Big Data, Small Bias: Harmonizing Diffusion MRI-Based Structural Connectomes to Mitigate Site-Related Bias in Data Integration

Rui Sherry Shen[1,2*], Drew Parker[2], Andrew An Chen[3], Benjamin E. Yerys[4,5,6], Birkan Tunç[4,5], Timothy P.L. Roberts[7], Russell T. Shinohara[8], Ragini Verma[1,2*]

[1]*Department of Bioengineering, University of Pennsylvania, Philadelphia, United States*
[2]*Diffusion & Connectomics In Precision Healthcare Research, 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]*Program in Advanced Imaging Research, Department of Radiology, Children's Hospital of Philadelphia, Philadelphia, United States*
[8]*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 changes associated with various disorders. However, small sample sizes in individual studies, along with highly heterogeneous disorder-related manifestations, underscore the need to pool datasets across multiple studies to be able to identify coherent and generalizable connectivity patterns linked to these disorders. Yet, combining datasets introduces site-related differences due to variations in scanner hardware or acquisition protocols. These differences highlight the necessity for statistical data harmonization to mitigate site-related effects on structural connectomes while preserving the biological information associated with participant demographics and the disorders. While several paradigms exist for harmonizing normally distributed neuroimaging measures, this paper represents the first effort to establish a harmonization framework specifically tailored for the structural connectome. We conduct a thorough investigation of various statistical harmonization methods, adapting them to accommodate the unique distributional characteristics and graph-based properties of structural connectomes. Through rigorous evaluation, we demonstrate that the generalized linear model with a log-linked gamma model (gamma-GLM) outperforms other approaches in modeling structural connectomes, enabling the effective removal of site-related biases in both edge-based and downstream graph analyses while preserving biological variability. Two real-world applications further highlight the utility of our harmonization framework in addressing challenges in multi-site structural connectome analysis. Specifically, harmonization with gamma-GLM enhances the generalizability of connectome-based machine learning predictors to new datasets and increases statistical power for detecting group-level differences. Our work provides essential guidelines for harmonizing multi-site 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

Diffusion-weighted MRI (dMRI) [1] offers a non-invasive approach for characterizing the brain's structural connectivity, represented as a structural connectome. Connectomic studies have greatly advanced our understanding of the underlying effects of various brain disorders, including autism spectrum disorder (hereafter, autism) [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 efforts [7-11]. In response, the scientific community has embraced big data initiatives by pooling data from multiple clinical sites, with the goal of offering a more comprehensive representation of disorder heterogeneity.

However, substantial site-related variability has been observed in dMRI data collected from different sites, driven by different scanners, acquisition protocols, preprocessing pipelines, and technical differences even with the same manufacturers [12-14]. This site-related variability affects downstream analyses, compromising the comparability of connectome data and obscuring biologically meaningful associations in statistical analyses [15-17]. Various

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

harmonization solutions have been proposed to mitigate site-related biases, targeting the discrepancies at different stages of the data collection and analysis pipelines. For instance, efforts to standardize acquisition protocols and parameters have reduced site-related effects [18-20]. However, this approach cannot resolve site differences retrospectively when data is already 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]. Others have attempted to mitigate site-related variations by harmonizing raw dMRI data at the signal level. Mirzaalian et al. [23] introduced a technique that transformed 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 demonstrated effectiveness 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 for addressing site effects at the dMRI signal level assume that downstream derived measures, 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 further introduce site-related biases in connectomes. 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-normal distribution. As a result, connectome harmonization requires specialized methods to address this complexity. While various statistical models have been proposed for harmonizing neuroimaging data, their applicability to structural connectomes remains largely 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-based scalar maps such as fractional anisotropy (FA) and mean diffusivity (MD), and Yu et al. [31] applied ComBat to resting-state functional connectivity. Other widely used harmonization methods include generalized linear model (GLM) [32, 33] and CovBat [34]. Yet, most of these harmonization methods are designed for data that follow a normal distribution. Several studies have applied them to structural connectomes without validating the underlying data assumptions [35], which could lead to incorrect conclusions or unintended consequences in multi-site studies. To date, no harmonization tools have been specifically designed or validated for structural connectomes.

In this work, we address this gap by proposing a harmonization framework for structural connectomes. We propose several statistical harmonization frameworks to accommodate connectome data and conduct a comprehensive evaluation of their effectiveness in eliminating site-related differences while preserving connectivity information and biological associations with age and clinical conditions. To our knowledge, this study is the first effort to propose a tailored harmonization framework specifically for 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 generalizability 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 pooled structural connectome data from six datasets, including the Philadelphia Neurodevelopmental Cohort (PNC) [36], the Center for Autism Research (CAR) [37], Lurie MEGlab [38], and three clinical sites from Autism Brain Imaging Data Exchange II (ABIDEII-NYU, ABIDEII-SDSU, ABIDEII-TCD) [3]. Our study included 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
Lurie MEGlab	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 created 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.

