DeepResBat: deep residual batch harmonization accounting for covariate distribution differences

Lijun An[1,2,3], Chen Zhang[1,2,3], Naren Wulan[1,2,3], Shaoshi Zhang[1,2,3], Pansheng Chen[1,2,3], Fang Ji[1], Kwun Kei Ng[1], Christopher Chen[4], Juan Helen Zhou[1,2,5], B.T. Thomas Yeo[1,2,3,5,6], Alzheimer's Disease Neuroimaging Initiative* & Australian Imaging Biomarkers and Lifestyle Study of Aging*

[1] Centre for Sleep and Cognition (CSC) & Centre for Translational Magnetic Resonance Research (TMR), Yong Loo Lin School of Medicine, National University of Singapore, Singapore [2] Department of Electrical and Computer Engineering, National University of Singapore, Singapore [3] N.1 Institute for Health & Institute for Digital Medicine (WisDM), National University of Singapore, Singapore [4] Department of Pharmacology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore [5] NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore [6] Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA, USA

Address correspondence to:

B. T. Thomas Yeo
CSC, TMR, ECE, N.1 & WISDM
National University of Singapore
Email: thomas.yeo@nus.edu.sg

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) and the Australian Imaging Biomarkers and Lifestyle Study of Aging (AIBL) database (www.aibl.csiro.au). As such, the investigators within the ADNI and AIBL contributed to the design and implementation of ADNI and AIBL and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Abstract

Pooling MRI data from multiple datasets requires harmonization to reduce undesired inter-site variabilities, while preserving effects of biological variables (or covariates). The popular harmonization approach ComBat uses a mixed effect regression framework that explicitly accounts for covariate distribution differences across datasets. There is also significant interest in developing harmonization approaches based on deep neural networks (DNNs), such as conditional variational autoencoder (cVAE). However, current DNN approaches do not explicitly account for covariate distribution differences across datasets. Here, we provide mathematical results, suggesting that not accounting for covariates can lead to suboptimal harmonization outcomes. We propose two DNN-based harmonization approaches that explicitly account for covariate distribution differences across datasets: covariate VAE (coVAE) and DeepResBat. The coVAE approach is a natural extension of cVAE by concatenating covariates and site information with site- and covariate-invariant latent representations. DeepResBat adopts a residual framework inspired by ComBat. DeepResBat first removes the effects of covariates with nonlinear regression trees, followed by eliminating site differences with cVAE. Finally, covariate effects are added back to the harmonized residuals. Using three datasets from three different continents with a total of 2787 participants and 10085 anatomical T1 scans, we find that DeepResBat and coVAE outperformed ComBat, CovBat and cVAE in terms of removing dataset differences, while enhancing biological effects of interest. However, coVAE hallucinates spurious associations between anatomical MRI and covariates even when no association exists. Therefore, future studies proposing DNN-based harmonization approaches should be aware of this false positive pitfall. Overall, our results suggest that DeepResBat is an effective deep learning alternative to ComBat.

1 Introduction

There is growing interest in combining MRI data across multiple sites, such as ENGIMA (Thompson et al., 2017) and ABCD (Volkow et al., 2018) studies. These so-called mega-analyses significantly advance neuroimaging research by increasing statistical power (Bethlehem et al., 2022; Marek et al., 2022), enhancing generalizability (He et al., 2022; Lu et al., 2022), and detecting subtle effects (Vogel et al., 2021; Tian et al., 2023). When pooling data across datasets, post-acquisition harmonization is necessary for removing undesirable variabilities across datasets, while preserving relevant biological information. A major source of undesirable cross-dataset heterogeneity is scanner differences across datasets (Magnotta et al., 2012; Chen et al., 2014; Hawco et al., 2018). In addition, the distributions of biological variables (e.g., demographics and clinical diagnosis) may also vary across datasets. These biological variables (also referred to as "covariates") can have a large impact on MRI data (Hua et al., 2010), whose effects should be preserved after harmonization.

A popular approach for harmonizing MRI data is via mixed effects modeling, such as ComBat (Fortin et al., 2017, 2018; Yu et al., 2018). ComBat removes additive and multiplicative site differences while including biological variables as covariates. For example, to perform a mega-analysis using several Alzheimer's Disease (AD) dementia datasets, the ComBat model might be set up with hippocampal volume as the dependent variable, site as an independent variable, as well as age, sex and clinical diagnosis as covariates. Additive and multiplicate site effects are removed from the hippocampal volume, while the residual effects of age, sex and clinical diagnoses are retained. Several ComBat variants have been proposed to enhance harmonization performance (Garcia-Dias et al., 2020; Pomponio et al., 2020; Wachinger et al., 2021). However, most ComBat variants harmonize brain regions separately, limiting their ability to eliminate nonlinear site differences spanning the brain regions.

Deep neural networks (DNNs) are promising for eliminating nonlinear site differences distributed across the brain (Hu et al., 2023). Variational autoencoder (VAE)-based approaches (Moyer et al., 2020; Russkikh et al., 2020; Zuo et al., 2021; An et al., 2022) use an encoder to generate site-invariant latent representations from input MRI data. Site information is then concatenated to the latent representations to reconstruct the MRI data via a decoder. Generative adversarial networks (Dewey et al., 2019; Zhao et al., 2019; Modanwal et al., 2020; Bashyam et al., 2021), normalizing flow (Wang et al., 2021; Beizaee et al., 2023) and federated learning (Dinsdale et al., 2022) have also been explored. However, existing

DNN approaches typically overlook the inclusion of covariates, which are explicitly controlled in mixed effects harmonization models (Fortin et al., 2017, 2018; Chen et al., 2021). Since covariate distribution differences are unavoidable across datasets, neglecting covariates during harmonization can inadvertently remove relevant biological information, instead of reducing undesired dataset differences, leading to worse downstream performance. In Section 2.1, we show how a theoretical machine learning result (Tachet et al., 2020) can be used to understand this phenomenon.

In this study, we propose two deep learning approaches: covariate VAE (coVAE) and deep residual batch effects harmonization (DeepResBat), which account for covariate distribution differences across datasets. coVAE extends conditional VAE (cVAE; Moyer et al., 2020) by concatenating covariates and site information with site- and covariate-invariant latent representations. On the other hand, DeepResBat adopts a residual framework inspired by the classical ComBat approach. DeepResBat first removes the effects of covariates using nonlinear regression trees, followed by eliminating unwanted site differences from the residuals with cVAE. Finally, covariate effects are added back to the harmonized residuals. We found that coVAE hallucinated spurious associations between anatomical MRI and covariates even when no association existed, suggesting that DNN-based harmonization approaches can introduce false positives during harmonization. On the other hand, DeepResBat effectively mitigated this false positive issue.

The contributions of this study are multi-fold. First, we showed theoretically that ignoring covariate differences across datasets can lead to suboptimal harmonization outcomes. Second, we introduced a DNN-based harmonization approach DeepResBat that could account for covariate differences across datasets. DeepResBat outperformed ComBat (Fortin et al., 2017), CovBat (Chen et al., 2021) and cVAE (Moyer et al., 2020) across multiple evaluation experiments, including enhancing biological effects of interest, while removing unwanted dataset differences. Third, we demonstrated for the first time that DNN-based harmonization approaches could potentially hallucinate relationships between covariates and MRI measurements even when none existed. Therefore, future studies proposing DNN-based harmonization approaches should be aware of this false positive pitfall. Although the current study focused on MRI data, our results are generally applicable to any field where instrumental harmonization is necessary, e.g., molecular biology (Johnson et al., 2007) and climate science (Iturbide et al., 2019).

2 Methods

2.1 Motivation for accounting for covariates during harmonization

Distribution differences in covariates, such as demographics and clinical diagnoses, across datasets are inevitable. Most deep learning harmonization approaches directly align distributions of latent representations across datasets without explicitly modeling covariate differences (Dewey et al., 2019; Zuo et al., 2021; Beizaee et al., 2023; Liu et al., 2023). Without explicitly accounting for these covariates, the covariates differences can be misinterpreted as undesirable dataset differences and wrongly removed by the harmonization algorithms. Here, we will formalize this phenomenon using a theoretical result from the machine learning literature (Tachet et al., 2020).

More specifically, suppose we have two datasets S and T with target label Y and input data X. The goal is to predict Y using X. Suppose we have feature extractor g that takes in X as the input with feature representation Z as the output. The feature representation Z is then entered into classifier h to predict the target label. Let $\epsilon_S(h \circ g)$ and $\epsilon_T(h \circ g)$ be the expectation of classification errors when applying g followed by h to datasets S and T respectively. Let Y_S and Y_T be the target label distributions in datasets S and T respectively. Let Z_S and Z_T be the distributions of the feature representation in datasets S and T respectively. Then, Tachet des Combes and colleagues show that the following inequality is true:

$$\epsilon_S(h \circ g) + \epsilon_T(h \circ g) \geq \frac{1}{2}\left(\sqrt{\mathrm{JS}(Y_S \parallel Y_T)} - \sqrt{\mathrm{JS}(Z_S \parallel Z_T)}\right)^2, \quad (1)$$

where $\mathrm{JS}(\cdot \,||\, \cdot)$ is the Jensen-Shannon divergence of two distributions. We note that the lowest possible error bound is zero. Therefore, assuming distributional differences between the target labels of the two datasets (i.e., $\mathrm{JS}(Y_S \parallel Y_T) > 0$), then a lower bound of zero can only be achieved if the same distribution differences exist between the feature representations of the two datasets, i.e., $\mathrm{JS}(Z_S \parallel Z_T) = \mathrm{JS}(Y_S \parallel Y_T)$. In other words, dataset-invariant representations (i.e., $\mathrm{JS}(Z_S \parallel Z_T) = 0$) lead to suboptimal classification performance (greater than zero error bound in Eq. (1)) if there exists distributional differences between the target labels of the two datasets ($\mathrm{JS}(Y_S \parallel Y_T) > 0$).

