ATV data.

have revealed significant development effects and disorder-related disruptions in brain structures and functions (Gilmore et al., 2018; van den Heuvel and Sporns, 2019). Therefore, the specific optimization strategy for harmonization methods needs further investigation for these special populations. Third, regarding clinical multicenter imaging data, neither the conventional statistics-based harmonization methods nor the novel DL-based method can be appropriately established when the patients and controls were separately scanned by different scanners. In such cases, the measurement bias and sampling bias highly overlap and cannot be corrected. However, we found that the proposed DL-model could work with a very small sample of traveling subjects, indicating its potential advantage in biased sample data over conventional statistics-based methods. Given that site effects may influence the case-control differences between patients and controls (Xia et al., 2019), the application of DL-based methods requires further validation on multisite brain disorder datasets, and perhaps specific models should be trained for different disorders. Fourth, conventional statistics-based harmonization methods have the advantage of employing flexible matrix designs to adapt to different types of databases, such as longitudinal (Beer et al., 2020) and distributed datasets (Chen et al., 2022b). Correspondingly, the proposed DL-based method provides a general framework to accommodate these special datasets, where the loss function can be appropriately modified by adding additional constraints for different types of training data to promote the convergence of the model and improve the interpretability of latent representations. Fifth, similar to most DL methods, the proposed DeRed framework comprises several convolution and pooling operations. Therefore, the harmonized data can be objectively smoothed based on neighboring information during encoding and decoding. Although this procedure overcomes local noise during harmonization, further validations of the data distribution and design optimization for the DL network are needed. Finally, compared with conventional statistics-based methods, DL-based methods require more time for model training. However, this increased training time is acceptable, as it is counted in hours, and the application of the method requires no additional time once the models are established. Future development on computational hardware, particularly GPUs, will further shorten the required

training time and thus support more complex DL-based harmonization methods.

Credit Author Statement

Dezheng Tian: Conceptualization, Methodology, Software, Data Curation, Formal analysis, Writing - Original Draft, Writing - Review & Editing. **Zilong Zeng:** Methodology, Software, Validation. **Xiaoyi Sun:** Software, Validation, Writing - Review & Editing. **Qiqi Tong:** Writing - Review & Editing. **Huanjie Li:** Writing - Review & Editing. **Hongjian He:** Writing - Review & Editing. **Jia-Hong Gao:** Writing - Review & Editing. **Yong He:** Resources, Supervision, Project administration, Funding acquisition. **Mingrui Xia:** Conceptualization, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Data and code availability

All MRI data used in this study are publicly available to anyone agreeing to the Open Access Data Use Terms at the DecNef Project Brain Data Repository website (https://bicr-resource.atr.jp/srpbsts/). The source code and trained models are available on GitHub (https://github.com/DezhengTian/DeRed-Harmonization).

Acknowledgments

This work was supported by National Natural Science Foundation of China (Nos. 82071998, 82021004, 81671767, and 81620108016), Beijing Nova Program (No. Z191100001119023), Fundamental Research Funds for the Central Universities (No. 2020NTST29), the National Key R&D Program of China (No. 2018YFA0701400), the Changjiang Scholar Professorship Award (No. T2015027), and Beijing United Imaging Research Institute of Intelligent Imaging Foundation (No. CRIBJZD202102).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2022.119297.

