Harmonization of Diffusion MRI-based Structural Connectomes for Multi-site Studies

Rui Sherry Shen[1,2*], Drew Parker[2], Andrew An Chen[3], Benjamin E. Yerys[4,5,6], Birkan Tunç[4,5], Russell T. Shinohara[7], Ragini Verma[1,2*]

[1]*Department of Bioengineering, University of Pennsylvania, Philadelphia, United States*
[2]*Perelman School of Medicine, University of Pennsylvania, Philadelphia, United States*
[3]*Department of Public Health Sciences, Medical University of South Carolina, Charleston, United States*
[4]*Center for Autism Research, Children's Hospital of Philadelphia, Philadelphia, United States*
[5]*Department of Child and Adolescent Psychiatry and Behavioral Science, Children's Hospital of Philadelphia, Philadelphia, United States*
[6]*Advancing Transition and Learning for Adult Success Center, Children's Hospital of Philadelphia, Philadelphia, United States*
[7]*Penn Statistics in Imaging and Visualization Center, Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, United States*

Abstract

Diffusion MRI-based structural connectomes are increasingly used to investigate brain connectivity in various disorders. However, small sample sizes in individual studies, along with the highly heterogeneous and complex disease-related effects on structural connectome data, underscore the need to pool datasets across multiple sources to identify meaningful patterns and establish reproducible imaging biomarkers. Despite ongoing collaborative efforts in multi-site data collection and integration, site-related differences—often arising from scanner and acquisition variability—can introduce biases into pooled datasets when developing structural connectome-based markers. This highlights the need for statistical data harmonization to mitigate site-related effects in structural connectomes while preserving the biological information of demographics and disorders. Our work represents the first effort to develop and evaluate harmonization frameworks specifically designed for structural connectomes. We conduct a thorough investigation of various statistical harmonization methods, including ComBat, CovBat, and GLM, adapting them to accommodate the unique distributional characteristics of structural connectomes. Through rigorous evaluation, we demonstrate that the gamma-GLM framework outperforms others in modeling structural connectomes, effectively removing site-related biases in both edge-based and downstream graph-based analyses while preserving biological variability. Two real-world applications further underscore the utility of our harmonization framework in addressing challenges in multi-site structural connectome analysis. Specifically, harmonization with gamma-GLM enhances the transferability of connectome-based brain age predictors to new datasets and improves statistical power via data integration in multi-site studies examining group-level differences. Our work provides essential guidelines for harmonization of multi-site studies on structural connectomes, paving the way for more robust discoveries through collaborative research in the era of team science and big data.
Keywords: Strucutral connectome, Harmonization, Diffusion MRI, Gamma Generalized Linear Model, ComBat, CovBat, Multi-site Analysis, Big Data

1. Introduction

Structural connectivity in the brain is often mapped non-invasively through fiber tractography using diffusion-weighted MRI (dMRI) [1] and represented as a connectivity matrix, known as the structural connectome. This approach has significantly advanced the study of various brain disorders, including autism spectrum disorder [2, 3], schizophrenia [4], and traumatic brain injury [5, 6]. However, high inter-individual heterogeneity and typically small sample sizes of single-site studies have often led to divergent findings across research on these disorders [7-11]. In response, the scientific community has initiated efforts to pool data from multiple clinical sites, aiming to provide a more comprehensive representation of the population and uncover consistent findings with larger sample sizes. As part of these efforts, data harmonization across sites is essential to eliminate site-specific variations while preserving the biological and demographic integrity of the data.

Previous studies have observed substantial variability in dMRI data collected across different sites, driven by technical differences in scanner manufacturers, acquisition protocols, and preprocessing pipelines [12-14]. These discrepancies can introduce site-related bias in downstream analyses of structural connectomes, compromising data

* Corresponding author. *E-mail address:* ruishen@seas.upenn.edu , *ragini@pennmedicine.upenn.edu*

comparability in collaborative research and obscuring biologically meaningful associations [15-17]. Various solutions have been proposed to address these issues at different stages of the data collection and analysis pipelines. For instance, efforts to standardize acquisition protocols and parameters have been made to reduce site-related effects [18-20]. However, this approach cannot resolve site differences in retrospective data that has already been collected, and considerable multi-site variations may still arise from residual factors, such as hardware imperfections, operator-specific characteristics, and other uncontrollable variables [18, 21, 22]. There have been attempts to mitigate site-related variations by harmonizing dMRI signals. Mirzaalian et al. [23] introduced a technique that transforms dMRI signals into spherical harmonic space, enabling site-to-site mappings by matching their rotation-invariant spherical harmonic (RISH) features. Subsequently, deep learning methods based on RISH features have also been developed for harmonization of dMRI signals [24, 25]. These RISH-based harmonization approaches have proven effective in mitigating site-related effects in dMRI signals from similar acquisitions. Another approach, proposed by Huynh et al. [26], employs the method of moments (MoM) to directly harmonize dMRI signals without transforming them into other domains, offering greater flexibility for integrating dMRI data with varying numbers of gradient directions.

Despite these advances, most existing approaches address site effects only at the dMRI signal level, assuming that downstream outcomes, such as structural connectomes, will be harmonized accordingly. However, Kurokawa et al. [27], using a traveling-subject design, demonstrated that site-related effects in structural connectomes and their graph metrics can persist even after dMRI signal harmonization. Moreover, different tractography algorithms and connectome construction methods are often adopted in data processing pipelines, particularly for retrospective datasets constrained by acquisition limitations, which can introduce further site-related biases after harmonizing dMRI signals. This emphasizes the need for more effective handling of inter-site variability at the connectivity matrix level to adequately mitigate site-related biases in structural connectomes before conducting graph analyses or other investigations in multi-site studies.

Structural connectomes represent complex brain connectivity that is high-dimensional and graph-structured, with connectivity information following a non-Gaussian distribution. As a result, connectome harmonization requires specialized methods to address this complexity. While various statistical models have been proposed for harmonizing neuroimaging-based measures, their applicability to structural connectomes remains unexplored. ComBat, originally designed to address 'batch effects' in genomic data [28], uses parametric empirical Bayes [29] to adjust for mean and variance of site-related effects, while accounting for biological covariates such as age and sex. Fortin et al. [30] extended ComBat to harmonize dMRI scalar maps such as fractional anisotropy (FA) and mean diffusivity (MD), and Yu et al. [31] applied ComBat to resting-state functional connectivity. In addition to ComBat, other harmonization methods such as the generalized linear model (GLM) [32, 33] and CovBat [34] have been widely used. However, most of these methods are limited to harmonizing measures that follow normal distribution. To date, no harmonization tools have been specifically designed and validated for structural connectomes.

In this work, we address this gap by proposing a harmonization framework for structural connectomes. We adapt several statistical harmonization models to accommodate connectome data and conduct a comprehensive evaluation of their effectiveness in eliminating site-related differences while preserving connectivity information and underlying biology associated with factors such as age and disorders. To the best of our knowledge, this is the first effort to develop harmonization frameworks specifically tailored to structural connectomes. Furthermore, we present two practical use cases that demonstrate how harmonization approaches can overcome common challenges in multi-site studies. In the first case, we assess the ability of harmonization methods to improve the transferability of machine learning models to new datasets, particularly in the context of brain age prediction. In the second case, recognizing that most connectome studies include both control and patient groups, we examine the effectiveness of harmonization methods in multi-site data integration to enhance statistical power for detecting clinical group differences. Finally, we provide general guidelines for the harmonization of multi-site structural connectomes, facilitating data integration to yield robust, replicable and generalizable findings.

2. Methods

2.1. Overview of Datasets

We will use structural connectome data pooled from six datasets, combining them in various data configurations to assess different harmonization approaches. These datasets include the Philadelphia Neurodevelopmental Cohort (PNC) [35], the Center for Autism Research (CAR) [36], TimRobert [37], and three clinical sites from Autism Brain Imaging Data Exchange II (ABIDEII-NYU, ABIDEII-SDSU, ABIDEII-TCD) [3]. Our study includes a total of 1503 participants (890 males, 613 females), comprising 1194 typically developing children (TDC) and 309 children with autism spectrum disorders (ASD). The detailed demographic information and acquisition parameters of each dataset are described in Table 1 .

Table 1 Demographic information and acquisition parameters for six datasets.

Dataset	TDC			ASD			Scanner	dMRI					T1-w MP-RAGE				
	Age range (years)	Sex (M, F)	Sample size	Age range (years)	Sex (M, F)	Sample size		b values (s/mm^2)	Gradient directions	TR (ms)	TE (ms)	Resolution (mm)	TR (ms)	TE (ms)	TI (ms)	Flip angle	Resolution (mm)
PNC	8.17-21.00 (15.06±3.33)	M=421 F=533	954	-	-	-	Siemen Trio 3T	0, 1000	64	8100	82	1.875× 1.875 × 2	1810	3.5	1100	9°	0.9375 × 0.9375 × 1
CAR	6.26-17.90 (11.81±2.85)	M=112 F=30	142	6.37-18.82 (12.05±2.73)	M=129 F=22	151	Siemens Verio 3T	0, 1000	30	11000	76	2 × 2 × 2	1900	2.54	900	9°	0.8 × 0.8 × 0.9
TimRobert	6.17-13.83 (9.97±2.04)	M=32 F=6	38	6.17-16.92 (10.65±2.18)	M=61 F=9	70	Siemens Verio 3T	0, 1000	30	11000	76.4	2 × 2 × 2	1900	2.87	1050	9°	1 × 1 × 1
ABIDEII-NYU	6.66-12.90 (9.55±1.77)	M=19 F=0	19	6.05-17.93 (8.33±2.17)	M=35 F=4	39	Siemens Allegra	0, 1000	64	5200	78	3 × 3 × 3	2530	3.25	1100	7°	1.3 × 1 × 1.33
ABIDEII-SDSU	8.10-17.70 (13.50±3.05)	M=21 F=2	23	7.40-18.00 (12.95±3.35)	M=24 F=7	31	GE MR750	0, 1000	61	8500	84.9	1.875 × 1.875 × 2	8.136	3.172	600	8°	1 × 1 × 1
ABIDEII-TCD	10.25-20.00 (16.18±2.97)	M=18 F=0	18	10.00-19.50 (14.47±3.32)	M=18 F=0	18	Philips Achieva	0, 1500	61	20244	79	1.94 × 1.94 × 2	8.4	3.9	1150	8°	0.9 × 0.9 × 0.9

[1]Mean and standard deviation of age are shown in parentheses.