To relate Eq. (1) to harmonization, we can think of g as a harmonization procedure (instead of a feature extractor), and h as a downstream task to predict covariates Y after harmonization. Therefore, if the distributions of covariates Y are different across datasets (i.e., $\mathrm{JS}(Y_S \parallel Y_T) > 0$), blindly matching the distributions of brain imaging measures across

datasets in the harmonization process (i.e., $\mathrm{JS}(Z_S \parallel Z_T) = 0$) is suboptimal for the downstream task, i.e., error bound in Eq. (1) is greater than 0.

For example, suppose we would like to harmonize two datasets with different distributions of healthy elderly participants and participants with Alzheimer's disease (AD) dementia. Then, it is important to account for these distributional differences when harmonizing the datasets. This is typically not an issue for mixed effects harmonization approaches (such as ComBat and CovBat) since covariates are typically explicitly included. However, most deep learning approaches do not account for covariate distribution differences between datasets, which can potentially result in suboptimal downstream task performance.

2.2 Datasets and preprocessing

In this study, we proposed DNN models for harmonizing T1 anatomical MRI data. We will test the models using data from separate research initiatives: the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack et al., 2008; Jack et al., 2010), the Australian Imaging, Biomarkers and Lifestyle (AIBL) study (Ellis et al., 2009, 2010) and the Singapore Memory Ageing and Cognition Centre (MACC) Harmonization cohort (Hilal et al., 2015; Chong et al., 2017; Hilal et al., 2020). All data collection and analysis procedures were approved by the respective Institutional Review Boards (IRBs), including the National University of Singapore IRB for the analysis presented in this paper. All three datasets encompass a range of modalities collected at multiple timepoints, such as MRI scans, cognition assessments, and clinical diagnoses.

We utilized ADNI1 and ADNI2/Go data from ADNI (Jack et al., 2008; Jack et al., 2010). For ADNI1, the MRI scans were collected from 1.5 and 3T scanners from different vendors (more details in Table S1). For ADNI2/Go, the MRI scans were acquired on 3T scanners. A total of 1,735 participants underwent at least one T1 MRI scan, resulting in 7,955 scans scanned at multiple timepoints. 68 cortical and 40 subcortical regions of interest (ROI) were defined based on FreeSurfer (Fischl et al., 2002; Desikan et al., 2006). The volumes of the cortical and subcortical ROIs were provided by ADNI using multiple preprocessing steps (http://adni.loni.usc.edu/methods/mri-tool/mri-pre-processing/), followed by the FreeSurfer version 4.3 (ADNI1) and 5.1 (ADNI2/Go) recon-all pipelines, yielding a total of 108 volumetric measures.

In the case of AIBL (Ellis et al., 2009, 2010), the MRI scans were collected from 1.5T and 3T Siemens (Avanto, Tim Trio, and Verio) scanners (see Table S2 for more information). There were 495 participants with at least one T1 MRI scan, resulting in 933

MRI scans across multiple timepoints. The FreeSurfer 6.0 recon-all pipeline was employed to extract the volumes from 108 cortical and subcortical ROIs.

In the case of MACC (Hilal et al., 2015; Chong et al., 2017; Hilal et al., 2020), the MRI scans were collected from a Siemens 3T Tim Trio scanner. There were 557 participants with at least one T1 MRI scan. There were 1197 MRI scans across the different timepoints of the 557 participants. Similar to AIBL, we utilized the FreeSurfer 6.0 recon-all pipeline to extract the volumes of 108 cortical and subcortical ROIs.

2.3 Workflow

To compare different harmonization approaches, we harmonized brain ROI volumes between ADNI and AIBL, as well as ADNI and MACC. Figure 1 illustrates the workflow in this study using AIBL as an example. The procedure was the same for harmonizing ADNI and MACC.

Following our previous study (An et al., 2022), the Hungarian matching algorithm (Kuhn, 1955) was first applied to match a subset of participants with similar age, sex, MMSE, and clinical diagnosis distribution between ADNI and AIBL datasets (Figure 1A). The distributions before and after matching are shown in Figure S1. When matching the ADNI and AIBL datasets, we obtained 257 matched participant pairs. When matching the ADNI and MACC datasets, 277 matched participant pairs were obtained. Notably, not all time points had corresponding MMSE and clinical diagnosis information. Therefore, care was taken to ensure that all timepoints in the matched participants had both MMSE and clinical diagnosis. We ensured that each participant's scans were all categorized as either "matched" or "unmatched" without splitting the participant's scans across categories. The quality of the matching procedure was assessed through statistical tests, whose p values are reported in Tables S3 to S9. All p values were greater than 0.05.

The resulting unmatched participants were used to train and tune various harmonization approaches (Figure 1B). After model fitting, the trained harmonization models were applied to harmonize matched and unmatched participants. Three evaluation experiments were performed (Figure 1C). First, as a common evaluation practice (Hu et al., 2023), a machine learning model was trained to investigate whether harmonization could effectively reduce dataset differences by predicting which dataset a participant came from (more details in Section 2.8). Second, we evaluated whether harmonization led to stronger associations between harmonized ROI volumes with the covariates of interest (more details in Section 2.9). Finally, an exhaustive false-positive permutation test was carried out, involving

240,000 GPU hours and 360,000 CPU hours. This test aimed to assess whether deep harmonization models might introduce spurious associations (i.e., false positives) between anatomical MRI and randomly permuted covariates when no association exists (more details in Section 2.10). In evaluations where training was necessary (dataset prediction and false positive experiments), unmatched participants were used as the training and validation sets for the evaluation experiments, while the matched participants were used as the test set.

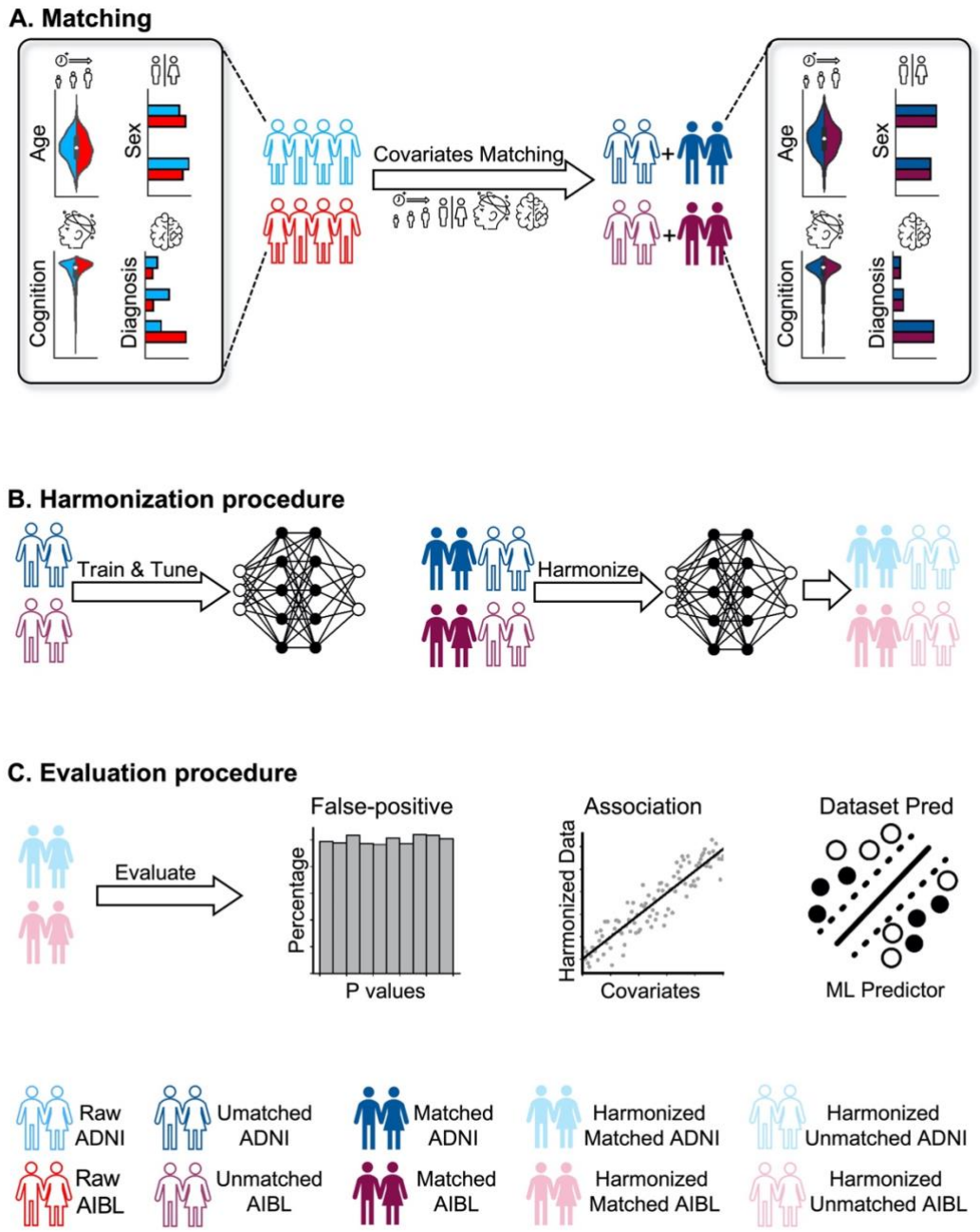

Figure 1. Workflow of study. We illustrate the workflow using ADNI and AIBL. The same procedure was applied to ADNI and MACC. (A) ADNI and AIBL participants were matched based on age, sex, mini mental-state examination (MMSE) and clinical diagnosis. The unmatched participants were used for training and tuning harmonization and evaluation

models. The matched participants served as the test set for harmonization evaluation. (B) Left: Train harmonization models with unmatched participants. Right: Apply trained harmonization models on both unmatched and matched participants. (C) Three sets of evaluation experiments were tested on matched harmonized participants: dataset prediction experiment, association analysis and false positive experiment. In evaluations where training was necessary (dataset prediction and false positive experiments), unmatched participants were used as the training and validation sets for the evaluation experiments, while the matched participants were used as the test set.