References

Ashburner, J., 2012. SPM: a history. Neuroimage 62, 791–800.
Bau, D., Zhu, J.Y., Strobelt, H., Lapedriza, A., Zhou, B., Torralba, A., 2020. Understanding the role of individual units in a deep neural network. Proc. Natl. Acad. Sci. U. S. A. 117, 30071–30078.
Beer, J.C., Tustison, N.J., Cook, P.A., Davatzikos, C., Sheline, Y.I., Shinohara, R.T., Linn, K.A.Alzheimer's Disease Neuroimaging, I., 2020. Longitudinal ComBat: A method for harmonizing longitudinal multi-scanner imaging data. Neuroimage 220, 117129.
Cao, M., Huang, H., He, Y., 2017a. Developmental connectomics from infancy through early childhood. Trends Neurosci. 40, 494–506.
Cao, S., Li, Y., Deng, W.W., Qin, B.Y., Zhang, Y., Xie, P., Yuan, J., Yu, B.W., Yu, T., 2017b. Local brain activity differences between herpes zoster and postherpetic neuralgia patients: a resting-state functional MRI study. Pain Physician 20, E687–E699.
Casey, B.J., Cannonier, T., Conley, M.I., Cohen, A.O., Barch, D.M., Heitzeg, M.M., Soules, M.E., Teslovich, T., Dellarco, D.V., Garavan, H., Orr, C.A., Wager, T.D., Banich, M.T., Speer, N.K., Sutherland, M.T., Riedel, M.C., Dick, A.S., Bjork, J.M., Thomas, K.M., Chaarani, B., Mejia, M.H., Hagler, D.J., Cornejo, M.D., Sicat, C.S., Harms, M.P., Dosenbach, N.U.F., Rosenberg, M., Earl, E., Bartsch, H., Watts, R., Polimeni, J.R., Kuperman, J.M., Fair, D.A., Dale, A.M., Workgrp, A.I.A., 2018. The adolescent brain cognitive development (ABCD) study: Imaging acquisition across 21 sites. Dev. Cogn. Neuros-Neth 32, 43–54.
Chen, A.A., Beer, J.C., Tustison, N.J., Cook, P.A., Shinohara, R.T., Shou, H.Alzheimer's Disease Neuroimaging, I., 2022a. Mitigating site effects in covariance for machine learning in neuroimaging data. Hum. Brain Mapp. 43, 1179–1195.
Chen, A.A., Luo, C., Chen, Y., Shinohara, R.T., Shou, H.Alzheimer's Disease Neuroimaging, I., 2022b. Privacy-preserving harmonization via distributed ComBat. Neuroimage 248, 118822.
Chen, J, Sun, Y, Fang, Z, Lin, W, Li, G, Wang, LUNC UMN Baby Connectome Project Consortium, 2022c. Harmonized neonatal brain MR image segmentation model for cross-site datasets. Biomed Signal Process Control 69, 102810.
Cigdem, O., Demirel, H., Unay, D., 2019. The performance of local-learning based clustering feature selection method on the diagnosis of parkinson's disease using structural MRI. 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), pp. 1286–1291.
Dewey, B.E., Zhao, C., Reinhold, J.C., Carass, A., Fitzgerald, K.C., Sotirchos, E.S., Saidha, S., Oh, J., Pham, D.L., Calabresi, P.A., van Zijl, P.C.M., Prince, J.L., 2019. DeepHarmony: a deep learning approach to contrast harmonization across scanner changes. Magn. Reson. Imaging 64, 160–170.
Fornito, A., Zalesky, A., Breakspear, M., 2015. The connectomics of brain disorders. Nat. Rev. Neurosci. 16, 159–172.
Fortin, J.P., Cullen, N., Sheline, Y.I., Taylor, W.D., Aselcioglu, I., Cook, P.A., Adams, P., Cooper, C., Fava, M., McGrath, P.J., McInnis, M., Phillips, M.L., Trivedi, M.H., Weissman, M.M., Shinohara, R.T., 2018. Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 167, 104–120.
Fortin, J.P., Parker, D., Tunc, B., Watanabe, T., Elliott, M.A., Ruparel, K., Roalf, D.R., Satterthwaite, T.D., Gur, R.C., Gur, R.E., Schultz, R.T., Verma, R., Shinohara, R.T., 2017. Harmonization of multi-site diffusion tensor imaging data. Neuroimage 161, 149–170.
Garcia-Dias, R., Scarpazza, C., Baecker, L., Vieira, S., Pinaya, W.H.L., Corvin, A., Redolfi, A., Nelson, B., Crespo-Facorro, B., McDonald, C., Tordesillas-Gutierrez, D., Cannon, D., Mothersill, D., Hernaus, D., Morris, D., Setien-Suero, E., Donohoe, G., Frisoni, G., Tronchin, G., Sato, J., Marcelis, M., Kempton, M., van Haren, N.E.M., Gruber, O., McGorry, P., Amminger, P., McGuire, P., Gong, Q.Y., Kahn, R.S., Ayesa-Arriola, R., van Amelsvoort, T., de la Foz, V.O.G., Calhoun, V., Cahn, W., Mechelli, A., 2020. Neuroharmony: a new tool for harmonizing volumetric MRI data from unseen scanners. Neuroimage 220, 117–127.