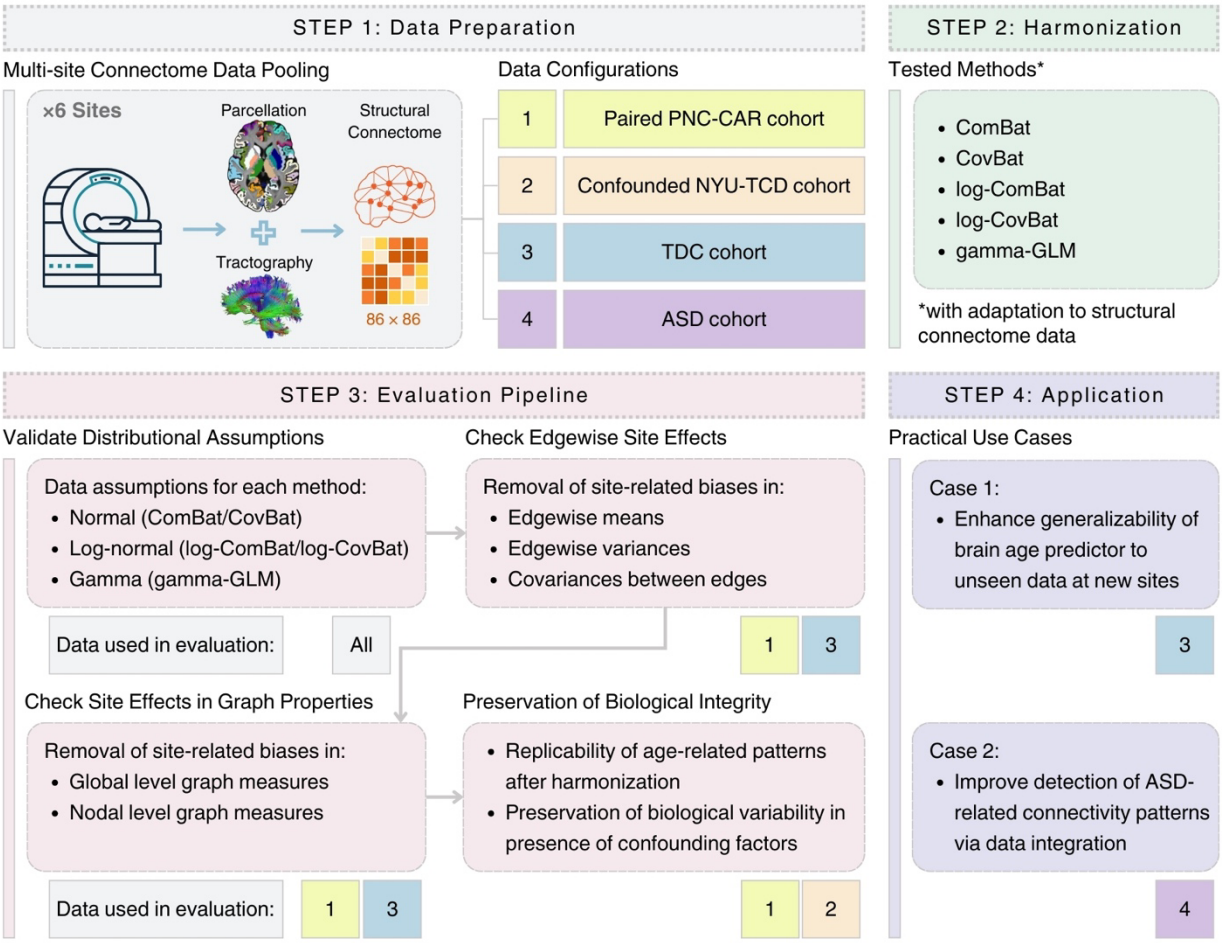

Figure 1 Overview of the harmonization and evaluation framework for structural connectomes. After pooling MRI data from six sites and extracting structural connectomes, we created four distinct data configurations by combining the cohorts in various ways (Step 1). Different harmonization methods were applied to each data configuration for the method comparison (Step 2). We evaluated 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 demonstrated 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 selected participants from the TDC group from our two largest datasets, PNC and CAR, ensuring they were closely matched in age and sex. This approach aimed to mimic the traveling-subject research paradigm [23], minimizing confounding effects that could impact the harmonization results. As a result, we included 100 participants (76 males, 24 females) for each site. Participants from the PNC site ranged in age from 8.17 to 18.08 years, with a mean age of 12.31 years and a standard deviation of 2.70 years. Similarly, participants from the CAR site ranged from 8.08 to 17.83 years, with a mean age of 12.27 years and a standard deviation of 2.73 years.

2.1.2. Data Configuration 2: Confounded NYU-TCD cohort

We used all TDC participants from the ABIDEII-NYU and ABIDEII-TCD datasets to create a cohort where age and site were confounded. The ABIDEII-NYU dataset included participants within the youngest age range among the six datasets, spanning 6.66 to 12.90 years, with a mean age of 9.55 years and a standard deviation of 1.77 years. In contrast, the ABIDEII-TCD dataset comprised the oldest participants, ranging from 10.25 to 20.00 years, with a

mean of 16.18 years and a standard deviation of 2.97 years. Since both cohorts included only male participants, no sex-related biases were present.

2.1.3. Data Configuration 3: TDC cohort

We pooled all TDC data from our datasets, resulting in a total of 1194 participants (623 males, 571 females). The age of participants ranged 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 represented a more generic case in which age and sex were unbalanced with respect to site, potentially introducing statistical confounding.

2.1.4. Data Configuration 4: ASD cohort

All datasets, except for PNC, included neuroimaging data from children with ASD. We pooled the ASD data from these five datasets to create a cohort of 309 ASD participants (267 males and 42 females) with 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 were harmonized by adjusting for site effects estimated from the corresponding TDC cohort.

2.2. Data Processing

2.2.1. Pre-processing and Quality Control

We conducted a series of preprocessing steps for dMRI scans, including denoising using joint local principal component analysis (PCA) [39], correction for eddy currents and movements with FSL EDDY [40], and skull stripping with BET2 [41]. We preprocessed 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]. Constrained spherical deconvolution (CSD) of dMRI was then conducted using MRtrix3 [44], with the fiber response function estimated using D'Hollander's algorithm [45, 46] and the fiber orientation distribution (FOD) fitted using the method proposed by Jeurissen and others [47]. We performed probabilistic fiber tracking with a step size of 1 mm and an angle threshold of 60 degree utilizing the iFOD2 algorithm [48] and anatomically constrained tractography (ACT) [48], with 250 seeds placed at random inside each voxel of the grey-matter white-matter interface (GMWMI). We conducted manual quality control to exclude data with poor acquisition quality or failed registration.

2.2.2. Construction of Structural Connectomes

We parcellated the brain into 86 regions of interest (ROIs), including 68 cortical regions from the Desikan-Killiany atlas [49] and 18 subcortical regions from FreeSurfer's atlas [50]. We calculated the number of streamlines intersecting each pairwise combination of these ROIs, yielding an 86x86 connectivity matrix. Post-processing of structural connectomes involved removing self-connections and symmetrizing the connectivity matrix. Additionally, we normalized the structural connectomes by dividing edge weights by each subject's total GMWMI volume. Since structural connectomes derived from probabilistic tractography may include numerous spurious connections that could affect downstream analysis, we applied consistency-based thresholding [51] to filter out spurious connections and retained only the most consistent ones across the population. For our primary analyses, we used a density threshold of 40%, while analysis of different consistency levels was also provided in Supplementary Materials.

2.3. Harmonization Methods

Based on different distributional assumptions, we investigated five statistical approaches for structural connectome harmonization. Firstly, we tested two widely used harmonization tools, ComBat and CovBat, both of which assumed normal distribution of the data. We then modified the two approaches and introduced log-ComBat and log-CovBat harmonization, which utilized a log-normal data assumption. Finally, we developed a generalized linear model with gamma distributed variables (gamma-GLM) for connectome harmonization, adopting a gamma distribution to fit the connectome data. Table 2 compares these harmonization methods in terms of their capabilities, as also demonstrated in previous literature [28, 34, 52]. All harmonization methods were performed with sex, age, and age^2 included as covariates.