2.4 Training, validation and test procedure

As described in section 2.3, the test set for evaluation (Figure 1C) consisted of matched participants. On the other hand, the unmatched participants were utilized for training both harmonization (Figure 1B) and evaluation models (e.g., dataset prediction and false positive experiments in Figure 1C).

In the case of mixed effects models (ComBat and CovBat), there is no hyperparameter, so the data of all unmatched participants were used to fit the models. On the other hand, for the deep harmonization models, we have to tune the hyperparameters. Therefore, we divided the unmatched participants into 10 distinct groups. Notably, all timepoints belonging to a participant were assigned to a single group, thus avoiding any splitting of timepoints of a participant across different groups. For the training and tuning of the deep harmonization models on unmatched participants, a 9-1 train-validation split was employed, with 9 groups used for training and one group used as the validation set for hyperparameter tuning. This process was repeated 10 times, with each group serving as the validation set once. Subsequently, in the harmonization step (Figure 1B), we obtained 10 sets of trained harmonization models. These models were then applied to the unharmonized data, resulting in the generation of 10 harmonized data sets from both unmatched and matched participants.

In the subsequent evaluation step (Figure 1C), the harmonized data of the unmatched participants were used to train and tune the evaluation models for dataset prediction (Section 2.8) and false positive analysis (Section 2.10), following the same 9-1 train-validation split previously described. The harmonized data of the matched participants were designated as the test set to evaluate the harmonization performance. In the case of association analysis (Section 2.9), no evaluation model needed to be trained, so we directly applied general linear models (GLMs) and multivariate analysis of variance (MANOVA) to the harmonized data of the matched participants to obtain association results.

2.5 Baseline harmonization models

Here, we considered ComBat (Johnson et al., 2007), CovBat (Chen et al., 2021), and cVAE (Moyer et al., 2020) as baseline models.

2.5.1 ComBat

ComBat is a mixed effects model that controls for additive and multiplicative site effects (Johnson et al., 2007). Here we utilized the R implementation of the algorithm (https://github.com/Jfortin1/ComBatHarmonization). The ComBat model is as follows:

$$x_{ijv} = \alpha_v + Y_{ij}^T \beta_v + \gamma_{iv} + \delta_{iv}\epsilon_{ijv}\ , \tag{2}$$

where i is the site index, j is the participant index, and v indexes the brain ROI volumes. x_{ijv} is the volume of the v-th brain ROI of participant j from site i. γ_{iv} is the additive site effect. δ_{iv} is the multiplicative site effect. ϵ_{ijv} is the residual error term following a normal distribution with zero mean and variance δ_v^2. Y_{ij} is the vector of covariates of participant j from site i. In this study, we chose age, sex, MMSE, and clinical diagnosis as covariates.

The ComBat parameters α_v, β_v, γ_{iv} and δ_{iv} were estimated for each brain region using the unharmonized ROI volumes of all unmatched participants (Figure 1B). The estimated parameters can then be applied to map a new participant i from site j to intermediate space with brain regional volume x_{ijv} and covariates Y_{ij}.

$$x_{ijv}^{ComBat} = \frac{x_{ijv} - \hat{\alpha}_v - Y_{ij}^T \hat{\beta}_v - \hat{\gamma}_{iv}}{\hat{\delta}_{iv}} + \hat{\alpha}_v + Y_{ij}^T \hat{\beta}_v, \tag{3}$$

where ˆ indicates that the parameter was estimated from the *unmatched unharmonized* ROI volumes from ADNI and AIBL. A separate ComBat model was fitted for harmonizing ADNI and MACC brain regional volumes.

2.5.2 CovBat

CovBat is a mixed effect harmonization model built on top of ComBat to remove site effects in mean, variance, and covariance (Chen et al., 2021). We utilized the authors' R implementation of the algorithm (https://github.com/andy1764/CovBat_Harmonization). There are four main steps in CovBat harmonization. First, ComBat (Section 2.5.1) is applied to the volume of each brain region. to obtain ComBat adjusted residuals:

$$e_{ijv}^{ComBat} = \frac{x_{ijv} - \hat{\alpha}_v - Y_{ij}^T \hat{\beta}_v - \hat{\gamma}_{iv}}{\hat{\delta}_{iv}}, \tag{4}$$

For participant i of site j, the ComBat-adjusted residuals of all regional volume can be concatenated into a column vector e_{ij}^{ComBat}. These vectors can in turn be concatenated across all participants of all sites into the matrix E^{Combat}, where the number of columns is equal to the total number of participants and the number of rows is equal to the number of brain regional volume.

Second, principal component analysis (PCA) is applied to E^{Combat} to obtain q principal component (PC) scores and PCs, where q is the rank of the matrix E^{Combat}. Therefore, we can write e_{ij}^{ComBat} as

$$e_{ij}^{ComBat} = \sum_{k=1}^{q} \xi_{ijk} \hat{\phi}_k \,, \tag{5}$$

where ξ_{ijk} is the k-th PC score of participant i of site j, and $\hat{\phi}_k$ is k-th PC.

Third, each of the top K PC scores were harmonized using a second round of ComBat, thus yielding $\hat{\xi}_{ijk}$. K was selected to explain the 95% percentage of variance. The remaining PC scores were not harmonized. Finally, the harmonized brain ROI volumes are projected to intermediate space after model fitting:

$$e_{ij}^{CovBat} = \sum_{k=1}^{K} \hat{\xi}_{ijk} \hat{\phi}_k + \sum_{k=K}^{q} \xi_{ijk} \hat{\phi}_k \,, \tag{6}$$

$$x_{ijv}^{CovBat} = e_{ijv}^{CovBat} + \hat{\alpha}_v + Y_{ij}^T \hat{\beta}_v, \tag{7}$$

Consistent with ComBat, we chose age, sex, MMSE, and clinical diagnosis as covariates. CovBat was fitted using all unmatched participants.

For a new participant, e_{ijv}^{ComBat} was computed using the ComBat parameters estimated from the unmatched participants. The e_{ij}^{ComBat} of the new participant was then projected onto the principal components (obtained from the unmatched particiapnts) to obtain principal component scores ξ_{ijk}. The top K scores were then harmonized using the second round of ComBat parameters estimated from the unmatched participants to obtain $\hat{\xi}_{ijk}$. Finally, Equations (6) and (7) were applied to obtain the harmonized ROI volumes of the new participant.

Similar to ComBat, two separate CovBat models were fitted: one for harmonizing ADNI and AIBL brain regional volumes, and one for harmonizing ADNI and MACC brain regional volumes.

2.5.3 cVAE

Moyer et al. (2020) introduced the conditional variational autoencoder (cVAE) model for harmonizing diffusion MRI data. In this paper, we applied the cVAE model to harmonize brain ROI volumes. Figure 2A shows the architecture of the cVAE model. The input brain volumes x (x denotes brain volumes of all regions: $x_1, x_2 \dots x_v, \dots x_{108}$) were processed through an encoder deep neural network (DNN) to obtain the latent representation z. The one-hot site vector s was concatenated with the latent representation z and then fed into the decoder DNN, producing the reconstructed brain volumes $\hat{x}$. To encourage the independence of the learned representation z from the site s, the cost function incorporated the mutual information $I(z, s)$. The resulting loss function could be expressed as follows:

$$L_{cVAE} = L_{recon} + \alpha L_{prior} - \gamma L_{adv} + \lambda I(z, s), \quad (8)$$

where L_{recon} was the mean square error (MSE) between x and $\hat{x}$, thus encouraging similarity between the harmonized $\hat{x}$ and unharmonized volumes x. Additionally, Moyer et al. introduced the term L_{adv}, which was the soft-max cross-entropy loss of an adversarial discriminator aiming to differentiate x and $\hat{x}$, thereby further promoting their similarity. Finally, L_{prior} was the KL divergence term between representation z and a multivariate Gaussian distribution with zero mean and identity covariance matrix (Sohn et al., 2015) to promote regularity and control over the latent space. This prior term will encourage the distributions of latent representations to be aligned across datasets, which can be problematic if there exists covariate differences between datasets (Section 2.1).

The decoder and encoder components of the model were implemented as fully connected feedforward neural networks, where each layer was connected to the subsequent layer. Consistent with Moyer et al., the tanh activation function (Maas et al., 2013) was employed. During training, the variable s represented the true site information associated with the input brain volumes x. After training, by setting s to zero, the input x could be mapped to an intermediate space. Training of the model was performed using the data from 90% of the unmatched participants and hyperparameters were tuned using the data from 10% of the unmatched participants as validation set (Section 2.4).