[2]Abbreviations: M, male; F, female, TR, repetition time; TE, echo time; TI, inversion time

In order to evaluate the effectiveness of harmonization methods in different scenarios, we create four data configurations by combining the cohorts in various ways, as detailed below. The overall experiment design and the usage of different data configurations are visualized in Figure 1.

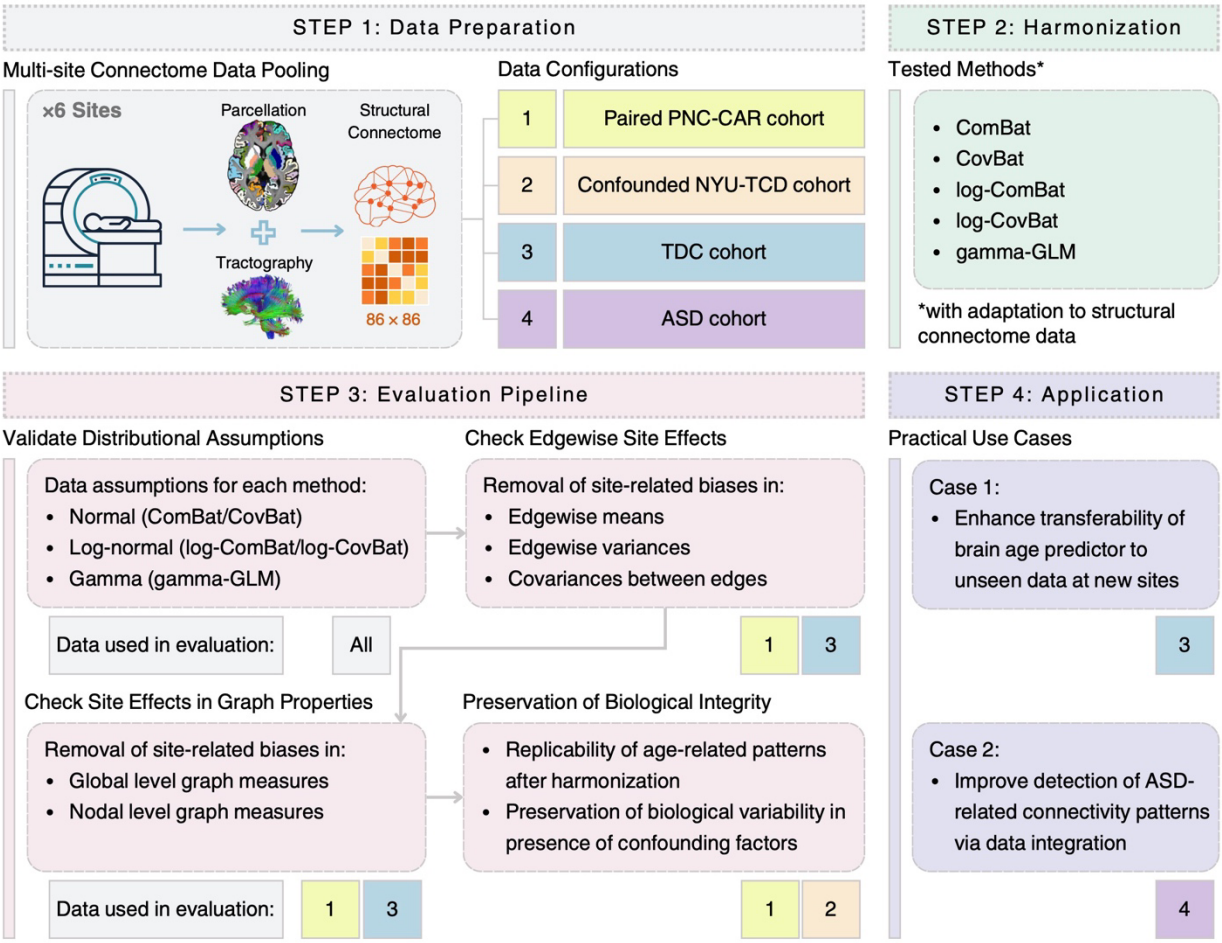

Figure 1 After pooling MRI data from six sites and extracting structural connectomes, we create four distinct data configurations by combining the cohorts in various ways (Step 1). We then apply different harmonization methods to each data configuration (Step 2). Next, we evaluate these harmonization methods from four perspectives: validation of distributional assumptions, removal of edgewise site effects, mitigation of site effects on graph properties, and preservation of biological integrity (Step 3). Finally, we demonstrate the practical utility of structural connectome harmonization methods through two use cases, highlighting their contributions to multi-site studies (Step 4). The data configurations used in each evaluation step are indicated by their corresponding numbers.

2.1.1. Data Configuration 1: Paired PNC-CAR cohort

We carefully select typically developing children from our two largest datasets, the PNC and CAR sites, ensuring they are closely matched in age and sex. This approach aims to mimic the traveling-subject research paradigm, minimizing confounding effects that could impact the harmonization results. As a result, we form two groups from different sites, each consisting of 100 participants (76 males, 24 females). The age of participants from the PNC site ranges from 8.17 to 18.08 years, with a mean of 12.31 years and a standard deviation of 2.70 years at enrollment. Participants from the CAR site have an age range from 8.08 to 17.83 years, with a mean of 12.27 years and a standard deviation of 2.73 years.

2.1.2. Data Configuration 2: Confounded NYU-TCD cohort

We use all typically developing children from the ABIDEII-NYU and ABIDEII-TCD datasets to create a cohort where age and site are confounded. The ABIDEII-NYU dataset includes the youngest participants among the six datasets, with an age range of 6.66 to 12.90 years, a mean of 9.55 years, and a standard deviation of 1.77 years. In contrast, the ABIDEII-TCD dataset consists of the oldest participants, with an age range of 10.25 to 20.00 years, a

mean of 16.18 years, and a standard deviation of 2.97 years. As this cohort only have male participants, no sex-related biases are present.

2.1.3. Data Configuration 3: TDC cohort

We pool all TDC data from our datasets, resulting in a total of 1194 participants (623 males, 571 females). The age of participants ranges from 6.17 to 21.00 years, with a mean age of 14.41 years and a standard deviation of 3.53 years. This cohort represents a more generic case in which age and sex are unbalanced with respect to site, potentially introducing statistical confounding.

2.1.4. Data Configuration 4: ASD cohort

All datasets, except for the PNC site, also contain neuroimaging data from children with ASD. We pool the ASD data from these five datasets to create a cohort of 309 participants (267 males, 42 females) aged from 6.05 to 19.50 years, with a mean age of 11.49 years and a standard deviation of 3.05 years. The structural connectomes in the ASD cohort will be harmonized by adjusting for the site effects estimated from the corresponding TDC cohort.

2.2. Data Processing

2.2.1. Pre-processing and Quality Control

We conduct a series of preprocessing steps for dMRI data, including denoising using joint local principal component analysis (PCA), correction for eddy currents and movements, and skull stripping with BET2. The constrained spherical deconvolution (CSD) of dMRI is then conducted using DIPY, with the fiber response function estimated using dhollander algorithm [38, 39] and the fiber orientation distribution (FOD) fitted using the method proposed by Jeurissen and others [40]. We perform probabilistic fiber tracking using the iFOD2 algorithm and anatomically constrained tractography (ACT) [41], with 500 seeds placed at random inside each voxel of the grey-matter white-matter interface (GMWMI). We preprocess the high-resolution T1-w images with FreeSurfer recon-all pipeline (http://surfer.nmr.mgh.harvard.edu) [42], followed by registration to the dMRI data of each subject using restricted deformable SyN algorithm in ANTs [43]. We conduct manual quality control to exclude data with poor acquisition quality or failed registration.

2.2.2. Construction of Structural Connectomes

We parcellate the brain into 86 regions of interest (ROIs), including 68 cortical regions from the Desikan-Killiany atlas [44] and 18 subcortical regions from the FreeSurfer's atlas [45]. We calculate the number of streamlines intersecting each pairwise combination of these ROIs, yielding an 86x86 connectivity matrix. Post-processing of the structural connectomes involves removing self-connections and symmetrizing the connectivity matrix. Additionally, we normalize the structural connectomes by dividing edge weights by each subject's total WMGMI volume. Since structural connectomes derived from probabilistic tractography may include numerous spurious connections that could affect downstream analysis, we apply consistency-based thresholding to filter out spurious connections and retain only the most consistent ones. For our primary analyses, the consistency-based thresholding at 40% density has been applied to retain edges that are highly consistent across the population. All harmonization methods are performed with sex, age, and age^2 included as covariates.

2.3. Harmonization Methods

We investigate five statistical approaches, adapting them for the harmonization of structural connectomes. Table 2 compares these harmonization methods in terms of their capabilities, as also demonstrated in previous literature [28, 34, 46]. In the following sections, we provide detailed descriptions of the adaptation of each technique to connectomes.

Table 2 Comparison of mainstream harmonization methodologies.

Harmonization method	Addresses mean site effects	Addresses variance site effects	Addresses covariance site effects	Address skewness	Assure non-negative edge values	Assumption for data distribution
ComBat	Yes	Yes				Normal
CovBat	Yes	Yes	Yes			Normal

log-ComBat	Yes	Yes		Yes	Yes	Log-normal
log-CovBat	Yes	Yes	Yes	Yes	Yes	Log-normal
gamma-GLM	Yes			Yes	Yes	Gamma

For each method, we assume that structural connectomes have been collected from M imaging sites, where $m = 1, 2, \dots, M$ denotes the index of the site, and each site comprises N_m samples, with $n = 1, 2, \dots, N_m$ representing the index of the subject. Let Y_{mne} represent the observed edge value before harmonization at a given edge index $e \in E(G)$ within the structural connectome of subject n from site m.