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 assumed 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 comprised N_m samples, with $n = 1, 2, \dots, N_m$ representing the index of the subject. Y_{mne} represents 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. In the following sections, we provided detailed descriptions of each technique for structural connectome harmonization.

2.3.1. Applying 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 extended 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 followed the approach of Onicas et al. [35] 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. Applying CovBat to Connectomes

ComBat treats each edge in the structural connectomes independently, without accounting for the covariance between edges. However, site effects might be correlated edgewise, exhibiting 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 first applies ComBat to remove the site-specific mean and variance, yielding 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:

$$\boldsymbol{\epsilon}_{\boldsymbol{mn}}^{\boldsymbol{ComBat}} = \sum_{l=1}^{L} \xi_{mnl} \boldsymbol{\Phi}_{\boldsymbol{l}}$$

Here, $\boldsymbol{\Phi}_{\boldsymbol{l}}$ 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, they can be harmonized 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:

$$\boldsymbol{e}_{\boldsymbol{mn}}^{\boldsymbol{CovBat}} = \sum_{k=1}^{K} \xi_{mnk}^{CovBat} \boldsymbol{\Phi}_{\boldsymbol{k}} + \sum_{l=K+1}^{L} \xi_{mnl}^{CovBat} \boldsymbol{\Phi}_{\boldsymbol{l}}$$

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

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

Similar to the post-processing step in ComBat, we reassigned zero values to harmonized connections that originally had zero edge values before harmonization.

2.3.3. Adapting log-ComBat for 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 [53-55]. 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:

$$\ln(Y_{mne}) = \alpha_e + \boldsymbol{X}_{\boldsymbol{mn}}^{\boldsymbol{T}} \boldsymbol{\beta}_{\boldsymbol{e}} + \gamma_{me} + \delta_{me}\epsilon_{mne}$$

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

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

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 added a small positive constant c to Y_{mne}. 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. Developing log-CovBat for Connectomes

We also developed log-CovBat, a modification of the CovBat algorithm that incorporates log-scaling to address the skewness of edge weights in structural connectomes. We first applied a log-transformation to all edge values, then performed CovBat on the log-transformed edges. The final harmonized edge values were obtained by applying the inverse logarithm (exponentiation), as formulated below:

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

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. Developing gamma-GLM for Connectomes

The generalized linear model (GLM) provides an alternative for regression when the assumptions of normality and homoskedasticity are not met [56]. By employing a gamma distribution, GLM offers greater flexibility in modeling a wide range of data shapes, particularly well-suited for right-skewed data. 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 of Y_{mne} is given by:

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

The parameters k_e and θ_e can be fitted using GLM regression. We modeled the effects of site and other covariates on the expected structural connectomes using a log-link function to ensure positive edge values after harmonization, which can be expressed as:

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

The mean site effects γ_{me}^* for site m at edge e can be then estimated. We regressed out the estimated mean site effects γ_{me}^* from the edge values Y_{mne} and applied the inverse logarithm transformation (exponentiation):

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

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 conducted goodness-of-fit tests on the structural connectome data to assess the validity of these data assumptions. Specifically, we fitted each statistical distribution to the observed structural connectomes at each site. Kolmogorov–Smirnov (KS) test was used 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 examined whether each harmonization method successfully eliminated site-related biases in the mean, variance, and covariance of edgewise connectome strength. This evaluation utilized the paired PNC-CAR cohort (Data Configuration 1) and the multi-site TDC cohort (Data Configuration 3). Prior to edgewise evaluation, covariates for sex, age, and age^2 were regressed out. To assess mean site effects, we performed a non-parametric Kruskal-Wallis test to identify edgewise group differences across sites. Variance site effects were evaluated using the Brown-Forsythe test on the residuals of edgewise strength. Brown-Forsythe test is a non-parametric equivalent of Levene's test to determine whether variances were equal across sites. For site-specific covariance effects, empirical covariance matrices were computed for each site. The Frobenius distance between within-site covariance matrices was then calculated for each site pair, both before and after harmonization. Box-M test was used to determine whether the site-specific covariance matrices were statistically equivalent.

2.5.2. On Removing Site Effects in Derived Graph Topological Measures

Structural connectomes represent brain connectivity as graphs, therefore, effective harmonization approaches should address not only edgewise site effects but also site-related variations in graph properties, which are typically

assessed using graph topological measures [57]. We computed graph topological measures at two levels: node-level features (node strength, betweenness centrality, local efficiency, clustering coefficient) and global graph measures (global strength, intra- and inter-hemispheric strength, characteristic path length, global efficiency, modularity). Site-related differences in these measures were analyzed before and after applying harmonization methods using a two-sample t-test for the paired PNC-CAR cohort (Data Configuration 3) and a one-way ANOVA test for the multi-site TDC cohort (Data Configuration 3), after controlling for sex, age, and age^2. Effect sizes for site differences were 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 remove non-biological variability introduced by different sites and scanners while preserving the statistical power to detect biologically meaningful associations. To validate this, we calculated the Spearman correlation coefficients between edgewise connectivity strength and participants' chronological age in the paired PNC-CAR cohort (Data Configuration 1), controlling for sex as a covariate. We expected that the harmonized structural connectomes would retain the age-connectivity relationships observed in the pre-harmonization data for each site. The replicability of age-related patterns was quantified by the true positive rate (sensitivity) of connections showing significant age correlations in the harmonized connectomes compared to the pre-harmonization data. Additionally, we ranked the connections by the absolute values of their age correlations and computed the overlap between the top k correlations in the harmonized and pre-harmonization data within each site. We visualized the results using concordance at the top (CAT) curves [58] by varying across all k values. 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 presents a challenge for harmonization, as removing site-related variation can inadvertently change age-related variation if not handled carefully. To evaluate how different harmonization methods handle this issue, we conducted a Spearman correlation analysis to examine the relationship between edgewise connections and participants' chronological age within each site using the confounded NYU-TCD cohort (Data Configuration 2). Similar to the previous analysis, CAT plots were used to visualize the replicability of these age associations before and after applying various harmonization methods, while the true positive rate quantified the replicability of connections with significant age correlations. Additionally, we computed the Frobenius distance between the edgewise age correlation matrices of the two sites. We expected this distance would remain stable before and after harmonization, indicating that age-related differences were preserved between sites. Finally, we jointly tested the replicability of age associations after combining data from both sites post-harmonization. We also included a scenario where the two sites were combined without proper harmonization to emphasize the critical role of harmonization in addressing 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 evaluated the effectiveness of harmonization frameworks through two practical use cases: 1) improving the generalizability of prediction models to new datasets; we used brain age prediction as an example; and 2) enhancing statistical power for clinical studies through data integration; here we focused 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 [59-63]. While many researchers have used machine learning to model neurotypical developmental trajectories to predict brain age, these predictors often struggle with generalizability across datasets due to site and scanner variability. We used brain age prediction as a case study to assess whether harmonization methods could 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. This evaluation was conducted on the TDC cohort (Data Configuration 3).