Hyperparameter tuning in the validation set involved optimizing a weighted sum of the reconstruction loss (MSE between x and $\hat{x}$) and the accuracy of participant-level dataset prediction: ½ MSE + Dataset Prediction Accuracy. The reconstruction loss (MSE) was halved to ensure comparability with the dataset prediction accuracy. Dataset prediction accuracy was determined by training an XGBoost classifier on the training set and evaluating

it on the validation set. To identify the best set of hyperparameters, the HORD algorithm (Eriksson et al., 2020; Regis & Shoemaker, 2013; Ilievski et al., 2017) was employed using the validation set (Table 1). The trained deep neural network (DNN) was then utilized for subsequent analyses after 1000 epochs of training.

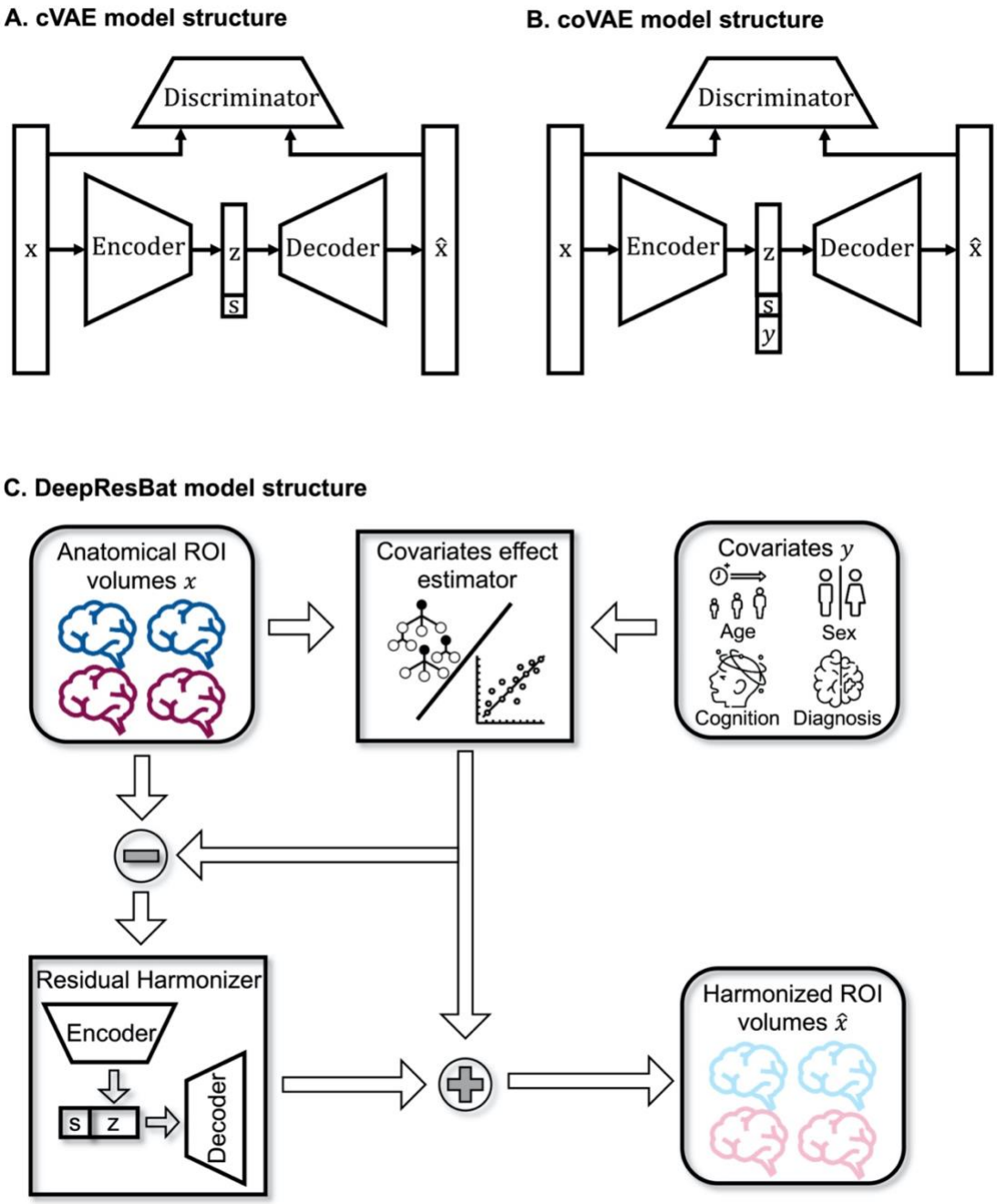

Figure 2. Model structure for cVAE, coVAE, and DeepResBat. (A) Model structure for the cVAE model. Encoder, decoder, and discriminator were all fully connected feedforward DNNs. s was the site we wanted to map the brain volumes to. (B) Model structure for the coVAE model. Site s and covariates y were input into the decoder to preserve covariates effects. Therefore, the main difference between cVAE and coVAE is the inclusion of covariates y. (C)

Model structure for the DeepResBat model. The covariates effect estimator was an ensemble of XGBoost and linear models. Once the effects of covariates were removed (subtraction sign), the residual harmonizer was a cVAE model taking covariates-free residuals as input. The covariates effects were then added back to the cVAE output, yielding a final set of harmonized ROI volumes.

Hyperparameter	Search Range
Initial learning rate	1e-2 – 1e-1
Learning rate step	10 - 999
Dropout rate	0 – 0.5
α	0.01 - 1
γ	0.01 - 10
λ	0.01 - 1
Nodes for each layer	32 - 512
Number of layers	2 - 4
Node for z	32 - 512

Table 1. Hyperparameters search ranges for cVAE on the validation set. We note that a learning rate decay strategy was utilized. After K training epochs (where K = learning rate step), the learning rate was reduced by a factor of 10.

2.6 coVAE

As highlighted in Section 2.1, accounting for covariates is essential for effective harmonization. Therefore, we extended the cVAE model to incorporate covariates as input, resulting in the covariate-VAE (coVAE) model, as shown in Figure 2B. More specifically, we concatenated site s and covariates Y to obtain $[s, Y]$. The loss function was the same as cVAE (Eq. (8)) except that the mutual information loss term in Eq. (8) was modified to become $I(z, [s, Y])$:

$$L_{coVAE} = L_{recon} + \alpha L_{prior} - \gamma L_{adv} + \lambda I(z, [s, Y]). \quad (9)$$

Therefore, instead of minimizing the mutual information between z and s, we minimize the mutual information between z and $[s, Y]$. For the reconstruction, we concatenated latent representation z with site s and covariates Y as input to the decoder network.

Recall that the L_{prior} term was the KL divergence term between representation z and a multivariate Gaussian distribution with zero mean and identity covariance matrix. This prior term therefore implicitly encouraged the alignment of the latent representations z between datasets. In the case of cVAE, the latent representation z would contain covariate information. Therefore, cVAE would force the alignment of latent representations z even if

the covariate distributions were different across datasets, which would be suboptimal (see Section 2.1). By contrast, because coVAE seek to minimize mutual information between z and $[s, Y]$, so the latent representation z would theoretically not contain covariate information. In this scenario, aligning the covariate-free and site-free latent representation z would make sense.

Consistent with ComBat (Section 2.5.1) and CovBat (Section 2.5.2), we chose age, sex, MMSE, and clinical diagnosis as covariates. The categorical covariates sex and clinical diagnosis were one-hot encoded for coVAE. Training of coVAE was performed using the data from 90% of the unmatched participants and hyperparameters were tuned using the data from 10% of the unmatched participants as validation set (Section 2.4). We used the HORD algorithm to search within the hyper-parameter ranges specified in Table 1 based on the validation set. Brain ROI volumes were mapped to intermediate space for subsequent analyses after 1000 training epochs.

Although coVAE appeared to be an intuitive straightforward extension of cVAE, as will be shown in the false positive analysis (Section 3.3), coVAE suffered from significant false positive rates. Therefore, in the next section, we proposed a second harmonization approach that could account for covariate distribution differences across datasets.

2.7 DeepResBat

Figure 2C illustrates our proposed DeepResBat approach. To motivate DeepResBat, we can write ComBat's Eq. (2) into a more general form:

$$x_{ijv} = f_v(Y_{ij}) + g_v(i), \tag{10}$$

where i is the site index, j is the participant index, and v indexes the brain ROI volumes. x_{ijv} is the v-th brain volume of participant j from site i. Y_{ij} are the covariates of participant j from site i. In ComBat, f_v is linear with $f_v(Y_{ij}) = \alpha_v + Y_{ij}^T \beta_v$, while $g_v(i) = \gamma_{iv} + \delta_{iv}\epsilon_{ijv}$ accounts for both additive and multiplicative site effects.

To improve on ComBat, DeepResBat utilized nonlinear functions for f and g. There are three stages for DeepResBat (Figure 2C). We first estimated the covariate effects f using a nonlinear regression approach (Section 2.7.1). The covariate-free residuals from the first stage ($x - f$) were then harmonized using a generic deep learning approach, instantiated as cVAE in the current study (Section 2.7.2). The covariate effects from the first stage were then added back to the harmonized brain volumes from the second stage (Section 2.7.3).

Consistent with ComBat, we chose age, sex, MMSE, and clinical diagnosis as covariates. Training of DeepResBat was performed using the data from 90% of the unmatched participants and hyperparameters were tuned using the data from 10% of the unmatched participants as validation set (Section 2.4). We used the HORD algorithm to search within the hyper-parameter ranges specified in Table 1 based on the validation set. Brain ROI volumes were mapped to intermediate space for subsequent analyses after 1000 training epochs.