2.3.1. Adaptation of ComBat to Connectomes

The ComBat model was originally developed by Johnson et al. [28] to mitigate batch effects when integrating multiple microarray datasets for gene expression analysis. We extend ComBat to harmonize structural connectomes by formulating each edge value as follows:

$$Y_{mne} = \alpha_e + \boldsymbol{X_{mn}^T \beta_e} + \gamma_{me} + \delta_{me}\epsilon_{mne}$$

In this model, α_e represents intercept of the connectivity value at edge e, the vector $\boldsymbol{X_{mn}}$ incorporates covariates of interest, such as age and sex, with $\boldsymbol{\beta_e}$ representing the vector of regression coefficients corresponding to $\boldsymbol{X_{mn}}$ at edge e. γ_{me} and δ_{me} denote the mean and variance of the site effect for site m at edge e. Combat assumes that the error term ϵ_{mne} follows an independent normal distribution $\epsilon_{mne} \sim N(0, \sigma_e^2)$. It first estimates $\hat{\alpha}_e, \hat{\beta}_e$ using ordinary least-squares by constraining $\Sigma_m N_m \hat{\gamma}_{me} = 0$ for all edges. To estimate empirical statistical distributions of γ_{me} and δ_{me}, ComBat employs an empirical Bayes approach with the assumption that all edges share a common yet site-specific prior distribution, with hyperparameters estimated empirically using the method of moments. This allows information from all edges to be leveraged to infer the statistical properties of the site effect. Specifically, γ_{me} is assumed to have an independent normal prior distribution, and δ_{me} follows an independent inverse gamma prior distribution. The estimates γ_{me}^* and δ_{me}^* are obtained by computing the conditional posterior means. Finally, ComBat adjusts for the estimated site effects to obtain the harmonized edge:

$$ComBat(Y_{mne}) = \frac{Y_{mne} - \hat{\alpha}_e - \boldsymbol{X_{mn}^T \widehat{\beta}_e} - \gamma_{me}^*}{\delta_{me}^*} + \hat{\alpha}_e + \boldsymbol{X_{mn}^T \widehat{\beta}_e}$$

ComBat harmonization may introduce spurious connections for those originally with 0 edge values (indicating no connection between two nodes), altering the physical interpretation of the connectivity. To correct this, we follow the approach of Onicas et al. [47] by reassigning zeros to those connections. Specifically, for each participant, we apply a binary connectivity matrix from the pre-harmonized data to mask the harmonized structural connectomes.

2.3.2. Adaptation of CovBat to Connectomes

ComBat treats each edge in the structural connectomes independently, without accounting for the covariance between edges. However, in structural connectomes, site effects might be correlated edgewise and exhibit different covariances across sites. To address these potential covariance effects, Chen et al. [34] adapted the ComBat framework into CovBat. CovBat posits that the error vector $\boldsymbol{\epsilon_{mn}} = (\epsilon_{mn1}, \epsilon_{mn2}, \dots, \epsilon_{mne}, \dots, \epsilon_{mnE})^T \sim N(\boldsymbol{0}, \boldsymbol{\Sigma_m})$, where E denotes total number of edges. CovBat initially applies ComBat to remove the mean and variance of the sites and obtain ComBat-adjusted residuals:

$$\epsilon_{mne}^{ComBat} = \frac{Y_{mne} - \hat{\alpha}_e - \boldsymbol{X_{mn}^T \widehat{\beta}_e} - \gamma_{me}^*}{\delta_{me}^*}$$

It then performs principal component analysis (PCA) on the residuals. These residuals can be projected onto the principal components, expressed as:

$$\boldsymbol{\epsilon_{mn}^{ComBat}} = \sum_{l=1}^{L} \xi_{mnl} \boldsymbol{\Phi_l}$$

Here, $\boldsymbol{\Phi_l}$ represent the eigenvectors of the covariance of the residual data, and $L = \min\left(\sum_{m=1}^{M} N_m,\ E\right)$ denotes the rank. ξ_{mnl} indicates the principal component scores. Assuming that the covariance site effect is captured within these principal component scores, we can harmonize them to mitigate the covariance site effects. Suppose:

$$\xi_{mnl} = \mu_{ml} + \rho_{ml}\epsilon_{mnl}$$

Where μ_{ml} and ρ_{ml} are location and scale shifts on principal component scores for site m at principal component l. The CovBat adjusts these site-related shifts in principal component scores:

$$\xi_{mnl}^{CovBat} = (\xi_{mnl} - \widehat{\mu_{ml}})/\widehat{\rho_{ml}}$$

The covariance-adjusted residuals can be obtained by projecting the adjusted principal component scores back into the original space, with a hyperparameter K controlling the desired portion of principal component scores to be adjusted:

$$\boldsymbol{e_{mn}^{CovBat}} = \sum_{k=1}^{K} \xi_{mnk}^{CovBat} \boldsymbol{\Phi_k} + \sum_{l=K+1}^{L} \xi_{mnl}^{CovBat} \boldsymbol{\Phi_l}$$

In this study, we select K such that the principal components explain 95% of the variation. The final harmonized edge values can be represented by replacing the residuals:

$$CovBat(Y_{mne}) = e_{mne}^{CovBat} + \hat{\alpha}_e + \boldsymbol{X_{mn}^T}\widehat{\boldsymbol{\beta}}_{\boldsymbol{e}}$$

Similar to the post-processing in ComBat, we reassign zeros to connections that had zero edge values prior to harmonization.

2.3.3. Adaptation of log-ComBat to Connectomes

The ComBat algorithm relies on the assumption of data conforming to a normal distribution. However, this premise may not hold true for structural connectome data, where edge values frequently demonstrate notable skewness. In biomedical research, the log-scaling technique is commonly employed to transform skewed data to approximate a normal distribution [48, 49]. The log-scaling approach assumes that edgewise connectome strength follows a log-normal distribution or closely approximate it. Consequently, the ComBat algorithm can be reformulated as follows:

$$\ln(Y_{mne}) = \alpha_e + \boldsymbol{X_{mn}^T}\boldsymbol{\beta_e} + \gamma_{me} + \delta_{me}\epsilon_{mne}$$

and the harmonized edge values can be obtained from the inverse of a logarithm:

$$logComBat(Y_{mne}) = \exp\left(\frac{\ln\left(Y_{mne}\right) - \hat{\alpha}_e - \boldsymbol{X_{mn}^T}\widehat{\boldsymbol{\beta}}_{\boldsymbol{e}} - \gamma_{me}^*}{\delta_{me}^*} + \hat{\alpha}_e + \boldsymbol{X_{mn}^T}\widehat{\boldsymbol{\beta}}_{\boldsymbol{e}}\right)$$

log-ComBat ensures that harmonized edges remain non-negative. To handle numerical errors with the logarithm when original edge values Y_{mne} are zero, we add a small positive constant c to Y_{mne}. Those connections with zero edge values prior to harmonization will be reassigned to zeros to ensure no spurious connections are introduced by the harmonization process

2.3.4. Adaptation of log-CovBat to Connectomes

We also adapt the CovBat algorithm by incorporating log-scaling to tackle the skewness of the edge weights in structural connectomes. First, we apply CovBat to log-transformed edge values, and then we apply the inverse of logarithm to obtain the final harmonized edge values:

$$logCovBat(Y_{mne}) = \exp\left(CovBat(\ln(Y_{mne}))\right)$$

log-CovBat ensures that edges remain non-negative after harmonization. Similarly, we add a small positive constant c to original edge value Y_{mne} during log-CovBat harmonization and restore connections that originally had zero edge values back to zeros.

2.3.5. Using gamma-GLM with log-link for Connectomes

The generalized linear model (GLM) provides an alternative for regression when the assumptions of normality and homoskedasticity are not met [50]. By employing a gamma distribution with a log-link function, GLM offers greater flexibility in modeling a wide range of data shapes, particularly well-suited for positive and right-skewed data. We model the effects of site and other covariates on expected structural connectomes with a log-link function, expressed as:

$$\ln\left(E_{mn}[Y_{mne}]\right) = \alpha_e + \boldsymbol{X_{mn}^T \beta_e} + \gamma_{me}$$

In the gamma-GLM, the edge values for each connection are assumed to follow a gamma distribution, $Y_{mne} \sim \Gamma(k_e, \theta_e)$, where k_e and θ_e denote the shape and scale parameters, respectively. The expected value for edge e is given by:

$$E_{mn}[Y_{mne}] = k_e \theta_e$$

The parameters k_e and θ_e can be fitted using GLM regression, enabling the estimation of the mean site effects γ_{me}^* for site m at edge e can be then estimated. To obtain the final harmonized edge values, the estimated mean site effects γ_{me}^* are regressed out, followed by applying the inverse logarithm transformation:

$$gammaGLM(Y_{mne}) = \exp(\ln(Y_{mne}) - \gamma_{me}^*) = Y_{mne}/\exp\left(\gamma_{me}^*\right)$$

The gamma-GLM with a log-link function ensures that the harmonized edge values remain non-negative. As gamma fitting also requires positive edge values, we add a small positive constant c to the original edge value Y_{mne} and restore connections with zero edge values back to zeros after harmonization.

2.4. Validation of Distributional Assumptions on Structural Connectomes

The harmonization methods discussed above are based on three different distributional assumptions for the structural connectome data: the normal distribution (as required by ComBat and CovBat), the log-normal distribution (for log-ComBat and log-CovBat), or the gamma distribution (for gamma-GLM). We conduct goodness-of-fit tests on the structural connectome data to assess the validity of these data assumptions. Specifically, we fit each statistical distribution to the observed structural connectomes at each site. We employ the Kolmogorov–Smirnov (KS) test to quantify the discrepancy between the observed data distribution and the fitted distributions.

2.5. Evaluation of Harmonization Methods

2.5.1. On Removing Site Effects in Edgewise Connectome Analysis

We first examine whether each harmonization method can eliminate site-related biases in the mean, variance, and covariance of edgewise connectome strength. To evaluate mean site effects, we conduct a two-sample t-test to assess edgewise site differences in the paired PNC-CAR cohort (Data Configuration 1) and a one-way analysis of variance (ANOVA) for the TDC cohort (Data Configuration 3, where multiple sites are involved). Sex, age, and age^2 are included as covariates in both analyses. To evaluate variance site effects, we use Brown-Forsythe test on the residuals of edgewise strength, after regressing out sex, age, and age^2. Brown-Forsyth test is the non-parametric version of Levene's test to examine whether the comparing groups have equal variances. For site-specific covariance effects, we first calculate the empirical covariance for each site and then compute the Frobenius distance between the within-site covariance matrices for each pair of sites, both before and after harmonization. The Box-M test is conducted to evaluate whether the site-specific covariance matrices are equal.

2.5.2. On Removing Site Effects in Derived Graph Topological Measures

Several graph topological measures have been derived from structural connectomes using the edgewise connectivity information [51]. Here, we evaluate whether harmonizing edgewise strength using different approaches also mitigates the site effects in the derived graph topological measures. We calculate graph topological measures at two levels, including node-level features (node strength, betweenness centrality, local efficiency, clustering coefficient) and global graph measures (global strength, intra- and inter-hemisphere strength, characteristic path length, global

efficiency, modularity). We then examine site-related differences in these measures before and after applying the adapted harmonization methods. We perform a two-sample t-test to assess site differences for the paired PNC-CAR cohort (Data Configuration 3) and use a one-way ANOVA test for the multi-site TDC cohort (Data Configuration 3). Effect sizes are calculated using Cohen's d for comparisons between two sites and Cohen's f when involving more than two sites.