We first trained a support vector regression (SVR) model using structural connectomes from the PNC dataset to predict brain age. Features were extracted from the upper triangular portion of the connectivity matrices, with PCA applied for dimensionality reduction. The PNC dataset was randomly split 50:50 into a training set and a testing set. To optimize the model, we applied 5-fold cross-validation on the training set to tune hyperparameters. The final SVR model was trained on all subjects in the training set using the optimal hyperparameters. Model performance was evaluated using the testing set from the PNC dataset. To assess the model's generalizability, we tested 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 literature has identified several structural connectome patterns associated with ASD [2, 64-67]. However, most studies were conducted at single sites and often had small sample sizes, which limited their statistical power and ability to draw generalizable conclusions. Pooling data from multiple sources can introduce site-related biases that may obscure the detection of clinical associations. Therefore, harmonizing structural connectomes is essential for uncovering reliable ASD-related characteristics in multi-site studies.

To validate the effectiveness of harmonization algorithms, we applied site-specific coefficients, estimated from the TDC cohort (Data Configuration 3), to harmonize the ASD cohort (Data Configuration 4). Given the pronounced sex differences in autistic children and the small sample size of autistic females, we excluded female subjects from this analysis to minimize potential sex bias. We then calculated global graph topological measures of structural connectomes before and after harmonization, assessing group-level differences between the ASD and TDC groups, as well as group-by-age interactions in these measures. Additionally, we conducted the same group-level analyses on unharmonized data, with the site covariate regressed out at the level of graph measures and compared the results with those from the harmonized structural connectomes. We further performed Spearman correlation analysis between graph measures and ASD and developmental clinical assessments, including the Autism Diagnostic Observation Schedule (ADOS) calibrated severity scores, Social Responsiveness Scale (SRS-2), and Vineland Adaptive Behavior Scales (VABS), while controlling for age. Multiple comparisons were corrected using the False Discovery Rate (FDR). As the ComBat and CovBat methods may produce negative edge values, which lack physical relevance when computing graph topological measures in clinical contexts, we excluded these two methods from this analysis.

3. Results

3.1. Distributional Properties of Connectivity Strength in Connectomes

Each edge of the pre-harmonized structural connectomes was examined to determine whether it followed 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, showing only the edges with significant differences from the observed distributions ($p<0.05$, FDR-adjusted). Notably, the normality data assumptions underlying ComBat and CovBat were not met at most edges. Among the top 40% of highly consistent edges (1480 edges in total), 97% did not follow a normal distribution, 41% showed significant differences from the log-normal distribution, and only 5% were distributed significantly different from the gamma distribution fitting.

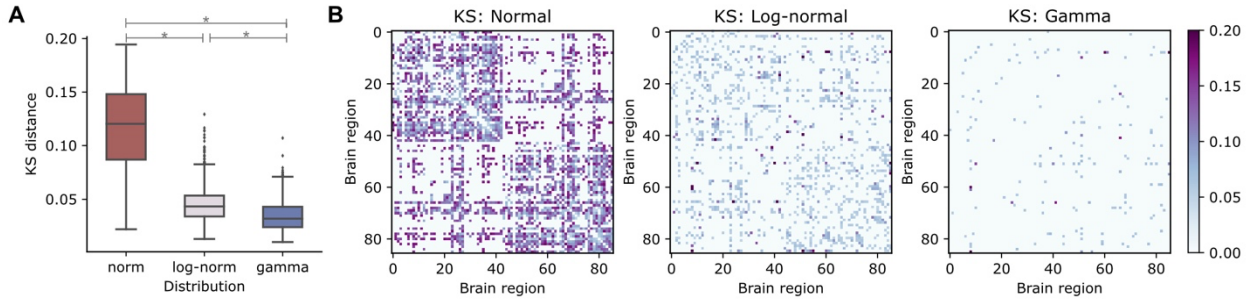

Figure 2 Goodness of fit test for PNC. A) Edgewise KS distances between the observed and three hypothesized distributions (normal, log-normal, and gamma) of connectivity strength in PNC. Asterisks denote significant differences between those hypothesized distributions (p < 0.0001, paired t-test). The gamma distribution provided a significantly better fit than the other two hypothesized models. B) Heatmaps of edgewise KS distances, showing only the 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. Structural connectome data were more likely to follow a gamma distribution, regardless of site or diagnostic group. We also evaluated the goodness of fit by varying different consistency threshold levels, as shown in Figure S2. For highly consistent edges, the gamma distribution provided a significantly better fit than the log-normal distribution ($p < 0.0001$, paired t-test). For less consistent edges (likely representing spurious connections), the fits of the log-normal and gamma distributions were comparable ($p > 0.05$, paired t-test). The normal distribution, however, resulted in significantly poorer fits compared to the other two distributions, regardless of the consistency level.

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. Figure 3A displays the MA-plot [68] of site differences between PNC and CAR. The MA-plot is a common method for visualizing differences in measurements taken from two settings, where 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. 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 Kruskal-Wallis H 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 obtained after consistency thresholding, 1035 (70%) edges initially showed significant differences in site means before harmonization ($p<0.05$, Kruskal-Wallis test). After ComBat harmonization, the number of edges with significant mean site differences dropped to just 11 edges (0.74%), and logComBat reduced the number to 5 edges (0.34%). After CovBat harmonization, only 3 edges (0.20%) showed significant site differences in means, and no significant site effects were observed with log-CovBat and gamma-GLM harmonization. For variance site effects, 706 edges (4.8%) exhibited significant site differences in pre-harmonization data ($p<0.05$, Brown-Forsythe test). After harmonization, 83 edges (5.6%) still showed significant site effects with ComBat, 79 edges (5.3%) remained significant with CovBat, and 4 edges (0.27%) with gamma-GLM. With log-ComBat and log-CovBat harmonization, all connections showed no statistically significant site-related effects on variance.