2.7.1 Covariates effects estimation

To estimate covariate effects, we first regressed out the linear effects of site for each brain ROI volume by fitting the following model,

$$x_{ijv} = \alpha_v + \gamma_{iv} + \epsilon_{ijv}, \tag{11}$$

where x_{ijv} is the v-th brain ROI volume for participant i of site j, α_v is the intercept term, γ_{iv} is the additive sites effect, and ϵ_{ijv} is the residual error term. The residual brain ROI volume $\tilde{x}_{ijv}$ without linear site effects is obtained by $\tilde{x}_{ijv} = x_{ijv} - \hat{\gamma}_{iv}$.

Before removing covariate effects, we first checked whether each covariate is actually related to any ROI brain volume in order to avoid the false positive issues exhibited by coVAE (Section 2.6). This check is performed in two stages. First, for each covariate, an XGBoost model (T. Chen & Guestrin, 2016) was trained to predict the covariate using all residual brain ROI volumes $\tilde{x}$ obtained from the previous step (Eq. (11)). XGBoost was chosen due to its efficacy with unstructured or tabular data (Grinsztajn et al., 2022; Shwartz-Ziv & Armon, 2022) and its simplicity, allowing for fast training. For each covariate, the XGBoost model was trained using the training set (90% of unmatched participants) and hyperparameters were tuned using the validation set (10% of unmatched participants). We randomly sampled 50% of participants from the validation set and computed the correlation between the prediction and ground truth covariate. Pearson's correlation and Spearman's correlation were used for continuous and discrete covariates respectively. This sampling procedure was repeated 100 times. If the p values of the correlations were less than 0.05 for more than 95% of the repetitions, then we retained the covariate for the next stage.

In the second stage, for each brain volume $\tilde{x}_v$, an XGBoost model was trained to predict the brain volume using all survived covariates $\tilde{Y}$. Once again, the training used the training set (90% of unmatched participants) and hyperparameters were tuned using the validation set (10% of unmatched participants). To ensure, we were not overfitting the

covariate estimator, we again sampled 50% of participants from the validation set and computed the correlation between the prediction and ground truth covariate. Pearson's correlation was used for evaluating brain ROI volumes' predictions. This sampling procedure was repeated 100 times. If the p values of the correlations were less than 0.05 for more than 95% of the repetitions, we retained the XGBoost model. If not, we fitted a linear model instead.

Therefore, regardless of whether we ended up using a linear model or XGBoost, we obtained the covariates effect estimator $f_v(\tilde{Y})$ for each brain region v. The estimated covariates effects could then be subtracted from the original brain ROI volume, yielding covariate-free residuals:

$$r_v = x_v - f_v(\tilde{Y}), \tag{12}$$

The residuals r_v were presumably free from covariate effects, but retained unwanted variations from each dataset, which could be removed with a generic deep learning based harmonization approach in the next stage (Section 2.7.2).

2.7.2 Covariate-free residuals harmonization

In the second stage of DeepResBat, the covariate-free residuals r_v were jointly fed into a deep learning based harmonization model $g(\cdot)$ for further harmonization. In the current study, we chose the cVAE model, although any deep learning harmonization model could be used. Similar to the cVAE baseline (Section 2.5.3), the cVAE was trained using the training set (90% of unmatched participants) and hyperparameters (Table 1) were tuned using the validation set (10% of unmatched participants) with the HORD algorithm.

Following training, the covariates-free residuals were mapped to an intermediate space:

$$\hat{r} = g(r), \tag{13}$$

where r was the covariate-free residuals and $\hat{r}$ was the harmonized residual brain volumes.

2.7.3 DeepResBat harmonization

The final harmonized brain ROI volumes were then obtained by adding the estimated covariates effect from stage 1 and harmonized residual from stage 2 for each brain ROI volume:

$$\hat{x}_v = \hat{r}_v + f_v(\tilde{Y}), \tag{14}$$

where $\hat{r}_v$ was the harmonized residuals (Section 2.7.2) and $f_v(\tilde{Y})$ was the estimated covariate effects (2.7.1).

2.8 Dataset prediction model

As an evaluation metric, we employed XGBoost to predict the source dataset of the harmonized brain volumes (Figure 1C). The inputs to the XGBoost model were the brain volumes normalized by each participant's total intracranial volume (ICV). Due to the 10-fold cross-validation procedure described in Section 2.4, recall that the unmatched participants were divided into 10 groups of training and validation sets. Therefore, in the case of cVAE, coVAE and DeepResBat, there were 10 harmonization models and 10 sets of harmonized data for each participant. In the case of ComBat and CovBat, the models were fitted on all unmatched participants, so there was only one set of harmonized data for each participant.

For each group of training and validation sets, an XGBoost classifier was trained using the training set and a grid search was conducted on the validation set to identify the optimal hyperparameters. To evaluate performance, the 10 XGBoost classifiers were used to predict the source dataset of the harmonized MRI volumes of the matched participants.

The prediction accuracy was calculated by averaging the results across all time points of each participant and the 10 classifiers before further averaging across participants. To evaluate the harmonization quality between ADNI and AIBL, this evaluation procedure was applied to the ADNI and AIBL participants. The same procedure was applied to evaluate the harmonization quality between ADNI and MACC datasets. Lower prediction accuracies indicated that greater dataset differences were removed, suggesting better harmonization quality.

2.9 Association analysis

Lower dataset prediction accuracies (Section 2.8) indicate greater dataset differences were removed, but the removed dataset differences might contain important biological information, which should not be removed. Therefore, association analysis was also conducted to evaluate the ability of preserving relevant biological information (Figure 1C) during harmonization. The variables of interest included age, sex, clinical diagnosis, and MMSE. Univariate and multivariate association analyses were performed on matched participants using unharmonized or harmonized regional brain volumes from different approaches. As mentioned in previous sections, brain ROI volumes were harmonized by

mapping to intermediate space. For the association analysis, we used the Python package *statsmodels*.

We again reminded the reader that in the case of cVAE, coVAE and DeepResBat, there were 10 harmonization models and 10 sets of harmonized data for each participant. Therefore, the 10 sets of harmonized data were averaged for each matched participant before the association analysis was performed. In the case of ComBat and CovBat, there was only one set of harmonized data for each matched participant, so no averaging was necessary.

2.9.1 Univariate association analysis

Of the 108 brain regions, 87 regions are grey matter ROIs. We analyzed the associations between the 87 grey matter ROI volumes and covariates with GLM. Grey matter volumes are well-studied biomarkers correlated to age, sex, cognition, and AD dementia (Hutton et al., 2009; Hua et al., 2010; Blessed et al., 2018; van de Mortel et al., 2021).

For each grey matter brain ROI volume and each harmonization approach, GLM models were fitted to evaluate the association between brain ROI volume and covariates. We fitted GLM separately for clinical diagnosis and MMSE to avoid cofounding. The GLM formulas were as follows:

$$\hat{x}_v \sim Age + Sex + MCI + AD \ + ICV, \tag{15}$$

and

$$\hat{x}_v \sim Age + Sex + MMSE \ + ICV, \tag{16}$$

where $\hat{x}_v$ was the v-th harmonized brain ROI volume, ICV was the estimated total intracranial volume. Z statistics for GLM's betas were utilized as evaluation metrics.

2.9.2 Multivariate association analysis

Multivariate analysis of variance (MANOVA) was applied to evaluate the multivariate association between 87 grey matter ROIs and covariates. As in Section 2.9.1, we conducted MANOVA separately for clinical diagnosis and MMSE. The fitted models were as follows:

$$\hat{x} \sim Age + Sex + MCI + AD \ + ICV, \tag{17}$$

and

$$\hat{x} \ \sim Age + Sex + MMSE \ + ICV, \tag{18}$$

where $\hat{x}$ referred to all harmonized grey matter brain ROI volumes and ICV was the estimated total intracranial volume.

2.10 False positive analysis

To evaluate whether the harmonization approaches will hallucinate associations between covariates and harmonized ROI volumes when none exists, we performed a permutation analysis to evaluate false positive rates. For example, if we permuted age across participants, then the resulting harmonized brain ROI volumes should not associate with the permuted age.

More specifically, when harmonizing ADNI and MACC, for each permutation, we randomly shuffled four covariates (age, sex, MMSE, and clinical diagnosis) together across (1) unmatched participants in the training set, (2) unmatched participants in the validation set and (3) matched participants. Harmonization models were then trained to harmonize brain ROI volumes based on the randomly shuffled covariates. As stated in Section 2.4, unmatched participants were used for training and tuning the harmonization models, and GLMs were performed in the matched harmonized participants.

We expect the association between the randomly permuted covariates and harmonized brain ROI volumes to not exist. Therefore, we ran GLMs to validate our assumption via association analysis. The GLMs were run for each harmonized brain ROI volume with randomly shuffled covariates on matched participants. We considered 87 grey matter ROIs (see Section 2.9.1) to run the GLM. A diagnosis GLM ($\hat{x}_v \sim Age + Sex + MCI + AD + ICV$) and a cognition GLM ($\hat{x}_v \sim Age + Sex + MMSE + ICV$) were fitted separately. Then for each permutation and each harmonized brain ROI volume, we obtained p values from the GLM corresponding to each covariate.

This permutation procedure was repeated 1000 times. For each brain ROI volume harmonized by each harmonization model, we calculate the percentage of nominally significant p values below 0.05 across the 1000 p values from the 1000 permutations. The expected percentage across all grey matter ROIs is 5%, with a confidence interval (CI) of 3.65% to 6.35% based on the normal approximation of the Binomial 95% CI (Eklund et al., 2016). Percentage higher than 6.35% indicated that there were false positives. The same procedure was repeated for ADNI and AIBL datasets.