2.6. Evaluation of Biological Variability Preservation

Ideally, the harmonization method should eliminate non-biological variability introduced by different sites and scanners, while retaining the statistical power to detect biologically meaningful associations. To validate this, we calculate the Pearson correlation between edgewise strength and participants' chronological age in the paired PNC-CAR cohort (Data Configuration 1), controlling for sex covariates. We expect that the harmonized structural connectomes can retain the age-connectivity relationships observed in the pre-harmonization data within each site. The restoration is quantified by the true positive rate (sensitivity) of detected connections exhibiting significant age correlations. Additionally, we rank the connections by the absolute values of their age correlations and compute the overlap between the top k correlations in the harmonized and the original data across all values of k, visualizing the results using concordance at the top (CAT) curves [52] within each site. A CAT curve closer to one indicates better overlap between the two sets of correlation relationships.

2.7. Evaluation of Harmonization with Confounding Factors

Confounding between age and site poses a major challenge for harmonization, as removing site-related variation can inadvertently change age-related variation if not handled carefully. To assess different harmonization methods in the presence of such confounding, we perform a Pearson's correlation analysis, testing the association between edgewise connections and participants' chronological age within each site for the confounded NYU-TCD cohort (Data Configuration 2). CAT plots are used to visualize the replicability of these age associations before and after applying different harmonization methods, and the true positive rate is employed to quantify the recovery of connections with significant age correlations. Additionally, we compute the Frobenius distance between the edgewise age correlation matrices of the two sites. We expect this distance to remain similar before and after harmonization, indicating that age-related differences are preserved between sites. Lastly, we jointly test the replicability of age associations with two sites combined after harmonization, and comparing the results using the paired PNC-CAR cohort (Data Configuration 1) and the confounded NYU-TCD cohort (Data Configuration 2). We also include the scenario where the two sites are combined without proper harmonization to highlight the critical role of harmonization in tackling confounding factors during data integration.

2.8. Practical Use Cases for Multi-site Connectome Harmonization

To further demonstrate the importance of harmonization in multi-site structural connectome analysis, we assess the effectiveness of harmonization frameworks through two practical use cases: 1) improving the transferability of prediction models to new datasets. We use brain age prediction as an example. 2) enhancing statistical power for clinical studies through data integration. Here we focus on detecting group differences and clinical associations in ASD.

2.8.1. Brain age prediction on new datasets

Brain development is accompanied by structural changes in brain organizations, and the gap between predicted brain age from neuroimaging data and chronological age has been linked to various health conditions [53-57]. While many researchers have used machine learning to model neurotypical developmental trajectories to predict brain age, these predictors often struggle with transferability across datasets due to site and scanner variability. We use brain age prediction as a case study to assess whether harmonization methods can eliminate site-related effects while preserving age-related biological variability in structural connectomes, thereby improving the prediction performance of machine learning models on new datasets. We use the TDC cohort (Data Configuration 3) for this evaluation.

We train a support vector regression (SVR) model to predict brain age from structural connectomes, using the TDC data from the PNC site. Features are extracted from the upper triangular portion of the connectivity matrices, with PCA applied for dimensionality reduction. We randomly split the PNC dataset into 50:50 for the training set and the testing set. To optimize the model, we apply 5-fold cross-validation on the training set to tune hyperparameters of the model. The final SVR model is trained on all subjects in the training set using the optimal hyperparameters. The performance of the model is reported using the testing set of the PNC dataset. We then assess the model's transferability by testing it on unseen data from the remaining sites, evaluating prediction performance through root mean square error (RMSE), mean absolute error (MAE), and Pearson correlation coefficient (R).

2.8.2. Enhancing statistical power for ASD studies

Previous research has identified several structural connectome findings in individuals with ASD [2, 58-61]. However, most studies are conducted at the single site and often suffer from small sample sizes, limiting their statistical power to draw generalizable conclusions. Pooling data from multiple sources can introduce site-related biases, which may obscure the detection of clinical associations. Therefore, harmonizing structural connectomes is crucial for revealing reliable ASD-related characteristics in multi-site datasets. To validate the effectiveness of harmonization algorithms, we apply site-specific coefficients estimated from the TDC cohort (Data Configuration 3) to harmonize the ASD cohort (Data Configuration 4). We calculate global graph topological measures of structural connectomes before and after harmonization, assessing group level differences as well as group-by-age interactions in these measures between the ASD and TDC groups. We compare harmonized structural connectomes to unharmonized data where site covariate is regressed out at the level of graph measures. Due to pronounced sex differences in autistic children and the small sample size of autistic females, we exclude female subjects from this analysis to minimize potential sex bias. We further conduct Spearman correlations between graph measures and autism clinical assessments, including Autism Diagnostic Observation Schedule (ADOS) calibrated severity scores, Social Responsive Scale (SRS-2), and Vineland Adaptive Behavior Scales (VABS), controlling for age. Multiple comparison is corrected based on False Discovery Rate (FDR). As ComBat and CovBat approaches will produce negative edge values which may lack physical relevance when computing graph topological measures in clinical contexts, we exclude these two methods from the analysis.

3. Results

We organized the results as follows: Section 3.1 examined the distributional properties of structural connectomes before harmonization, validating the data distribution assumptions underlying different harmonization methods. Section 3.2 showed evidence of substantial mean, variance and covariance site effects in edgewise connectivity strength, and evaluated how effectively harmonization methods reduce site variability. Section 3.3 explored the site effects on derived graph topological measures of structural connectomes. Section 3.4 assessed the replicability of biological patterns such as age associations. Section 3.5 tested the robustness of harmonization methods in a more generic context involving confounding factors. Sections 3.6 and 3.7 highlighted two practical applications of harmonization: enhancing the transferability of machine learning-based brain age predictors to new datasets and improving the statistical power of detection for clinical associations.

3.1. Distributional Properties of Connectivity Strength in Connectomes

Each edge of the pre-harmonized structural connectomes was examined to determine whether it follows one of three hypothesized probability distributions: normal, log-normal, or gamma. Figure 2A shows the edgewise KS distances between the observed distributions and three fitted hypothesized distributions in the PNC dataset, the largest site for the TDC group. A larger KS distance indicates a greater deviation of the hypothesized distribution from the observed data. We showed that the gamma distribution demonstrated the best fit, with significantly lower KS distances compared to the other two hypothesized distributions ($p<0.0001$, paired t-test), followed by the log-normal distribution. Figure 2B illustrates the heatmaps of KS distances at each edge, and only edges with significant differences from the observed distributions were shown in the heatmaps ($p<0.05$, FDR-adjusted). Among the top 40% of highly consistent edges (1480 edges in total), 97% of them did not follow a normal distribution, while 31%

of them distributed significantly different from the log-normal distribution, and fewer than 5% edges showed significant differences from the gamma distribution fitting.

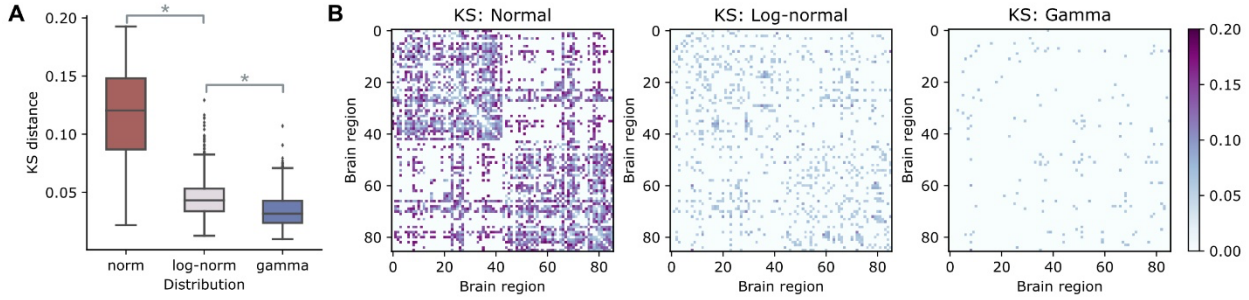

Figure 2 Goodness of fit test for the PNC site. A) Edgewise KS distances between the observed and three hypothesized distributions (normal, log-normal, and gamma) of connectivity strength in the PNC site. Asterisks denote significant differences between those hypothesized distributions (p < 0.0001, paired t-test). The gamma distribution provided a significantly better fit than the other two hypothesized models. B) Heatmaps of edgewise KS distances, showing only edges with significant discrepancy from the observed distributions in pre-harmonized structural connectomes (p < 0.05, FDR-adjusted).

Similar distributional properties of structural connectomes were also observed in the TDC group across the other five sites and held true for the ASD group as well, as shown in Figure S1. Notably, the normality data assumptions underlying ComBat and CovBat were not met at most edges. Instead, structural connectome data were more likely to follow a gamma distribution, regardless of site or diagnostic group.

3.2. Edgewise Evaluation of Site Effects on Structural Connectomes

We utilized the paired PNC-CAR cohort (Data Configuration 1) to assess the site effects on edgewise strength. We created the MA-plot [62], which is a common method for visualizing differences in measurements taken from two sites. In this plot, the edgewise mean connectivity strength was calculated for each site, and the between-site differences in log-transformed means were displayed as a function of the average log-transformed site means. If the structural connectome were free of edgewise site-related effects, the scatterplot would align along the horizontal line at zero, as indicated by the dashed lines. As shown in Figure 3A, we observed striking site differences in edgewise strength between two sites before harmonization. These site differences were largely reduced after applying any of those hypothesized harmonization approaches.

Figure 3B illustrates site-related biases in the mean and variance of edgewise connectivity strength, showing t-statistics for mean site effects (first row) and F* statistics from the Brown-Forsythe test for variance site effects (second row). The log-CovBat and gamma-GLM approaches demonstrated favorable performance in eliminating mean site effects, while log-ComBat and log-CovBat were particularly effective in harmonizing variance site effects. Specifically, among the 1480 edges after consistency thresholding, 1035 (70%) edges initially showed significant differences in site means before harmonization ($p<0.05$, two-sample t-test). Following ComBat harmonization, this number dropped to just 11 edges (0.74%), and logComBat reduced it to 5 edges (0.34%). After CovBat harmonization, only 3 edges (0.2%) showed significant site differences in means, and no significant site effects were observed with log-CovBat and gamma-GLM harmonization. For variance site effects, 741 edges (50%) exhibited significant site differences in pre-harmonization data ($p<0.05$, Brown-Forsythe test). After harmonization, 110 edges (7.4%) still showed significant site effects with ComBat, 104 edges (7%) remained significant with CovBat, and 7 edges (0.47%) with gamma-GLM. With log-ComBat and log-CovBat harmonization, all connections were free from site-related effects on variance.