Gilmore, J.H., Knickmeyer, R.C., Gao, W., 2018. Imaging structural and functional brain development in early childhood. Nat. Rev. Neurosci. 19, 123–137.
Grieve, S.M., Korgaonkar, M.S., Koslow, S.H., Gordon, E., Williams, L.M., 2013. Widespread reductions in gray matter volume in depression. Neuroimage Clin. 3, 332–339.
He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J., 2016a. Deep Residual Learning for Image Recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.
He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J., 2016b. Identity Mappings in Deep Residual Networks. European conference on computer vision, Springer, pp. 630–645.
Huang, X., Belongie, S., 2017. Arbitrary style transfer in real-time with adaptive instance normalization. Proceedings of the IEEE International conference on computer vision, pp. 1510–1519.
Iglesias, J.E., Augustinack, J.C., Nguyen, K., Player, C.M., Player, A., Wright, M., Roy, N., Frosch, M.P., McKee, A.C., Wald, L.L., Fischl, B., Van Leemput, K., Neuroimaging, A.D., 2015. A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: application to adaptive segmentation of in vivo MRI. Neuroimage 115, 117–137.
Johnson, W.E., Li, C., Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–127.
Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J.G., Hamarneh, G., 2017. BrainNetCNN: convolutional neural networks for brain networks; towards predicting neurodevelopment. Neuroimage 146, 1038–1049.
Ktena, S.I., Parisot, S., Ferrante, E., Rajchl, M., Lee, M., Glocker, B., Rueckert, D., 2018. Metric learning with spectral graph convolutions on brain connectivity networks. Neuroimage 169, 431–442.
Laird, A.R., 2021. Large, open datasets for human connectomics research: Considerations for reproducible and responsible data use. Neuroimage 244, 118579.
Li, H.J., Smith, S.M., Gruber, S., Lukas, S.E., Silveri, M.M., Hill, K.P., Killgore, W.D.S., Nickerson, L.D., 2020. Denoising scanner effects from multimodal MRI data using linked independent component analysis. Neuroimage 208, 116388.
Liu, S.Y., Thung, K.H., Qu, L.Q., Lin, W.L., Shen, D.G., Yap, P.T., 2021. Learning MRI artefact removal with unpaired data. Nat. Mach. Intell. 3, 60–67.
Lusch, B., Kutz, J.N., Brunton, S.L., 2018. Deep learning for universal linear embeddings of nonlinear dynamics. Nat. Commun. 9, 4950.
Maikusa, N., Zhu, Y., Uematsu, A., Yamashita, A., Saotome, K., Okada, N., Kasai, K., Okanoya, K., Yamashita, O., Tanaka, S.C., Koike, S., 2021. Comparison of traveling–subject and ComBat harmonization methods for assessing structural brain characteristics. Hum. Brain Mapp. 42, 5278–5287.
Melzer, T.R., Keenan, R.J., Leeper, G.J., Kingston-Smith, S., Felton, S.A., Green, S.K., Henderson, K.J., Palmer, N.J., Shoorangiz, R., Almuqbel, M.M., Myall, D.J., 2020. Test-retest reliability and sample size estimates after MRI scanner relocation. Neuroimage 211, 116608.
Moyer, D., Steeg, G.V., Tax, C.M.W., Thompson, P.M., 2020. Scanner invariant representations for diffusion MRI harmonization. Magn. Reson. Med. 84, 2174–2189.
Noble, S., Scheinost, D., Finn, E.S., Shen, X., Papademetris, X., McEwen, S.C., Bearden, C.E., Addington, J., Goodyear, B., Cadenhead, K.S., Mirzakhanian, H., Cornblatt, B.A., Olvet, D.M., Mathalon, D.H., McGlashan, T.H., Perkins, D.O., Belger, A., Seidman, L.J., Thermenos, H., Tsuang, M.T., van Erp, T.G.M., Walker, E.F., Hamann, S., Woods, S.W., Cannon, T.D., Constable, R.T., 2017a. Multisite reliability of MR-based functional connectivity. Neuroimage 146, 959–970.
Noble, S., Scheinost, D., Finn, E.S., Shen, X.L., Papademetris, X., McEwen, S.C., Bearden, C.E., Addington, J., Goodyear, B., Cadenhead, K.S., Mirzakhanian, H., Cornblatt, B.A., Olvet, D.M., Mathalon, D.H., McGlashan, T.H., Perkins, D.O., Belger, A., Seidman, L.J., Thermenos, H., Tsuang, M.T., van Erp, T.G.M., Walker, E.F., Hamann, S., Woods, S.W., Cannon, T.D., Constable, R.T., 2017b. Multisite reliability of MR-based functional connectivity. Neuroimage 146, 959–970.
Park, H.J., Friston, K., 2013. Structural and functional brain networks: from connections to cognition. Science 342, 1238411.