2.11 Deep neural network implementation

The DNNs developed in this paper were implemented using PyTorch (Paszke et al., 2017) and executed on NVIDIA RTX 3090 GPUs with CUDA 11.0. The DNNs were

optimized using the Adam optimizer (Kingma & Ba, 2017) with the default settings provided by PyTorch.

2.12 Statistical tests

To assess distribution differences in age and MMSE between matched participants of AIBL and ADNI (as well as MACC and ADNI), two-sided two-sample t-tests were employed. For sex and clinical diagnoses, chi-squared tests were utilized to examine any significant distinctions.

In the case of dataset prediction, the prediction performance was averaged over all time points of each participant and then across the 10 sets of models, resulting in a single prediction performance value for each participant. Therefore, this process yielded a vector of prediction performance for each dataset and harmonization approach, with each element corresponding to a particular participant. To compare the dataset prediction performance between the two harmonization approaches, a permutation test with 10,000 permutations was conducted. Each permutation involved randomly exchanging the entries between the performance vectors of the two approaches. A more detailed illustration of this permutation procedure can be found in Figure S2.

Multiple comparisons were corrected with a false discovery rate (FDR) of $q < 0.05$.

2.13 Data and code availability

Code for the various harmonization algorithms can be found here (GITHUB_LINK). Two co-authors (CZ and PC) reviewed the code before merging it into the GitHub repository to reduce the chance of coding errors.

The ADNI and the AIBL datasets can be accessed via the Image & Data Archive (https://ida.loni.usc.edu/). The MACC dataset can be obtained via a data-transfer agreement with the MACC (http://www.macc.sg/).

3 Results

3.1 DNN models removed more dataset differences than classical mixed effect models

Dataset prediction accuracies of matched participants are shown in Figure 3. Lower prediction accuracies indicated that greater dataset differences were removed, suggesting better harmonization quality.

Figure 3A shows the dataset prediction performance for matched ADNI and AIBL participants. Without harmonization, an XGBoost classifier achieved 100% accuracy in identifying which dataset a participant's data came from. After applying mixed effect harmonization approaches (ComBat and CovBat), the prediction accuracy significantly dropped to 0.695 ± 0.376 (mean ± std) for ComBat, and 0.670 ± 0.383 for CovBat, indicating a substantial reduction in dataset differences. Deep learning approaches showed improved dataset difference removal, with performance of 0.595 ± 0.381 for cVAE and 0.592 ± 0.368 for coVAE. Our proposed DeepResBat achieved an accuracy of 0.594 ± 0.375, which was not statistically different from the deep learning baselines (Table 2). Notably, all deep learning approaches exhibited significantly lower dataset prediction accuracies than mixed effect approaches, demonstrating the potential of deep learning for data harmonization. However, the dataset prediction accuracies of all deep learning approaches remained better than chance ($p = 1e-4$), indicating residual dataset differences.

Similar outcomes were observed for matched ADNI and MACC participants (Figure 3B). Without harmonization, the XGBoost classifier accurately predicted source datasets with 100% accuracy. Mixed effect approaches, including ComBat and CovBat, reduced dataset differences to some extent, yielding accuracies of 0.752 ± 0.361 for ComBat and 0.740 ± 0.370 for CovBat. All deep learning approaches (cVAE: 0.603 ± 0.391, coVAE: 0.558 ± 0.418, DeepResBat: 0.608 ± 0.381) exhibited more effective removal of dataset differences than mixed effect approaches (Table 3). Our proposed DeepResBat achieved similar accuracy to cVAE with no statistical difference, but performed worse than coVAE, indicating room for improvement. However, coVAE introduced significant false positives (as will be shown in Section 3.3) and is therefore not an acceptable approach. Finally, the dataset prediction accuracies of all deep learning approaches remained better than chance ($p = 1e-4$), indicating the presence of residual dataset differences.

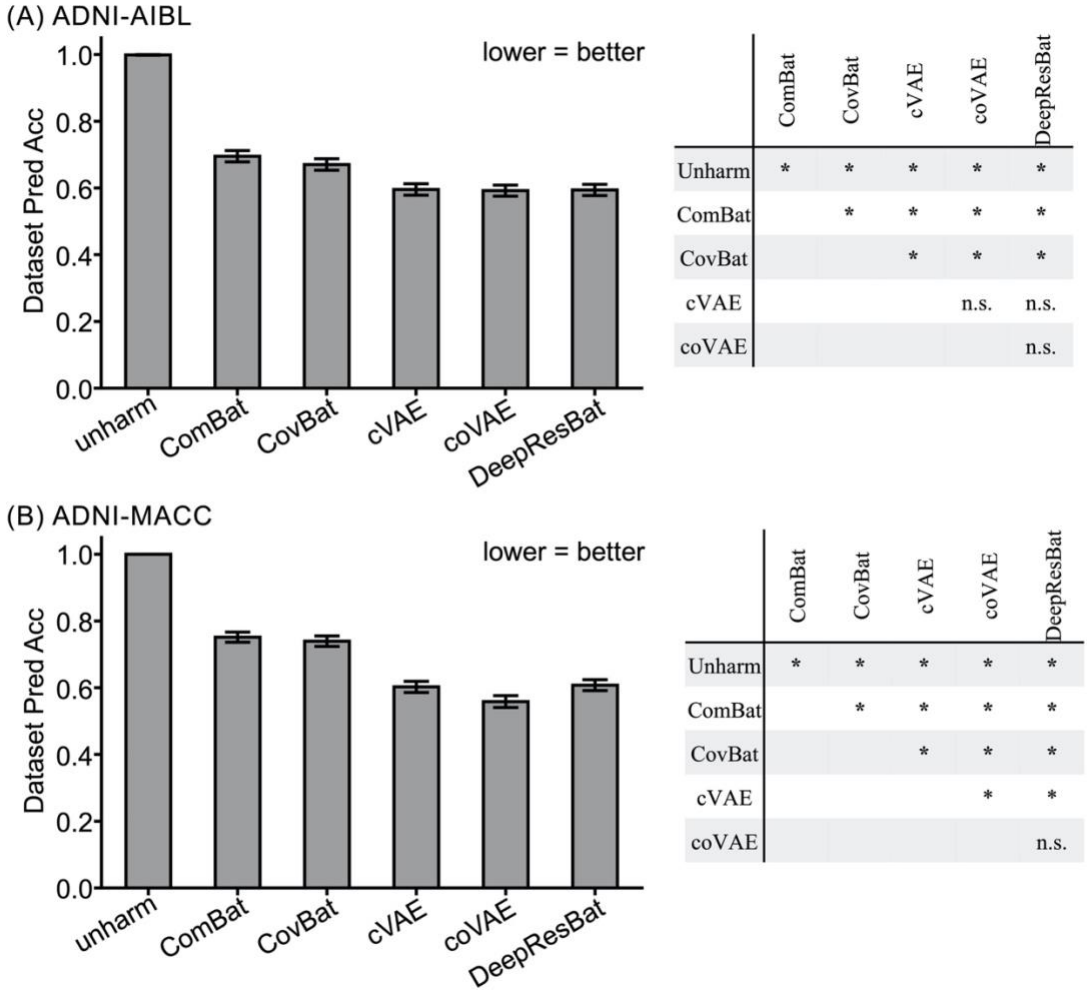

Figure 3. Dataset prediction accuracies. (A) Left: Dataset prediction accuracies for matched ADNI and AIBL participants. Right: p values of differences between different approaches. "*" indicates statistical significance after surviving FDR correction ($q < 0.05$). "n.s." indicates not significant. (B) Same as (A) but for matched ADNI and MACC participants. All p values are reported in Tables 2 and 3.

Dataset Prediction Accuracies (mean ± std)	p values					
	Unharm	ComBat	CovBat	cVAE	coVAE	DeepResBat
Unharmonized (1.000 ± 0.020)		**1e-4**	**1e-4**	**1e-4**	**1e-4**	**1e-4**
ComBat (0.695 ± 0.376)			**8e-4**	**1e-4**	**1e-4**	**1e-4**
CovBat (0.670 ± 0.383)				**1e-4**	**1e-4**	**1e-4**
cVAE (0.595 ± 0.381)					0.5945	0.8931
coVAE (0.592 ± 0.368)						0.8303
DeepResBat (0.594 ± 0.375)						

Table 2. Dataset prediction accuracies with p values of differences between different approaches for matched ADNI and AIBL participants. Statistically significant p values after FDR ($q < 0.05$) corrections are bolded.

Dataset Prediction Accuracies (mean ± std)	p values					
	Unharm	ComBat	CovBat	cVAE	coVAE	DeepResBat
Unharmonized (1.000 ± 1e-16)		**1e-4**	**1e-4**	**1e-4**	**1e-4**	**1e-4**
ComBat (0.752 ± 0.361)			**0.0095**	**1e-4**	**1e-4**	**1e-4**
CovBat (0.740 ± 0.370)				**1e-4**	**1e-4**	**1e-4**
cVAE (0.603 ± 0.391)					**1e-4**	0.7258
coVAE (0.558 ± 0.418)						**2e-4**
DeepResBat (0.608 ± 0.381)						

Table 3. Dataset prediction accuracies with p values of differences between different approaches for matched ADNI and MACC participants. Statistically significant p values after FDR ($q < 0.05$) corrections are bolded.

3.2 DeepResBat enhanced associations between harmonized brain volumes and covariates

Lower dataset prediction accuracies (Section 3.1) indicate greater dataset differences were removed, but the removed dataset differences might contain important biological information, which should not be removed. To evaluate whether relevant biological information is retained in the harmonization process, we performed univariate GLM and multivariate MANOVA analyses to evaluate the associations between harmonized volumes and covariates. Stronger associations between harmonized volumes and covariates suggest better enhancement of biological information after harmonization.