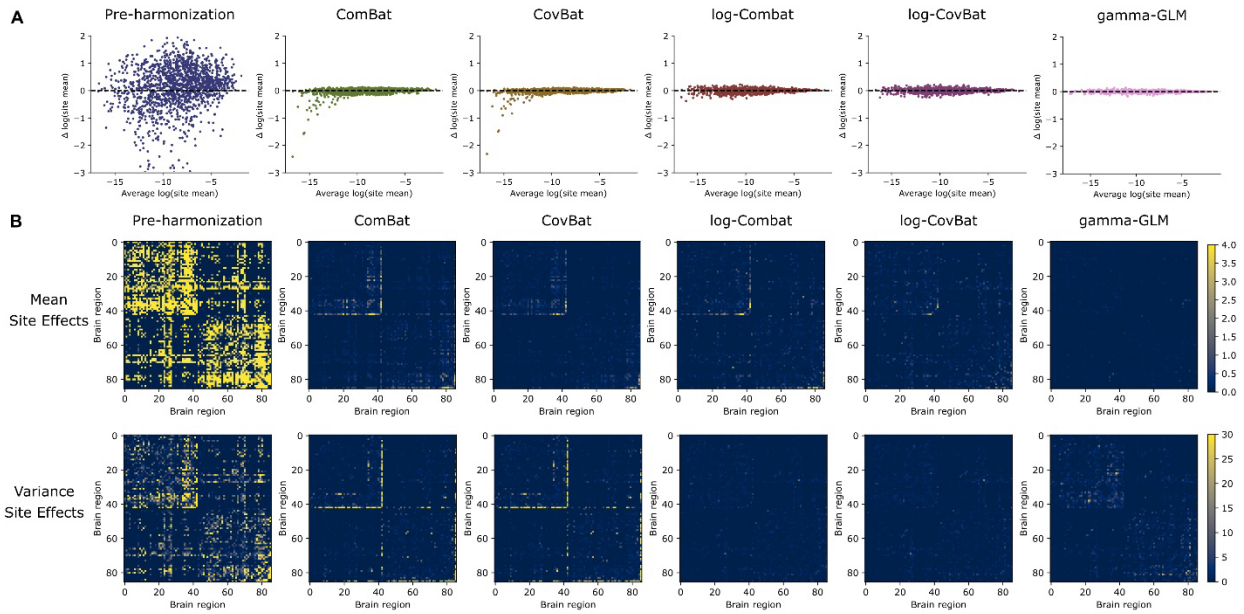

Figure 3 Site-related effects on mean and variance of edgewise strength in the paired PNC-CAR cohort (Data Configuration 1). A) MA-plots for visualization of site differences between the PNC site and the CAR site with paired subjects. The x-axis represents the averaged log-transformed means across sites and the y-axis represents between-site differences in log-transformed means. The horizontal line at zero indicates no site-related effects. B) Edgewise site effects on mean (first row) and variance (second row) connectivity strength, showing the t-statistics for mean site effects and F statistics from the Brown-Forsythe test for variance site effects.*

Table 3 shows the site-specific covariance effects on edgewise strength before and after harmonization, quantified by the Frobenius distance between covariance matrices of two sites. We regressed out the demographic factors, including sex, age, and age², jointly across both sites from edgewise strength using a linear model. The log-CovBat harmonization approach most effectively mitigated site effects on covariance in the paired PNC-CAR cohort (Data Configuration 1), followed by log-ComBat and gamma-GLM. However, results from the Box's M test indicated that covariance site effects on structural connectomes were not significant across all tested cases, including the pre-harmonization data.

Table 3 Frobenius distances between site-specific covariance matrices

Pairwise Site differences	Pre-harmonization	ComBat	CovBat	log-ComBat	log-CovBat	gamma-GLM
Paired PNC-CAR	0.1022	0.0924	0.0912	0.0728	0.0726	0.0786

3.3. Evaluation of Site Effects on Derived Graph Topological Measures

We examined the derived graph topological measures of structural connectomes to determine whether harmonizing edgewise connectivity strength can lead to harmonized downstream outcomes. Figure 4 illustrates the Cohen's d effect sizes of site differences across six global graph topological measures, with asterisks denoting significant site effects. All six global measures showed significant site effects before harmonization ($p<0.05$, two-sample t-test), with three strength-based graph measures exhibiting large effect sizes (Cohen's $d > 1.0$), characteristic path length displaying a moderate effect size (Cohen's $d=0.68$), and global efficiency showing a smaller effect size (Cohen's $d=0.30$). After ComBat harmonization of edgewise connections, site effects on four global measures were eliminated. However, site effects on characteristic path length and global efficiency remained significant ($p<0.05$, two-sample t-test), though their effect sizes were reduced to small (Cohen's $d=0.20$ for characteristic path length, Cohen's $d=0.27$ for global efficiency). Among CovBat, log-ComBat, log-CovBat, and gamma-GLM, all of which

addressed site effects on each global measure, leaving no measure with significant site-related differences. The gamma-GLM showed the best overall harmonization performance, yielding the smallest effect sizes across most measures.

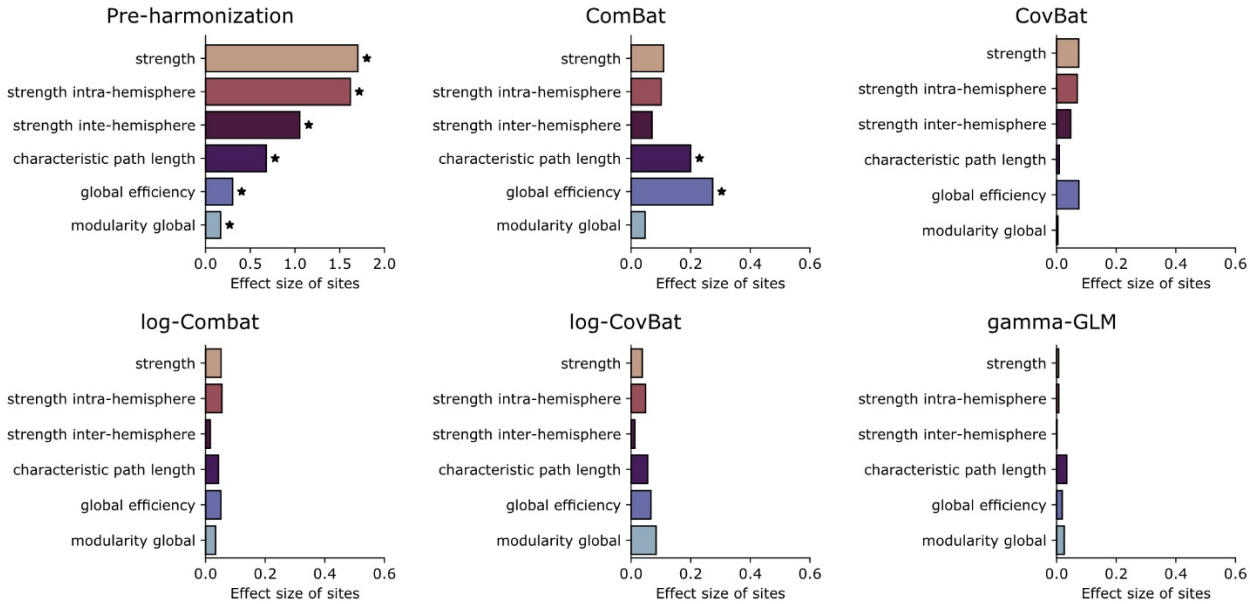

Figure 4 Effect size of site differences on global graph topological measures in the paired PNC-CAR cohort (Data Configuration 1). For each harmonization method, the Cohen's d effect sizes of sites were evaluated on six global graph topological measures (global strength, intra- and inter-hemisphere strength, characteristic path length, global efficiency, modularity). Significant site effects were indicated by asterisks (p < 0.05, two-sample t-test).

We then explored the site effects on nodewise graph topological features. Figure 5 presents the effect sizes of sites on four node-level graph features, with significant site effects marked by asterisks. The number of nodes with significant site effects was indicated in the top left corner of each plot. While all tested harmonization approaches addressed most site effects on node strength and clustering coefficient, ComBat and CovBat failed to correct site effects on betweenness centrality and local efficiency, leaving a substantial number of nodes with significant site effects (p<0.05, two-sample t-test).

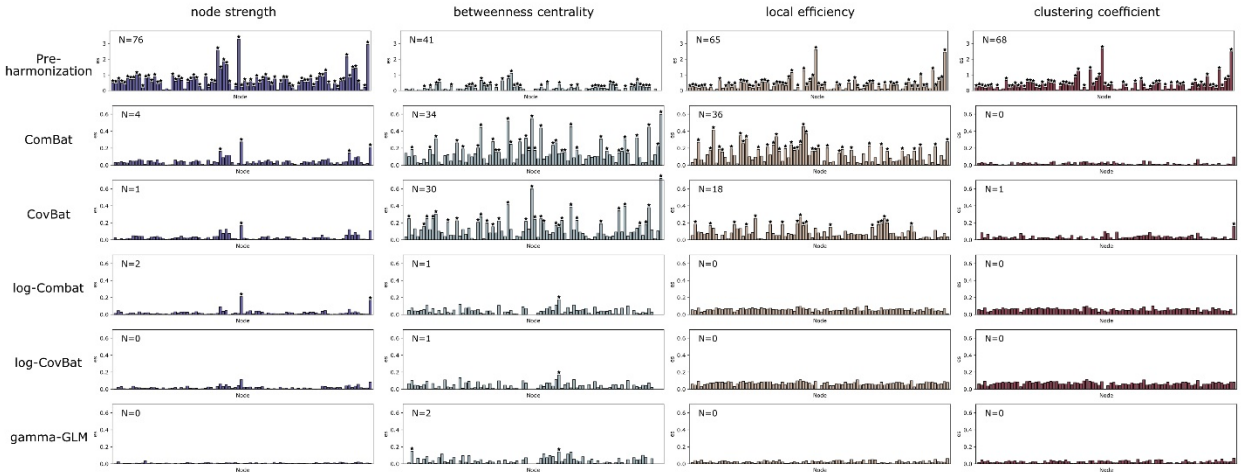

Figure 5 Effect size of site differences on nodal graph topological measures in the paired PNC-CAR cohort (Data Configuration 1). Four nodewise topological measures (node strength, betweenness centrality, local efficiency and clustering coefficient) were

evaluated. Significant site effects (p < 0.05, two-sample t-test) were indicated by asterisks. The number of nodes with significant site effects was noted in the top left corner of each plot.