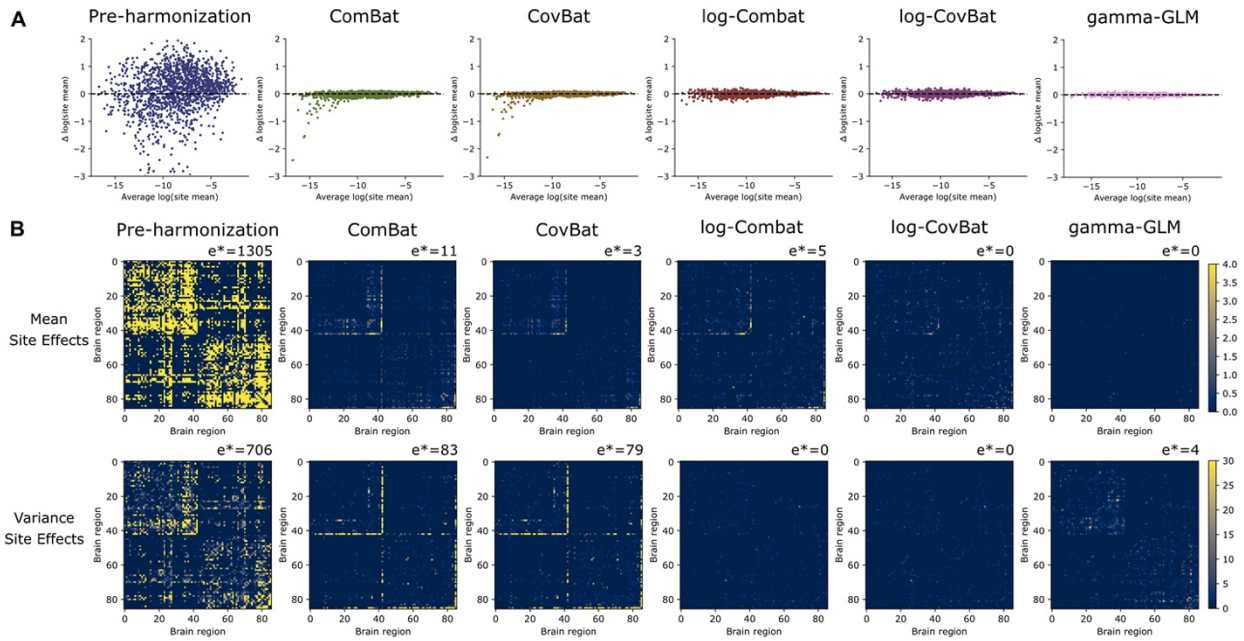

Figure 3 Site-related effects on mean and variance of edgewise strength in the paired PNC-CAR cohort (Data Configuration 1). A) MA-plots for visualization of site differences between PNC and CAR with paired subjects. The x-axis represents the averaged log-transformed means across sites and the y-axis 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 Kruskal-Wallis H statistics for mean site effects and F statistics from the Brown-Forsythe test for variance site effects. The number of edges with significant site effects was noted by e* in the top left corner of each plot.*

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 small effect size (Cohen's $d=0.30$). After ComBat harmonization of edgewise connections, 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.

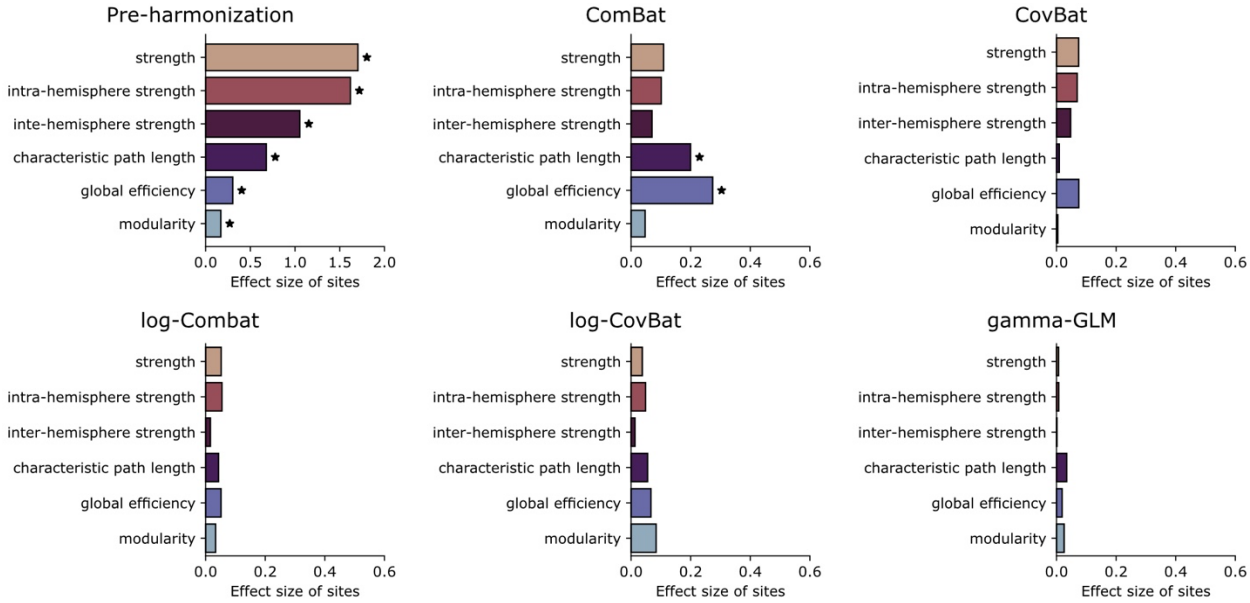

Figure 4 Effect size of site differences on global graph topological measures in the paired PNC-CAR cohort (Data Configuration 1). For each harmonization method, the Cohen's d effect sizes of sites were evaluated on six global graph topological measures (global strength, intra- and inter-hemisphere strength, 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).