Poldrack, R.A., Gorgolewski, K.J., 2014. Making big data open: data sharing in neuroimaging. Nat. Neurosci. 17, 1510–1517.

Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., Bashyam, V., Nasrallah, I.M., Satterthwaite, T.D., Fan, Y., Launer, L.J., Masters, C.L., Maruff, P., Zhuo, C.J., Volzke, H., Johnson, S.C., Fripp, J., Koutsouleris, N., Wolf, D.H., Gur, R., Gur, R., Morris, J., Albert, M.S., Grabe, H.J., Resnick, S.M., Bryan, R.N., Wolk, D.A., Shinohara, R.T., Shou, H.C., Davatzikos, C., 2020. Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. Neuroimage 208, 116450.

Radua, J., Vieta, E., Shinohara, R., Kochunov, P., Quide, Y., Green, M.J., Weickert, C.S., Weickert, T., Bruggemann, J., Kircher, T., Nenadic, I., Cairns, M.J., Seal, M., Schall, U., Henskens, F., Fullerton, J.M., Mowry, B., Pantelis, C., Lenroot, R., Cropley, V., Loughland, C., Scott, R., Wolf, D., Satterthwaite, T.D., Tan, Y., Sim, K., Piras, F., Spalletta, G., Banaj, N., Pomarol-Clotet, E., Solanes, A., Albajes-Eizagirre, A., Canales-Rodriguez, E.J., Sarro, S., Di Giorgio, A., Bertolino, A., Stablein, M., Oertel, V., Knochel, C., Borgwardt, S., du Plessis, S., Yun, J.Y., Kwon, J.S., Dannlowski, U., Hahn, T., Grotegerd, D., Alloza, C., Arango, C., Janssen, J., Diaz-Caneja, C., Jiang, W., Calhoun, V., Ehrlich, S., Yang, K., Cascella, N.G., Takayanagi, Y., Sawa, A., Tomyshev, A., Lebedeva, I., Kaleda, V., Kirschner, M., Hoschl, C., Tomecek, D., Skoch, A., van Amelsvoort, T., Bakker, G., James, A., Preda, A., Weideman, A., Stein, D.J., Howells, F., Uhlmann, A., Temmingh, H., Lopez-Jaramillo, C., Diaz-Zuluaga, A., Fortea, L., Martinez-Heras, E., Solana, E., Llufriu, S., Jahanshad, N., Thompson, P., Turner, J., van Erp, T.collaborators, E.C., 2020. Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA. Neuroimage 218, 116956.

Rao, A., Monteiro, J.M., Mourao-Miranda, J., Initiative, A.D., 2017. Predictive modelling using neuroimaging data in the presence of confounds. Neuroimage 150, 23–49.

Schumann, G., Loth, E., Banaschewski, T., Barbot, A., Barker, G., Buchel, C., Conrod, P.J., Dalley, J.W., Flor, H., Gallinat, J., Garavan, H., Heinz, A., Itterman, B., Lathrop, M., Mallik, C., Mann, K., Martinot, J.L., Paus, T., Poline, J.B., Robbins, T.W., Rietschel, M., Reed, L., Smolka, M., Spanagel, R., Speiser, C., Stephens, D.N., Strohle, A., Struve, M., Consortium, I., 2010. The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology. Mol Psychiatr 15, 1128–1139.