3.4. Evaluation of Biological Variability Preservation

We assessed the replicability of age-related biological variability after applying different harmonization methods. For each site, Pearson correlations between edgewise connectivity strength and subjects' chronological age were calculated before and after harmonization, controlling for sex as a covariate. Figure 6A shows the edgewise correlations for the PNC site (first row) and the CAR site (second row) before and after harmonization with paired subjects (Data Configuration 1), displaying only edges with significant age associations ($p < 0.05$, Pearson's R). Notably, the gamma-GLM method most closely resembled the pre-harmonization patterns of age associations. Table 4 quantifies the replicability of these age associations using the true positive rate (TPR) of edges that detected significant correlations, compared to the pre-harmonization connectomes. Our results showed that gamma-GLM outperformed other methods by recovering all the positive and negative correlations in the original structural connectomes across both sites, followed by log-ComBat and log-CovBat.

We visualized the replicability of age associations with CAT plots, with each curve representing the concordance of rankings for all edgewise age correlations before and after harmonization, controlling for sex (Figure 6A). In the absence of confounding, gamma-GLM performed the best at both PNC and CAR sites, showing an almost flat CAT curve near 1. The log-ComBat and log-CovBat methods exhibited good overlap, performing slightly behind gamma-GLM, while ComBat and CovBat ranked the last.

Table 4 Recovery of connections with significant age correlation

Cohort	Site	# positive correlations	TPR of positive correlations					# negative correlations	TPR of negative correlations				
		pre-harmonization	ComBat	CovBat	log-ComBat	log-CovBat	gamma-GLM	pre-harmonization	ComBat	CovBat	log-ComBat	log-CovBat	gamma-GLM
Paired PNC-CAR	PNC	228	94.30%	94.30%	99.56%	95.61%	100%	143	82.51%	82.51%	98.60%	97.90%	100%
	CAR	227	72.69%	74.89%	94.27%	92.95%	100%	253	90.12%	92.09%	92.49%	92.09%	100%
Confounded NYU-TCD	ABIDEII-NYU	19	84.21%	78.95%	94.74%	84.21%	100%	78	80.76%	78.21%	82.05%	73.08%	100%
	ABIDEII-TCD	36	86.11%	58.33%	94.44%	66.67%	100%	95	84.21%	70.53%	85.26%	80.00%	100%

Abbreviations: TPR, true positive rate; # positive (or negative) correlations, number of connections with positive (or negative) age correlations

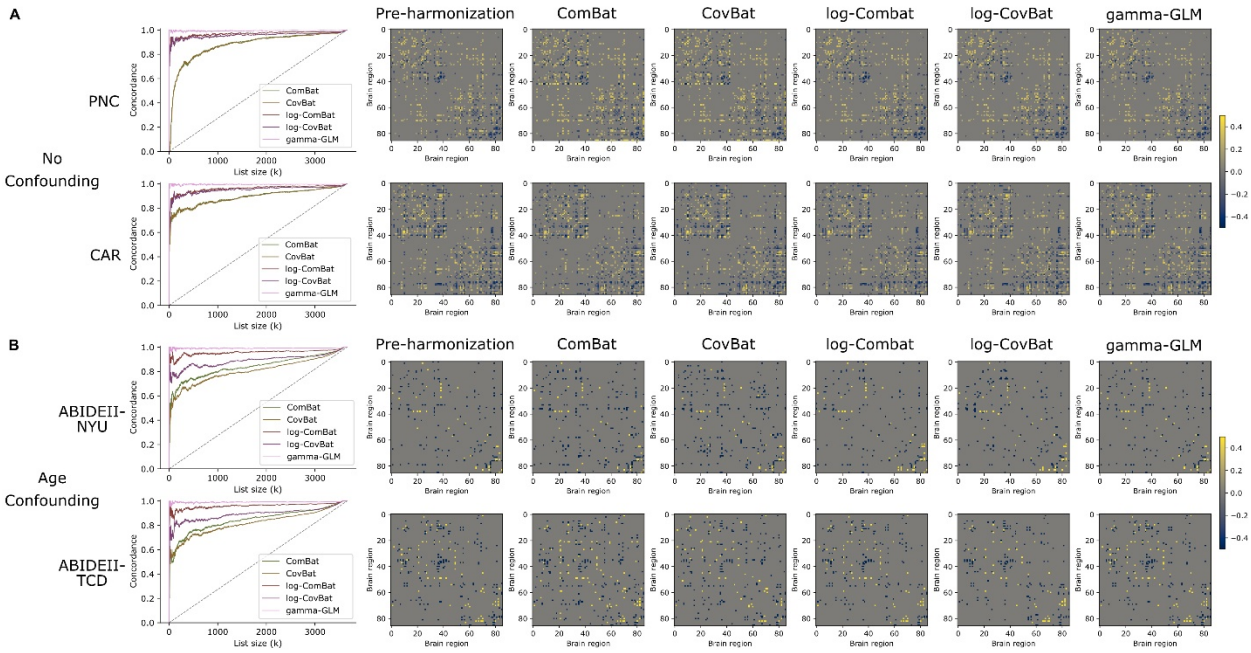

Figure 6 Replicability of age associations with edgewise connectivity strength after applying different harmonization methods. Two scenarios were tested: (A) the paired PNC-CAR cohort (Data Configuration 1) and (B) the confounded NYU-TCD cohort (Data Configuration 2). For each scenario, edgewise correlations between connectivity strength and age were shown for each site and harmonization method, displaying only significant age-associated connections ($p < 0.05$, Pearson's R). The CAT curves visualized the concordance of edgewise age associations before and after harmonization for each method. A CAT curve closer to one indicated better overlap in age associations.

3.5. Harmonization with Confounding Factors

The previous results were obtained by harmonizing two sites that paired for age and sex to minimize potential confounding of those variables across sites. However, this kind of pairing is not always feasible in multi-site studies, a better harmonization method should be applied on all available data across sites without the pairing process and should be robust to confounding. Figure 6B demonstrates the edgewise within-site correlations with age in the confounded NYU-TCD cohort (Data Configuration 2) before and after different harmonization approaches, with ABIDEII-NYU for the first row and ABIDEII-TCD for the second row. We noted that, compared to the paired cohort, the confounded cohort demonstrated different edgewise patterns between two sites, which may be due to the effects of age on structural connectomes are different across different age groups. The CAT plots show that the site-specific age-related patterns were best preserved for each site using gamma-GLM harmonization, followed by log-ComBat, log-CovBat, ComBat, and CovBat. The gamma-GLM harmonization recovered all edges with significant positive and negative correlations in presence of confounding between age and sites (Table 4).

To assess the age-effects between sites, we calculated the Frobenius distances between the edgewise correlation patterns of two sites, as shown in Table 5. Ideally, we expected the between-site distance for age-related patterns should remain the same before and after harmonization. The gamma-GLM most successfully persevered the age-related differences across sites, while log-ComBat and log-CovBat inadvertently reduced these differences. The ComBat and CovBat increased the age-related differences across sites, which may be due to the poor fitting using the normal distribution for edges.

Table 5 Frobenius distances between site-specific edgewise patterns of age associations

Pairwise Site differences	Pre-harmonization	ComBat	CovBat	log-ComBat	log-CovBat	gamma-GLM
Confounded NYU-TCD	11.7748	13.6246	12.8325	11.3402	11.4245	11.7764

Lastly, we evaluated the replicability of age associations when combining two sites. Figure S2 shows the CAT curves for two data configurations. Our results demonstrated that, in the presence of confounding, the replicability of age associations dropped dramatically compared to the paired cohort if sites are combined without proper harmonization, whereas properly harmonized ones maintained high replicability.

3.6. Use case 1: Brain age prediction on new datasets

In the following sessions, we evaluated the effectiveness of different harmonization methods in addressing several common challenges that encountered in multi-site structural connectome studies. For this analysis, we used the entire TDC cohort (Data Configuration 3), and the site effects before and after harmonization were reported in the Supplementary Materials.

We trained a brain age predictor using an SVR model on the PNC site's training set and apply it to the remaining sites before and after different harmonization methods. Table 6 shows the model's prediction performance at each new site, as well as for the combined data from all new sites. For reference, the original performance of the trained brain age predictor on the PNC site's test set was provided (underlined). Our results demonstrated that using proper harmonization to remove site-specific variations between the PNC and new sites could largely improve the transferability of the brain age predictor to new datasets. The gamma-GLM-harmonized data achieved the lowest MAE and RMSE, along with the highest Pearson R correlation, across most new sites with unseen data. Although log-ComBat attained the lowest MAE and RMSE in two sites (TimRobert, ABIDEII-TCD), it did not match gamma-GLM's performance in terms of Pearson R correlation.

Table 6 Impact of harmonization on the transferability of brain age predictors to new datasets

Dataset	Metric	Pre-harmonization	ComBat	CovBat	log-ComBat	log-CovBat	gamma-GLM
PNC	MAE	1.7321	1.6752	1.6433	1.7078	1.7001	1.7281
	RMSE	2.1681	2.0683	2.0400	2.1396	2.1329	2.1651
	Pearson R	0.7405	0.7662	0.7734	0.7483	0.7504	0.7414
CAR	MAE	3.7216	2.4807	2.4579	1.9644	1.9739	**1.8536**
	RMSE	4.3141	3.1131	3.0585	2.3928	2.4044	**2.2631**
	Pearson R	0.0548	0.3260	0.3175	0.6400	0.6446	**0.7074**
TimRobert	MAE	5.0619	2.7072	2.5526	**1.9518**	2.0532	2.0260
	RMSE	5.4438	3.2332	3.0941	**2.2854**	2.4712	2.3330
	Pearson R	0.2194	0.2418	0.2436	0.5063	0.4355	**0.5426**
ABIDEII-NYU	MAE	5.5198	3.1685	3.1896	1.8013	2.0526	**1.7870**
	RMSE	5.7817	3.7338	3.9420	2.2315	2.5171	**2.2105**
	Pearson R	-0.5111	-0.2693	-0.3182	0.4385	0.3965	**0.4676**
ABIDEII-SDSU	MAE	2.8504	3.4530	3.4925	2.5960	2.6739	**2.3822**
	RMSE	3.3662	4.1185	4.1908	3.1410	3.2959	**2.9947**
	Pearson R	0.2313	-0.4330	-0.4747	0.2078	0.1278	**0.3514**
ABIDEII-TCD	MAE	2.5921	2.5443	2.9942	**2.0126**	2.1363	2.0621
	RMSE	3.0745	3.2309	3.6377	**2.3909**	2.5667	2.4667
	Pearson R	0.5522	0.0682	-0.1160	0.6647	0.5909	**0.6804**
All New Sites	MAE	3.9080	2.6690	2.6702	2.0137	2.0719	**1.9419**
	RMSE	4.4883	3.3017	3.3100	2.4460	2.5346	**2.3653**
	Pearson R	0.2027	0.3935	0.3433	0.6790	0.6587	**0.7201**

[1]The performance of brain age predictors is reported on the testing set of the PNC site (underlined). The brain age predictors are directly applied to other sites to evaluate the transferability of the model to new datasets.