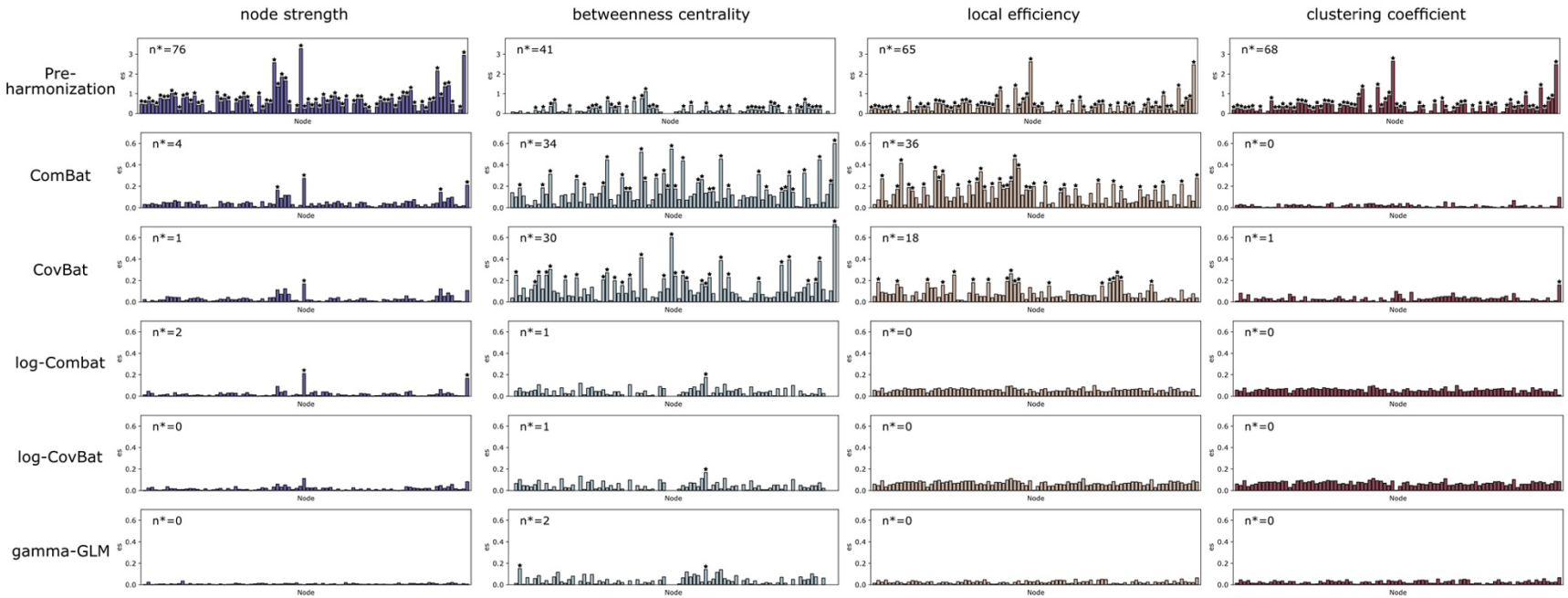

Figure 5 Effect size of site differences on nodal graph topological measures in the paired PNC-CAR cohort (Data Configuration 1). Four nodewise topological measures (node strength, betweenness centrality, local efficiency and clustering coefficient) were evaluated. Significant site effects ($p < 0.05$, two-sample t-test) were indicated by asterisks. The number of nodes with significant site effects was noted by n 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, Spearman 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 PNC (first row) and CAR (second row) with paired subjects (Data Configuration 1) before harmonization, displaying only edges with significant age associations ($p < 0.05$, Spearman R). The differences in edgewise correlation with ages between harmonized and pre-harmonization data were displayed for each harmonization approach. 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 confounds, 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 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	245	93.47%	92.65%	99.59%	97.14%	**100%**	179	79.89%	79.33%	97.77%	97.21%	**100%**
	CAR	223	73.54%	73.54%	90.13%	86.55%	**100%**	292	88.36%	89.73%	94.86%	94.52 %	**100%**
Confounded NYU-TCD	ABIDEII-NYU	16	75.00%	81.25%	87.50%	93.75%	**100%**	105	87.62%	76.19%	92.38%	83.81%	**100%**
	ABIDEII-TCD	35	80.00%	65.71%	94.29%	74.29%	**100%**	93	82.80%	67.74%	94.62.26%	91.40%	**100%**

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

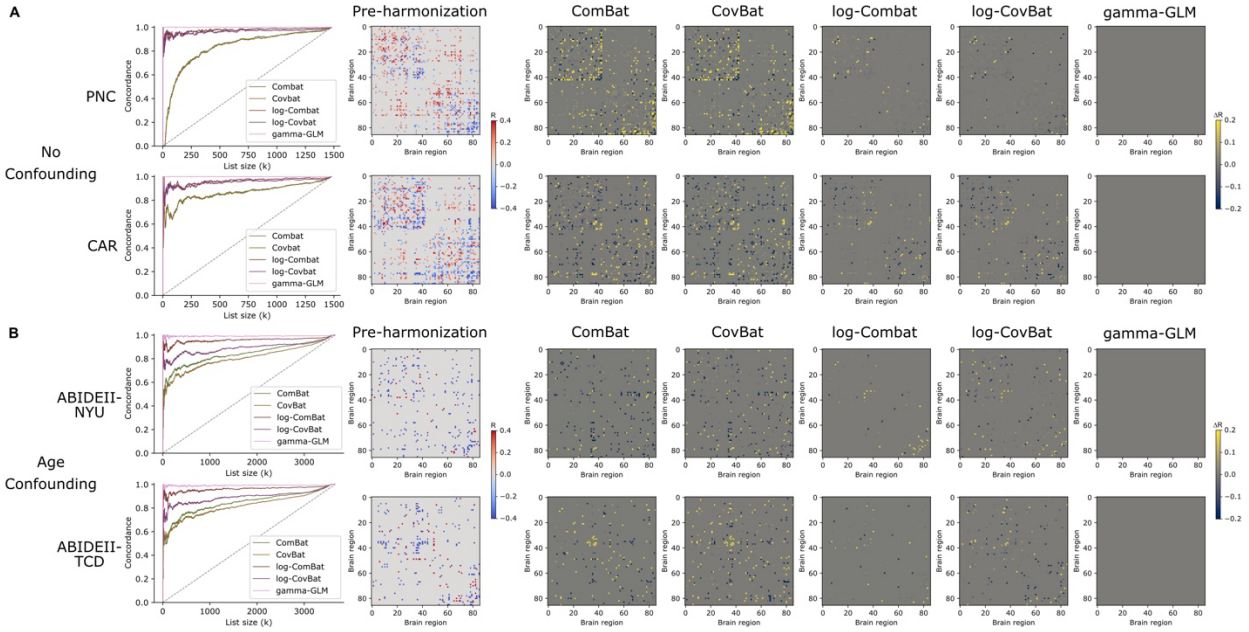

Figure 6 Visualization of edgewise age associations before and after applying different harmonization methods. Two scenarios were tested: (A) the paired PNC-CAR cohort (Data Configuration 1) and (B) the confounded NYU-TCD cohort (Data Configuration