Smallwood, R.F., Laird, A.R., Ramage, A.E., Parkinson, A.L., Lewis, J., Clauw, D.J., Williams, D.A., Schmidt-Wilcke, T., Farrell, M.J., Eickhoff, S.B., Robin, D.A., 2013. Structural brain anomalies and chronic pain: a quantitative meta-analysis of gray matter volume. J. Pain 14, 663–675.

Tanaka, S.C., Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., Kunimatsu, A., Okada, N., Hashimoto, R., Okada, G., Sakai, Y., Morimoto, J., Narumoto, J., Shimada, Y., Mano, H., Yoshida, W., Seymour, B., Shimizu, T., Hosomi, K., Saitoh, Y., Kasai, K., Kato, N., Takahashi, H., Okamoto, Y., Yamashita, O., Kawato, M., Imamizu, H., 2021. A multi-site, multi-disorder resting-state magnetic resonance image database. Sci Data 8, 227.

Tong, Q., Gong, T., He, H., Wang, Z., Yu, W., Zhang, J., Zhai, L., Cui, H., Meng, X., Tax, C.W.M., Zhong, J., 2020. A deep learning-based method for improving reliability of multicenter diffusion kurtosis imaging with varied acquisition protocols. Magn. Reson. Imaging 73, 31–44.

Tong, Q.Q., He, H.J., Gong, T., Li, C., Liang, P.P., Qian, T.Y., Sun, Y., Ding, Q.P., Lie, K.C., Zhong, J.H., 2019. Reproducibility of multi-shell diffusion tractography on traveling subjects: A multicenter study prospective. Magn. Reson. Imaging 59, 1–9.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15, 273–289.

van den Heuvel, M.P., Sporns, O., 2019. A cross-disorder connectome landscape of brain dysconnectivity. Nat. Rev. Neurosci. 20, 435–446.

Vovk, U., Pernus, F., Likar, B., 2007. A review of methods for correction of intensity inhomogeneity in MRI. Ieee T Med Imaging 26, 405–421.

Xia, M., He, Y., 2017. Functional connectomics from a "big data" perspective. Neuroimage 160, 152–167.

Xia, M.R., Si, T.M., Sun, X.Y., Ma, Q., Liu, B.S., Wang, L., Meng, J., Chang, M., Huang, X.Q., Cheni, Z.Q., Tang, Y.Q., Xu, K., Gong, Q.Y., Wang, F., Qiu, J., Xie, P., Li, L.J., He, Y., Wor, D.M.D.D., 2019. Reproducibility of functional brain alterations in major depressive disorder: Evidence from a multisite resting-state functional MRI study with 1,434 individuals. Neuroimage 189, 700–714.

Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., Kunimatsu, A., Okada, N., Yamagata, H., Matsuo, K., Hashimoto, R., Okada, G., Sakai, Y., Morimoto, J., Narumoto, J., Shimada, Y., Kasai, K., Kato, N., Takahashi, H., Okamoto, Y., Tanaka, S.C., Kawato, M., Yamashita, O., Imamizu, H., 2019. Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. PLoS Biol. 17, e3000042.

Yu, M.C., Linn, K.A., Cook, P.A., Phillips, M.L., McInnis, M., Fava, M., Trivedi, M.H., Weissman, M.M., Shinohara, R.T., Sheline, Y.I., 2018. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. Hum. Brain Mapp. 39, 4213–4227.

Zhao, F.Q., Wu, Z.W., Wang, L., Lin, W.L., Xia, S.R., Shen, D.G., Li, G.the UNC/UMN Baby Connectome Project Consortium, 2019. Harmonization of infant cortical thickness using surface-to-surface cycle-consistent adversarial networks. International Conference on Medical Image Computing and Computer-Assisted Intervention 11767 , pp. 475–483.

Zuo, L.R., Dewey, B.E., Liu, Y.H., He, Y.F., Newsome, S.D., Mowry, E.M., Resnick, S.M., Prince, J.L., Carass, A., 2021. Unsupervised MR harmonization by learning disentangled representations using information bottleneck theory. Neuroimage 243, 118569.