[2]The best-performing harmonization method for each metric is highlighted in bold.

[3]Abbreviations: MAE, mean absolute error; RMSE, root mean square error; Pearson R, Pearson's correlation.

3.7. Use case 2: Enhancing statistical power for ASD studies

Harmonization is essential when pooling data from multiple sources to enhance the statistical power of clinical studies. Our findings indicated that prior harmonization, the structural connectome data revealed no group-level

differences in global graph measures. Only a weak group-by-age interaction in clustering coefficients (standardized $\boldsymbol{\beta}$ =0.16, p=0.043) and a marginally significant interaction for global efficiency (standardized $\boldsymbol{\beta}$=0.16, p=0.052) were identified, while none were significantly associated to clinical scores. Post harmonization, we found that males with ASD showed significantly higher intra-hemispheric strength compared to the TDC males (Cohen's d=0.37, p=0.041). Significant group-by-age interactions were observed for clustering coefficient (standardized $\boldsymbol{\beta}$=0.27, p<0.001) and global efficiency (standardized $\boldsymbol{\beta}$=0.26, p=0.002), with both measures increasing with age in ASD but decreasing in TDC. Intra-hemispheric strength positively correlated with SRS-2 t-score (R=0.12, p=0.012), while clustering coefficient (R=0.15, p=0.006) and global efficiency (R=0.14, p=0.013) correlated positively with VABS communication subscale.

4. Discussion

This study addressed a key challenge in multi-site structural connectome research: how to effectively harmonize data collected from various sources to enable integration of multi-site studies and retrospective use of existing datasets. By doing so, it sought to provide a way to expand sample sizes, enabling a comprehensive representation of both neurotypical and neurodivergent populations to support more robust findings in the field. While previous studies have proposed various harmonization approaches at the level of dMRI signals [23, 26], they failed to remove site effects at the level of structural connectome [27], introducing biases in downstream graph analyses and other connectome-based investigations. Therefore, we believe these site effects should be specifically addressed and evaluated at the connectome level.

Through our experiments, we have demonstrated that applying harmonization techniques based on normal or log-normal distributions to connectomes data failed to remove mean site effects consistently at certain connections, potentially resulting in spurious findings in downstream analyses. Given the unique distributional properties of edgewise connectivity strength, connectome harmonization requires careful consideration in modeling and removing site-related effects. Standard harmonization methods in neuroimaging, such as ComBat, CovBat, or simple linear regression, assume a normal distribution and therefore should not be applied to connectomes, as most of their edges follow a gamma distribution rather than a normal distribution. Moreover, ComBat and CovBat may also produce negative values post harmonization, which lack physical meaning in structural connectomes where edges represent (normalized) numbers of streamlines connecting brain regions. Thus, the application of harmonization methods without validating the underlying data assumptions [47] could lead to incorrect conclusions or unintended consequences in multi-site studies.

Importantly, connectomes represent graph-structured brain connectivity, so harmonization should not only remove edgewise site effects but also eliminate site-related variations in their graph properties, commonly measured by graph topological measures [51]. Most prior work only evaluated graph measures at the global level [31, 47], but our results indicated that even though no site differences were observed in global measures, they may still be exhibited at the nodal level. For example, while CovBat effectively removed site effects in all common global graph measures, significant site effects persist in nodal measures such as betweenness centrality and local efficiency. Therefore, we recommend evaluating harmonization effectiveness for structural connectomes across all three levels—edgewise connectivity values, nodal, and global graph measures—to ensure that site effects are adequately eliminated. Our results showed that the harmonization framework based on gamma-GLM performed best across all three levels, followed by log-CovBat and log-ComBat approaches with our adaptation to structural connectomes. In contrast, ComBat and CovBat are not recommended for structural connectome harmonization.

The primary goal of harmonization is to integrate data across sites to facilitate the investigation of biologically based hypotheses using larger and more comparable structural connectome data. Therefore, it is crucial that biologically relevant properties are not inadvertently altered or removed during the harmonization process. Given that age-related connectivity changes are critical in studies of neurodevelopment and neurodegenerative processes [56, 63], we assessed edgewise associations with age, both pre- and post-harmonization. Our findings indicated that gamma-GLM-based harmonization framework most effectively preserved these biologically meaningful associations across sites and remained robust even when age was confounded with site effects. This may be due to the gamma-GLM-

based model providing a better fit for covariate effects in structural connectome data, further emphasizing its advantage in connectome harmonization over other approaches.

We further demonstrated the practical utility of structural connectome harmonization with two use cases. One commonly reported challenge in collaborative research is that predictive models developed with one dataset often fail to predict unseen data from the new sites [64, 65]. This occurs because new datasets may exhibit substantial differences in scanners, acquisition protocols, and other site-related effects compared to the training cohort, limiting the transferability of predictive models across sites. In this work, we created machine learning models to predict structural connectome-based brain age using data from one site and evaluate the transferability of these predictors on other sites. Our findings indicated that harmonizing structural connectomes across sites, particularly using gamma-GLM based model, enhanced the transferability of predictive models to unseen data at new sites. This approach can be further adapted for a federated learning setup in healthcare [66, 67]. In this context, researchers can first harmonize and standardize decentralized data sources to a central site to ensure comparability, thereby improving further transferability when deployed from and to the central site.

Another key application of harmonization is to increase the statistical power of clinical studies via integrating structural connectomes from multiple sources. This is especially beneficial for clinical studies involving patients, where data from a single site is often limited. By pooling patient data from various sources, researchers can create a larger and more diverse cohort, which are essential for detecting subtle differences in their connectivity patterns associated with different conditions. Therefore, the effectiveness of harmonization should be evaluated not only in the context of the control groups but also for the patient population. Our results on a clinical cohort with neurotypical and neurodivergent individuals showed that without proper harmonization, the associations between graph topological measures and neurodivergent traits may be obscured by site-related effects. Harmonization with gamma-GLM-based framework successfully uncovered the biological features related to disorders in structural connectomes and effectively enhanced statistical power compared to other approaches.

Finally, we provided practical guidance for researchers seeking to harmonize structural connectome data, as shown in Figure 1. Overall, the gamma-GLM-based harmonization approach is our top recommendation, as it outperformed other models in removing site effects in edgewise mean connectivity strength, as well as in global and nodal graph measures. Furthermore, it most effectively preserved demographic and disorder-related biological integrity throughout the harmonization process compared to other approaches. Users should also be aware that the gamma-GLM-based approach does not model site-specific variance and covariance. If image acquisition and reconstruction protocols vary significantly across sites, careful checks are necessary to ensure that this variance does not impact image properties [68]. Additionally, while we did not observe significant covariance site effects in this study, such effects may emerge in subnetwork analyses or when structural connectomes are modulated by diffusion metrics like fractional anisotropy (FA) or mean diffusivity (MD). In these special cases, log-ComBat or log-CovBat may be a more appropriate choice.

Despite such a comprehensive evaluations, there are several issues that warrant attention in future research. For instance, we tested only a few distributions from the exponential family and focused on modeling mean, variance, and covariance site effects. It would be valuable to explore additional distributions and statistical metrics, particularly when structural connectomes are created using more complex methods. In such cases, more flexible distributional regression models, like the Generalized Additive Model for Location, Scale, and Shape (GAMLSS) [69], might be suitable. Another area for future work lies in expanding demographic variables involved during harmonization. Incorporating factors like ethnicity and socioeconomic status could enhance the applicability of findings across diverse populations to promote health equity [70, 71]. Our harmonization and evaluation framework can also be adapted to better address the specific goals of structural connectome studies in these contexts.

In summary, we have outlined a path forward for integrating multi-site structural connectome datasets in collaborative research, supporting a range of advanced research goals. Owing to the complex nature of connectomes, it is crucial to examine the distributional properties of the connectome data prior to harmonization to ensure that the assumptions of each harmonization method are met. Based on this assessment, the appropriate harmonization framework should be adopted accordingly, such as gamma-GLM or log-Combat/log-Covbat that we have provided. All harmonization results should be carefully examined for site effects at various levels of analysis, including

edgewise connectivity strength, as well as global and nodal-level graph properties, to ensure unbiased interpretation of the data. Finally, extreme care is needed when harmonizing structural connectome data constructed using unconventional approaches, as well as when considering diverse demographic and clinical characteristics within the population.

References

1. Iturria-Medina, Y., et al., *Studying the human brain anatomical network via diffusion-weighted MRI and Graph Theory.* Neuroimage, 2008. **40**(3): p. 1064-1076.
2. Rudie, J.D., et al., *Altered functional and structural brain network organization in autism.* Neuroimage Clin, 2012. **2**: p. 79-94.
3. Di Martino, A., et al., *Enhancing studies of the connectome in autism using the autism brain imaging data exchange II.* Scientific data, 2017. **4**(1): p. 1-15.
4. Kubicki, M., et al., *A review of diffusion tensor imaging studies in schizophrenia.* Journal of psychiatric research, 2007. **41**(1-2): p. 15-30.
5. Shen, R.S., et al. *Graph matching based connectomic biomarker with learning for brain disorders*. in *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis: Second International Workshop, UNSURE 2020, and Third International Workshop, GRAIL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 2*. 2020. Springer.
6. Kim, J., et al., *Disrupted structural connectome is associated with both psychometric and real-world neuropsychological impairment in diffuse traumatic brain injury.* Journal of the International Neuropsychological Society, 2014. **20**(9): p. 887-896.
7. Button, K.S., et al., *Power failure: why small sample size undermines the reliability of neuroscience.* Nature reviews neuroscience, 2013. **14**(5): p. 365-376.
8. Smith, S.M. and T.E. Nichols, *Statistical challenges in "big data" human neuroimaging.* Neuron, 2018. **97**(2): p. 263-268.
9. Ioannidis, J.P., *Why most published research findings are false.* PLoS medicine, 2005. **2**(8): p. e124.
10. Wheeler, A.L. and A.N. Voineskos, *A review of structural neuroimaging in schizophrenia: from connectivity to connectomics.* Frontiers in human neuroscience, 2014. **8**: p. 653.
11. Picci, G., S.J. Gotts, and K.S. Scherf, *A theoretical rut: revisiting and critically evaluating the generalized under/over-connectivity hypothesis of autism.* Developmental science, 2016. **19**(4): p. 524-549.
12. Qi, S., et al., *The influence of construction methodology on structural brain network measures: A review.* Journal of neuroscience methods, 2015. **253**: p. 170-182.
13. Grech-Sollars, M., et al., *Multi-centre reproducibility of diffusion MRI parameters for clinical sequences in the brain.* NMR in Biomedicine, 2015. **28**(4): p. 468-485.
14. Zhu, T., et al., *Quantification of accuracy and precision of multi-center DTI measurements: a diffusion phantom and human brain study.* Neuroimage, 2011. **56**(3): p. 1398-1411.
15. Fornito, A., A. Zalesky, and M. Breakspear, *Graph analysis of the human connectome: promise, progress, and pitfalls.* Neuroimage, 2013. **80**: p. 426-444.
16. Cao, H., et al., *Toward leveraging human connectomic data in large consortia: generalizability of fMRI-based brain graphs across sites, sessions, and paradigms.* Cerebral Cortex, 2019. **29**(3): p. 1263-1279.
17. Abraham, A., et al., *Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example.* NeuroImage, 2017. **147**: p. 736-745.