2). For each scenario, edgewise correlations between connectivity strength and age were shown for each site before harmonization, displaying only significant age-associated connections (p < 0.05, Spearman R). The differences in edgewise correlations between harmonized data and pre-harmonization data were shown for each harmonization approach. 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 factors. 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 shown in the first row and ABIDEII-TCD in the second row. We noted that, compared to the paired cohort, the confounded cohort demonstrated different edgewise patterns between two sites. Since the two sites included participants from distinctly different age ranges, the age-related effects on structural connectomes may vary. 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 successfully recovered all edges with significant positive and negative correlations in presence of confounding factors between age and sites (Table 4).

We calculated the Frobenius distances between the edgewise age correlation patterns of two sites, as shown in Table 5. We found the Frobenius distances between two confounded sites were larger than those between two paired sites. 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 in both paired and confounded scenarios. In contrast, other harmonization approaches inadvertently reduced or increased the age-related differences across sites, likely due to poor fitting caused by an inappropriate choice of distributions for the 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
Paired PNC-CAR	14.4747	16.0879 (+1.6132)	16.1088 (+1.6341)	14.3714 (-0.1032)	14.3696 (-0.1051)	**14.4743 (-0.0004)**
Confounded NYU-TCD	19.4563	20.6006 (+1.1443)	20.2229 (+0.7666)	19.3874 (-0.0689)	19.5346 (+0.0783)	**19.4563 (-0.0000)**

Lastly, we evaluated the replicability of age associations when combining two sites. Figure S3 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. The gamma-GLM harmonization resulted in highest replicability regardless of age confounds.

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). We first repeated the previous evaluations to examine site effects on edgewise connectivity strength and graph topological measures, both before and after harmonization. The results are presented in Figure S3-S5 in the Supplementary Materials.

We trained a brain age predictor using an SVR model on the training set of PNC 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 performance of the trained brain age predictor on the testing set of PNC was provided (underlined). Our results demonstrated that using proper

harmonization to remove site-specific variations between PNC and new sites could largely improve the generalizability 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 (Lurie MEGlab, ABIDEII-TCD), it did not match gamma-GLM's performance in terms of Pearson R correlation.

Table 6 Impact of harmonization on the generalizability 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**
Lurie MEGlab	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 PNC (underlined). The brain age predictors are directly applied to other sites to evaluate the generalizability 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

In use case 2, we evaluated the effectiveness of harmonization methods in the pooled multi-site ASD cohort (Data Configuration 4) to examine their impact on detecting clinical group differences. Our findings indicated that prior to harmonization, the structural connectome data showed no significant group-level differences between ASD and TDC groups, nor any group-by-age interactions in global graph measures ($p>0.05$, FDR-adjusted). Additionally, none of the global graph measures demonstrated significant associations with clinical scores ($p>0.05$, FDR-adjusted). Post harmonization, we found that males with ASD showed significantly higher intra-hemispheric strength compared to the TDC males (Cohen's $d=0.37$, $p=0.041$, FDR-adjusted). Significant group-by-age interactions were observed for clustering coefficient (standardized $\boldsymbol{\beta}=0.27$, $p=0.005$, FDR-adjusted) and global efficiency (standardized $\boldsymbol{\beta}=0.26$, $p=0.006$, FDR-adjusted), with both measures increasing with age in ASD but decreasing in TDC. Intra-hemispheric strength showed a marginally significant correlation with the SRS-2 T-score ($R=0.12$, $p=0.072$, FDR-adjusted), while clustering coefficient ($R=0.15$, $p=0.038$, FDR-adjusted) and global efficiency ($R=0.14$, $p=0.038$, FDR-adjusted) were significantly and positively correlated with the VABS

communication subscale. However, ADOS and the two other domains of VABS (socialization and daily living) did not correlate with any global graph measures ($p>0.05$, FDR-adjusted).

4. Discussion

The scientific community is transitioning into the era of big data, where multiple research institutions and clinical centers collaborate to collect and maintain large-scale datasets. This study tackles the critical challenge of effectively merging multi-site structural connectome data without introducing site-related biases. Our harmonization framework provides a way to expand sample sizes, enabling a comprehensive representation of both healthy controls and patient populations to support more robust findings in the field. While previous studies introduced harmonization approaches at the level of dMRI signals [23, 26], they did not remove site effects at the level of the structural connectome [27], introducing bias in downstream graph analyses and other connectome-based investigations. Our study addressed the harmonization of the connectomes.

We provide practical guidance for researchers aiming to harmonize structural connectome data, following the pipeline outlined in Figure 1. Overall, we recommend the gamma-GLM approach as our top choice, as it outperformed other models in removing site effects while effectively preserving both demographic and disorder-related biological integrity throughout the harmonization process. In contrast, we found that harmonization techniques based on normal distributional assumption failed to remove site effects consistently for all connections in connectome data, potentially resulting in spurious findings in downstream analyses. Given the unique distributional properties of connectivity strength, it is essential to validate the underlying data assumptions when applying harmonization methods. Standard harmonization methods in neuroimaging, such as ComBat, CovBat, or simple linear regression, assume a normal distribution and are therefore unsuitable for structural connectomes, where most edges follow a gamma distribution. Moreover, ComBat and CovBat can yield negative values after harmonization, which lack physical meaning in structural connectomes, where edges represent (normalized) numbers of streamlines connecting brain regions. Applying these methods without careful consideration and validation [35] could lead to incorrect conclusions or unintended consequences in multi-site studies.

Notably, 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 [57]. Most prior work only evaluated graph measures at the global level [31, 35], but our results indicated that even though no site differences were observed in global measures, site differences can still exist at the nodal level. For example, while CovBat effectively removed site effects in all common global graph measures, substantial site effects persisted 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 prior to downstream analyses. 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.

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 essential that biologically relevant properties are not inadvertently altered or removed during the harmonization process. Given that age-related connectivity changes are critical in many studies of neurodevelopment and neurodegenerative processes [62, 69], we evaluated 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. The gamma-GLM model provides a better fit for covariate effects in structural connectome data, further emphasizing its advantage over other approaches for multi-site connectome data.