3.2.1 DeepResBat outperformed baselines for univariate analysis

Figures 4 and 5 show the results of the univariate GLM association analysis involving clinical diagnosis (Eq. (15)) in the ADNI-AIBL and ADNI-MACC matched participants respectively. Each dot in the plots represented a different brain region, so there are 87 dots in total. For age, MCI and AD dementia, more negative z values indicated greater atrophy due to aging and AD progression. Conversely, a lower MMSE indicates worse cognition, so more positive z values indicated greater atrophy related to worse cognition. For sex, the absolute z statistics were compared because there was no a priori expectation of positive or negative values, so a larger magnitude indicating a larger effect size. Therefore, in Figures 4 and 5, red indicates better performance by DeepResBat, while blue indicates worse performance.

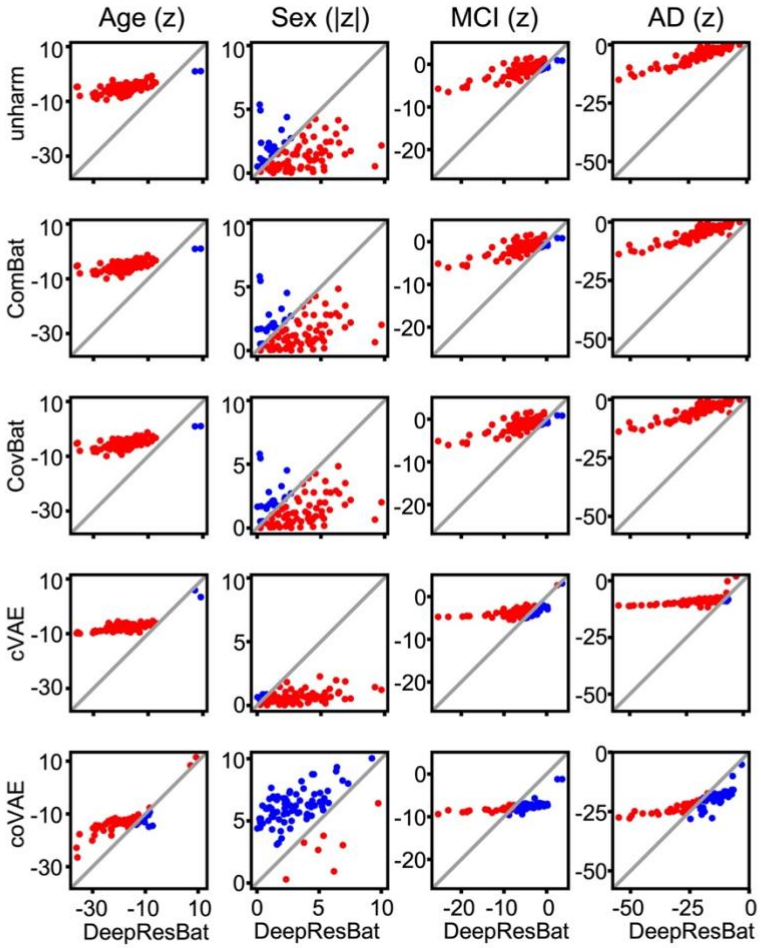

Figure 4. Comparison of z statistics from GLM involving clinical diagnosis for DeepResBat and baselines on matched ADNI and AIBL participants. Each row compares DeepResBat and one baseline approach: no harmonization (row 1), ComBat (row 2), CovBat (row 3), cVAE (row 4) and coVAE (row 5). Each column represents one covariate: age (column 1), sex (column 2), MCI (column 3) and AD dementia (column 4). Each subplot compares z statistics of DeepResBat against another baseline for a given covariate across 87 grey matter ROIs. Each dot represents one grey matter ROI. Red dots indicate better performance by DeepResBat. Blue dots indicate worse performance by DeepResBat.

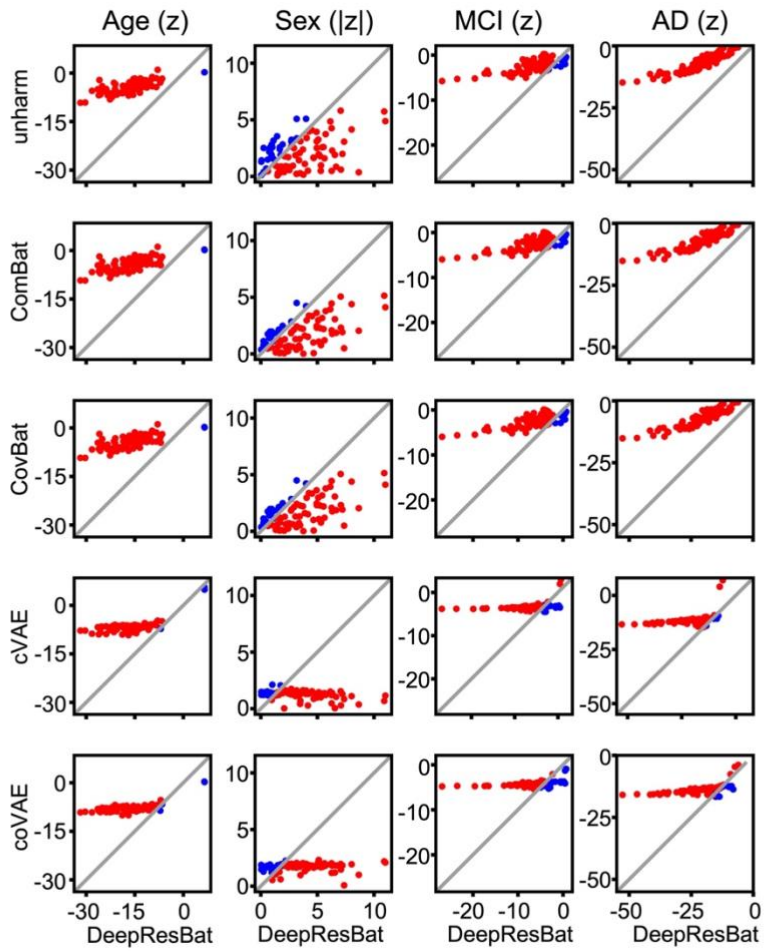

Figure 5. Comparison of z statistics from GLM involving clinical diagnosis for DeepResBat and baselines on matched ADNI and MACC participants. Each row compares DeepResBat and one baseline approach: no harmonization (row 1), ComBat (row 2), CovBat (row 3), cVAE (row 4) and coVAE (row 5). Each column represents one covariate: age (column 1), sex (column 2), MCI (column 3) and AD dementia (column 4). Each subplot compares z statistics of DeepResBat against another baseline for a given covariate across 87 grey matter ROIs. Each dot represents one grey matter ROI. Red dots indicate better performance by DeepResBat. Blue dots indicate worse performance by DeepResBat.

In the case of matched ADNI-AIBL participants, DeepResBat yielded stronger associations between brain volumes and all covariates with respect to no harmonization (Figure 4 row 1), ComBat (Figure 4 row 2), CovBat (Figure 4 row 3) and cVAE (Figure 4 row 4). When compared to coVAE, DeepResBat yielded weaker absolute z statistics for sex, similar z statistics for MCI and AD, and better z statistics for age. However, coVAE introduced false positives (as will be shown in Section 3.3) and is therefore not an acceptable approach. In the case of matched ADNI-MACC participants, DeepResBat yielded stronger associations between brain volumes and all covariates with respect to no harmonization (Figure 4 row 1) and all other baseline approaches (Figure 4 rows 2 to 5).

Similar conclusions were obtained for the univariate GLM association analyses involving MMSE (Eq. (16)) in the ADNI-AIBL (Figure S3) and ADNI-MACC (Figure S4) matched participants.

3.2.2 DeepResBat outperformed baselines for multivariate analysis

Figure 6 shows the results of the multivariate MANOVA association analysis involving clinical diagnosis (Eq. (17)) in the ADNI-AIBL and ADNI-MACC matched participants. The negative logarithm of p values was employed as the metric, where a larger negative logarithm of p values indicates a stronger association. Therefore, a higher bar in Figure 6 indicates better performance.

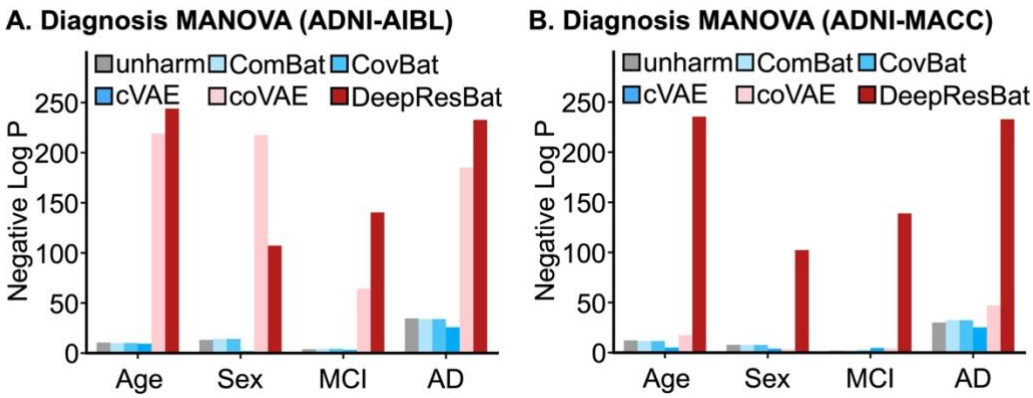

Figure 6. Significance bar plot by MANOVA involving clinical diagnosis. A larger negative of log p value indicates a stronger association, and thus better performance. (A) Bar plot for matched ADNI and AIBL participants. (B) Bar plot for matched ADNI and MACC participants.