18. Vollmar, C., et al., *Identical, but not the same: intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0 T scanners.* Neuroimage, 2010. **51**(4): p. 1384-1394.
19. Magnotta, V.A., et al., *Multicenter reliability of diffusion tensor imaging.* Brain connectivity, 2012. **2**(6): p. 345-355.
20. Jovicich, J., et al., *Multisite longitudinal reliability of tract-based spatial statistics in diffusion tensor imaging of healthy elderly subjects.* Neuroimage, 2014. **101**: p. 390-403.
21. Bonilha, L., et al., *Reproducibility of the structural brain connectome derived from diffusion tensor imaging.* PloS one, 2015. **10**(9): p. e0135247.
22. Buchanan, C.R., et al., *Test–retest reliability of structural brain networks from diffusion MRI.* Neuroimage, 2014. **86**: p. 231-243.
23. Mirzaalian, H., et al., *Inter-site and inter-scanner diffusion MRI data harmonization.* NeuroImage, 2016. **135**: p. 311-323.
24. Koppers, S., C. Haarburger, and D. Merhof. *Diffusion MRI signal augmentation: from single shell to multi shell with deep learning*. in *Computational Diffusion MRI: MICCAI Workshop, Athens, Greece, October 2016 19*. 2017. Springer.
25. Koppers, S., et al. *Spherical harmonic residual network for diffusion signal harmonization*. in *Computational Diffusion MRI: International MICCAI Workshop, Granada, Spain, September 2018 22*. 2019. Springer.
26. Huynh, K.M., et al., *Multi-site harmonization of diffusion MRI data via method of moments.* IEEE transactions on medical imaging, 2019. **38**(7): p. 1599-1609.
27. Kurokawa, R., et al., *Cross-scanner reproducibility and harmonization of a diffusion MRI structural brain network: A traveling subject study of multi-b acquisition.* NeuroImage, 2021. **245**: p. 118675.
28. Johnson, W.E., C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical Bayes methods.* Biostatistics, 2007. **8**(1): p. 118-127.
29. Morris, C.N., *Parametric empirical Bayes inference: theory and applications.* Journal of the American statistical Association, 1983. **78**(381): p. 47-55.
30. Fortin, J.-P., et al., *Harmonization of multi-site diffusion tensor imaging data.* Neuroimage, 2017. **161**: p. 149-170.
31. Yu, M., et al., *Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data.* Human brain mapping, 2018. **39**(11): p. 4213-4227.
32. Yamashita, A., et al., *Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias.* PLoS biology, 2019. **17**(4): p. e3000042.
33. Ma, D., et al., *Quantitative assessment of field strength, total intracranial volume, sex, and age effects on the goodness of harmonization for volumetric analysis on the ADNI database.* Human brain mapping, 2019. **40**(5): p. 1507-1527.
34. Chen, A.A., et al., *Mitigating site effects in covariance for machine learning in neuroimaging data.* Human brain mapping, 2022. **43**(4): p. 1179-1195.
35. Satterthwaite, T.D., et al., *Neuroimaging of the Philadelphia Neurodevelopmental Cohort.* NeuroImage, 2014. **86**: p. 544-553.
36. Tunç, B., et al., *Deviation from normative brain development is associated with symptom severity in autism spectrum disorder.* Molecular autism, 2019. **10**: p. 1-14.
37. Ghanbari, Y., et al., *Identifying group discriminative and age regressive sub-networks from DTI-based connectivity via a unified framework of non-negative matrix factorization and graph embedding.* Medical Image Analysis, 2014. **18**(8): p. 1337-1348.
38. Dhollander, T., et al. *Improved white matter response function estimation for 3-tissue constrained spherical deconvolution*. in *Proc. Intl. Soc. Mag. Reson. Med*. 2019.

39. Dhollander, T., D. Raffelt, and A. Connelly. *Unsupervised 3-tissue response function estimation from single-shell or multi-shell diffusion MR data without a co-registered T1 image*. in *ISMRM workshop on breaking the barriers of diffusion MRI*. 2016. Lisbon, Portugal.
40. Jeurissen, B., et al., *Investigating the prevalence of complex fiber configurations in white matter tissue with diffusion magnetic resonance imaging.* Human brain mapping, 2013. **34**(11): p. 2747-2766.
41. Tournier, J.D., F. Calamante, and A. Connelly. *Improved probabilistic streamlines tractography by 2nd order integration over fibre orientation distributions*. in *Proceedings of the international society for magnetic resonance in medicine*. 2010. John Wiley & Sons, Inc, New Jersey.
42. Fischl, B., *FreeSurfer.* Neuroimage, 2012. **62**(2): p. 774-781.
43. Avants, B.B., et al., *Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain.* Medical image analysis, 2008. **12**(1): p. 26-41.
44. Desikan, R.S., et al., *An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest.* Neuroimage, 2006. **31**(3): p. 968-980.
45. Fischl, B., et al., *Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain.* Neuron, 2002. **33**(3): p. 341-355.
46. McCullagh, P., *Generalized linear models*. 2019: Routledge.
47. Onicas, A.I., et al., *Multisite harmonization of structural DTI networks in children: An A-CAP study.* Frontiers in Neurology, 2022. **13**: p. 850642.
48. Manning, W.G. and J. Mullahy, *Estimating log models: to transform or not to transform?* Journal of health economics, 2001. **20**(4): p. 461-494.
49. Keene, O.N., *The log transformation is special.* Statistics in medicine, 1995. **14**(8): p. 811-819.
50. Nelder, J.A. and R.W. Wedderburn, *Generalized linear models.* Journal of the Royal Statistical Society Series A: Statistics in Society, 1972. **135**(3): p. 370-384.
51. Rubinov, M. and O. Sporns, *Complex network measures of brain connectivity: uses and interpretations.* Neuroimage, 2010. **52**(3): p. 1059-1069.
52. Irizarry, R.A., et al., *Multiple-laboratory comparison of microarray platforms.* Nature methods, 2005. **2**(5): p. 345-350.
53. Franke, K. and C. Gaser, *Ten years of BrainAGE as a neuroimaging biomarker of brain aging: what insights have we gained?* Frontiers in neurology, 2019. **10**: p. 789.
54. Liem, F., et al., *Predicting brain-age from multimodal imaging data captures cognitive impairment.* Neuroimage, 2017. **148**: p. 179-188.
55. Dosenbach, N.U., et al., *Prediction of individual brain maturity using fMRI.* Science, 2010. **329**(5997): p. 1358-1361.
56. Bethlehem, R.A., et al., *Brain charts for the human lifespan.* Nature, 2022. **604**(7906): p. 525-533.
57. Kaufmann, T., et al., *Common brain disorders are associated with heritable patterns of apparent aging of the brain.* Nature neuroscience, 2019. **22**(10): p. 1617-1623.
58. Rane, P., et al., *Connectivity in autism: a review of MRI connectivity studies.* Harvard review of psychiatry, 2015. **23**(4): p. 223-244.
59. Ha, S., et al., *Characteristics of brains in autism spectrum disorder: structure, function and connectivity across the lifespan.* Experimental neurobiology, 2015. **24**(4): p. 273.
60. Vissers, M.E., M.X. Cohen, and H.M. Geurts, *Brain connectivity and high functioning autism: a promising path of research that needs refined models, methodological convergence, and stronger behavioral links.* Neuroscience & Biobehavioral Reviews, 2012. **36**(1): p. 604-625.
61. Lewis, J.D., et al., *Network inefficiencies in autism spectrum disorder at 24 months.* Translational psychiatry, 2014. **4**(5): p. e388-e388.

62. Dudoit, S., et al., *Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.* Statistica sinica, 2002: p. 111-139.
63. Hagmann, P., et al., *White matter maturation reshapes structural connectivity in the late developing human brain.* Proceedings of the National Academy of Sciences, 2010. **107**(44): p. 19067-19072.
64. Goetz, L., et al., *Generalization—a key challenge for responsible AI in patient-facing clinical applications.* npj Digital Medicine, 2024. **7**(1): p. 126.
65. Rajpurkar, P., et al., *AI in health and medicine.* Nature medicine, 2022. **28**(1): p. 31-38.
66. Antunes, R.S., et al., *Federated learning for healthcare: Systematic review and architecture proposal.* ACM Transactions on Intelligent Systems and Technology (TIST), 2022. **13**(4): p. 1-23.
67. Li, X., et al., *Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results.* Medical image analysis, 2020. **65**: p. 101765.
68. Orlhac, F., et al., *A guide to ComBat harmonization of imaging biomarkers in multicenter studies.* Journal of Nuclear Medicine, 2022. **63**(2): p. 172-179.
69. Rigby, R.A. and D.M. Stasinopoulos, *Generalized additive models for location, scale and shape.* Journal of the Royal Statistical Society Series C: Applied Statistics, 2005. **54**(3): p. 507-554.
70. Norori, N., et al., *Addressing bias in big data and AI for health care: A call for open science.* Patterns, 2021. **2**(10).
71. Wang, R., P. Chaudhari, and C. Davatzikos, *Bias in machine learning models can be significantly mitigated by careful training: Evidence from neuroimaging studies.* Proceedings of the National Academy of Sciences, 2023. **120**(6): p. e2211613120.