We demonstrated the practical utility of structural connectome harmonization with two use cases. The first use case addressed a commonly reported challenge in collaborative research: predictive models developed with one dataset often fail to predict unseen data from new sites [70, 71]. 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 generalizability of predictive models across sites. Here, we created machine learning models to predict structural connectome-based brain age using data from one site and evaluate the generalizability of these predictors on other sites. We showed that harmonizing structural connectomes across sites, particularly using the gamma-GLM based model, enhanced the generalizability of predictive models to unseen data at new sites. This approach can be further applied to a federated learning setup in healthcare [72, 73]. In this context, researchers can first harmonize and standardize decentralized data sources to a central site to ensure comparability, thereby improving further generalizability when deployed from and to the central site.

The second use case of harmonization is to increase the statistical power of case-control cohort studies via pooling structural connectomes from multiple sources. This approach is especially beneficial for clinical studies involving hard-to-recruit clinical populations, where data from a single site is often insufficient. By pooling patient data from various sources, researchers can create larger and more diverse cohorts, which are crucial for detecting subtle connectivity differences associated with various clinical conditions. To achieve this, the effectiveness and robustness of harmonization methods must be carefully evaluated within the context of the patient population. Our results, based on a clinical cohort of ASD participants and controls, demonstrated that without proper harmonization, site-related effects could obscure associations between graph topological measures and neurodivergent traits. While there is no definitive ground truth for *accuracy* in such measurements, harmonization allowed sufficient improvement in estimation *precision* by mitigating noise introduced by site and scanner differences. In the present data, harmonization with gamma-GLM effectively resolved significant differences in structural connectomes between diagnostic groups that were previously undetectable.

Despite such a comprehensive set of experiments evaluating the pros and cons of different frameworks for connectome harmonization, there are several issues that warrant attention in future research. Users should note that the gamma-GLM approach does not model site-specific variance and covariance. When image acquisition and reconstruction protocols vary significantly across sites, careful checks are necessary to ensure that such variations do not affect the data properties [55]. Although we did not observe significant site-related differences in covariance in this study, these effects may emerge in subnetwork analyses or when structural connectomes are modulated by diffusion metrics like fractional anisotropy (FA) or mean diffusivity (MD). While we tackled the most used definition of structural connectomes and tested common distributions from the exponential family, researchers creating connectomes using more complex methods may need to explore additional distributions (e.g., zero-adjusted gamma distribution) and other statistical metrics. In such cases, more flexible distributional regression models, like the Generalized Additive Model for Location, Scale, and Shape (GAMLSS) [74], 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 [75, 76]. Our harmonization and evaluation framework can also be adapted to better address the specific goals of structural connectome studies in these contexts.

5. Conclusion

In summary, we outlined a path forward for harmonizing multi-site structural connectome datasets, thereby facilitating collaborative clinical research. Given the complex nature of connectomes, it is essential to examine the distributional properties of the connectome data. The appropriate harmonization framework, such as gamma-GLM or more flexible distributional regression models, should be selected to ensure that the integrity of the data is maintained post-harmonization. Additionally, all harmonization results should be carefully assessed for site effects at different levels of measures, such as edgewise connectivity strength and global and nodal-level graph properties, to ensure an unbiased interpretation of the data in downstream connectome studies.

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. Onicas, A.I., et al., Multisite harmonization of structural DTI networks in children: An A-CAP study. Frontiers in Neurology, 2022. 13: p. 850642.
36. Satterthwaite, T.D., et al., Neuroimaging of the Philadelphia Neurodevelopmental Cohort. NeuroImage, 2014. 86: p. 544-553.
37. 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.
38. 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.
39. Manjón, J.V., et al., Diffusion weighted image denoising using overcomplete local PCA. PloS one, 2013. 8(9): p. e73021.
40. Andersson, J.L. and S.N. Sotiropoulos, An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. Neuroimage, 2016. 125: p. 1063-1078.
41. Smith, S.M., Fast robust automated brain extraction. Human brain mapping, 2002. 17(3): p. 143-155.
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. Tournier, J.-D., et al., MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation. Neuroimage, 2019. 202: p. 116137.
45. Dhollander, T., et al. Improved white matter response function estimation for 3-tissue constrained spherical deconvolution. in Proc. Intl. Soc. Mag. Reson. Med. 2019.
46. 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.
47. 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.
48. 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, NJ.
49. 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.
50. Fischl, B., et al., Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron, 2002. 33(3): p. 341-355.
51. Roberts, J.A., et al., Consistency-based thresholding of the human connectome. NeuroImage, 2017. 145: p. 118-129.
52. McCullagh, P., Generalized linear models. 2019: Routledge.
53. 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.

54. Keene, O.N., The log transformation is special. Statistics in medicine, 1995. 14(8): p. 811-819.
55. 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.
56. 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.
57. Rubinov, M. and O. Sporns, Complex network measures of brain connectivity: uses and interpretations. Neuroimage, 2010. 52(3): p. 1059-1069.
58. Irizarry, R.A., et al., Multiple-laboratory comparison of microarray platforms. Nature methods, 2005. 2(5): p. 345-350.
59. 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.
60. Liem, F., et al., Predicting brain-age from multimodal imaging data captures cognitive impairment. Neuroimage, 2017. 148: p. 179-188.
61. Dosenbach, N.U., et al., Prediction of individual brain maturity using fMRI. Science, 2010. 329(5997): p. 1358-1361.
62. Bethlehem, R.A., et al., Brain charts for the human lifespan. Nature, 2022. 604(7906): p. 525-533.
63. 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.
64. Rane, P., et al., Connectivity in autism: a review of MRI connectivity studies. Harvard review of psychiatry, 2015. 23(4): p. 223-244.
65. 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.
66. 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.
67. Lewis, J.D., et al., Network inefficiencies in autism spectrum disorder at 24 months. Translational psychiatry, 2014. 4(5): p. e388-e388.
68. Dudoit, S., et al., Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica sinica, 2002: p. 111-139.
69. 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.
70. Goetz, L., et al., Generalization—a key challenge for responsible AI in patient-facing clinical applications. npj Digital Medicine, 2024. 7(1): p. 126.
71. Rajpurkar, P., et al., AI in health and medicine. Nature medicine, 2022. 28(1): p. 31-38.
72. 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.
73. 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.
74. 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.
75. Norori, N., et al., Addressing bias in big data and AI for health care: A call for open science. Patterns, 2021. 2(10).
76. 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.