In the case of matched ADNI-AIBL participants (left panel in Figure 6), DeepResBat yielded stronger associations between brain volumes and all covariates with respect to no harmonization, ComBat, CovBat and cVAE. When compared to coVAE, DeepResBat yielded weaker p values for sex, but stronger p values for age, MCI and AD. However, coVAE introduced false positives (as will be shown in Section 3.3) and is therefore not an acceptable approach. In the case of matched ADNI-MACC participants (right panel in Figure 6), DeepResBat yielded stronger associations between brain volumes and all covariates with respect to no harmonization and all other baseline harmonization approaches.

Interestingly, coVAE demonstrated inconsistent behavior for association with sex across different cohorts. In the ADNI-AIBL analyses, coVAE harmonized ROIs exhibited a strong association with sex (Figure 6A). By contrast, in the ADNI-MACC analyses, the association with sex was notably weak (Figure 6B). Conversely, DeepResBat displayed more

stable performance across cohorts. The inconsistency of coVAE was also present in the univariate GLM analyses (Figures 4 and 5).

Similar conclusions were obtained for the multivariate MANOVA association analyses involving MMSE (Eq. (18)) in the ADNI-AIBL and ADNI-MACC matched participants (Figure S5).

3.3 CoVAE, but not DeepResBat, exhibited spurious associations between permuted covariates and harmonized brain volumes

The harmonization models were retrained after permuting all four covariates: age, sex, diagnosis and MMSE (Section 2.10). The GLM association analyses (Section 2.9.1) were then rerun. Figure 7 illustrates the percentage of nominally significant p values (i.e., $p < 0.05$) across 87 grey matter ROIs from 1000 permutations. More specifically, for each harmonized brain ROI volume, we calculated the percentage of nominally significant p values (i.e., $p < 0.05$) across 1000 p values corresponding to the 1000 permutations. The expected percentage of nominally significant p values across all grey matter ROIs should be 5%, with a 95% confidence interval (CI) of 3.65% to 6.35% based on the normal approximation of the Binomial 95% CI (Eklund et al., 2016).

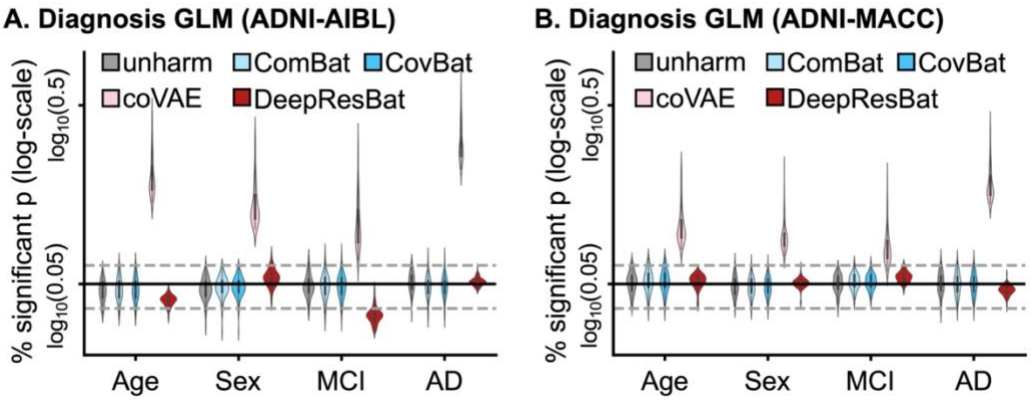

Figure 7. Percentage of nominally significant p values (i.e., $p < 0.05$) from GLM with clinical diagnosis after 1000 permutations of covariates. More specifically, each data point in the violin plot represents a brain ROI volume. Percentage is calculated based on the number of permutations in which p value of corresponding covariate was nominally significant (i.e., $p < 0.05$) divided by 1000 permutations. Percentage (vertical axis) is shown on a log scale. The black solid line is the expected percentage (which is 0.05), while the grey dashed lines indicated 95% confidence intervals. (A) GLM analysis involving clinical diagnosis for matched ADNI and AIBL participants. (B) GLM analysis involving clinical diagnosis for matched ADNI and MACC participants.

In both ADNI-AIBL (Figure 7A) and ADNI-MACC (Figure 7B), not performing any harmonization yielded 5% nominally significant p values. ComBat, CovBat and DeepResBat also did not suffer from any spurious associations. However, coVAE yielded an inflated false positive rate in both ADNI-AIBL and ADNI-MACC, since the percentages of nominally significant p values were much greater than 5% for all covariates (Figure 7).

To provide a different visualization of the results, Figure 8 shows the frequency distribution of p values for coVAE and DeepResBat in matched ADNI and AIBL participants. Each line in Figure 8 corresponded to a single brain ROI. The p values were divided into bins with a width of 0.05. Therefore, in the ideal scenario, the distribution of p values should follow a uniform distribution at a height of 0.05. For coVAE (Figure 8A) the frequency of p values within the 0-0.05 bin greatly exceeded 5% for all covariates. For DeepResBat (Figure 8B), there was a uniform distribution of p values for sex and AD. However, the distribution of p values for age and MCI were more conservative with less than 5% of p values in the 0-0.05 bin. Similar results were obtained for matched ADNI and MACC participants (Figure 9). For coVAE (Figure 9A) the frequency of p values within the 0-0.05 bin greatly exceeded 5% for all covariates. For DeepResBat (Figure 9B), the distributions of p values were uniform for all covariates. Visualization of p values distributions for no harmonization, ComBat and CovBat can be found in Figures S6 to S11.

Similar conclusions were obtained for the GLM involving MMSE (Figures S12 to S14). Furthermore, instead of permuting all four covariates, we also considered permuting only clinical diagnosis (Figure S15) or MMSE (Figure S16), yielding similar conclusions.

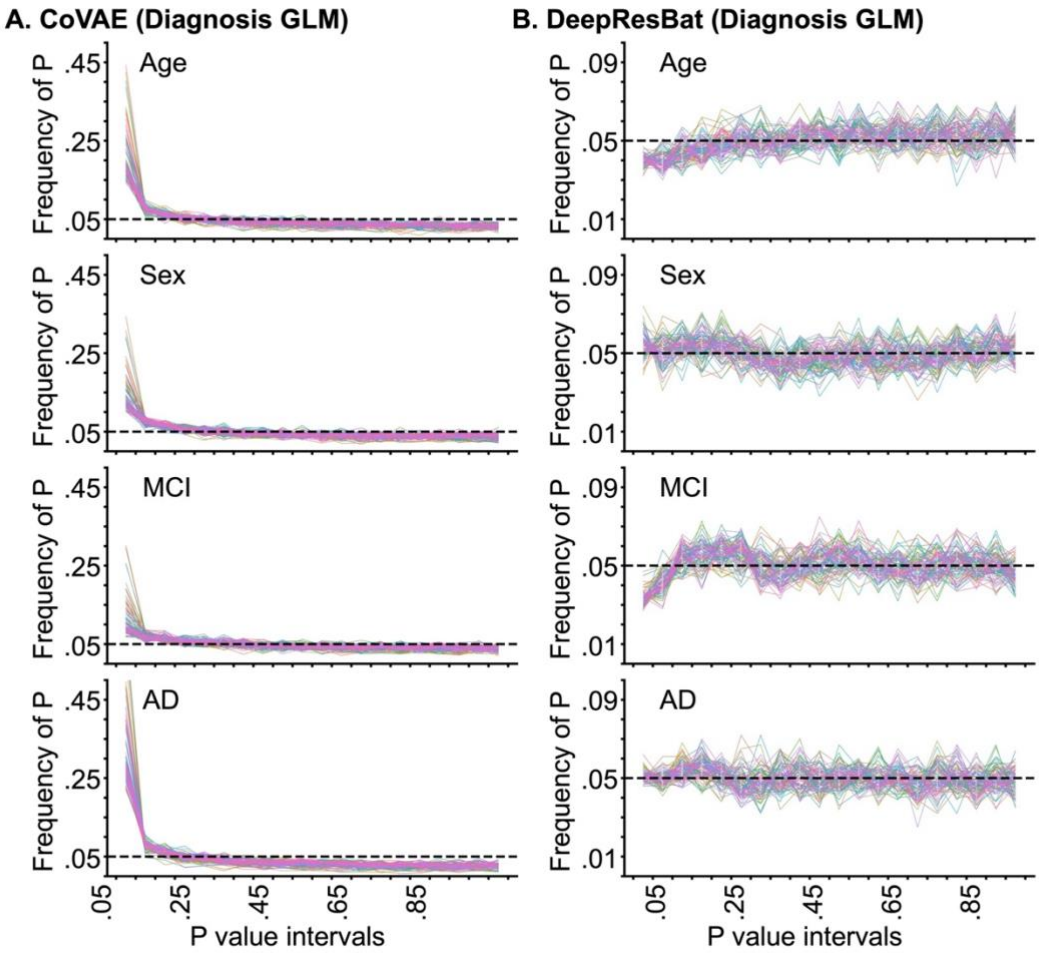

Figure 8. Frequency of p values of coVAE and DeepResBat for matched ADNI and AIBL participants by GLM involving clinical diagnosis based on 1000 permutations. Each line corresponds to a single brain ROI. P values were binned in intervals of 0.05. Therefore, in the ideal scenario, the distributions of p values should follow a uniform distribution with a height of 0.05. (A) Frequency of p values for coVAE. (B) Frequency of p values for DeepResBat.

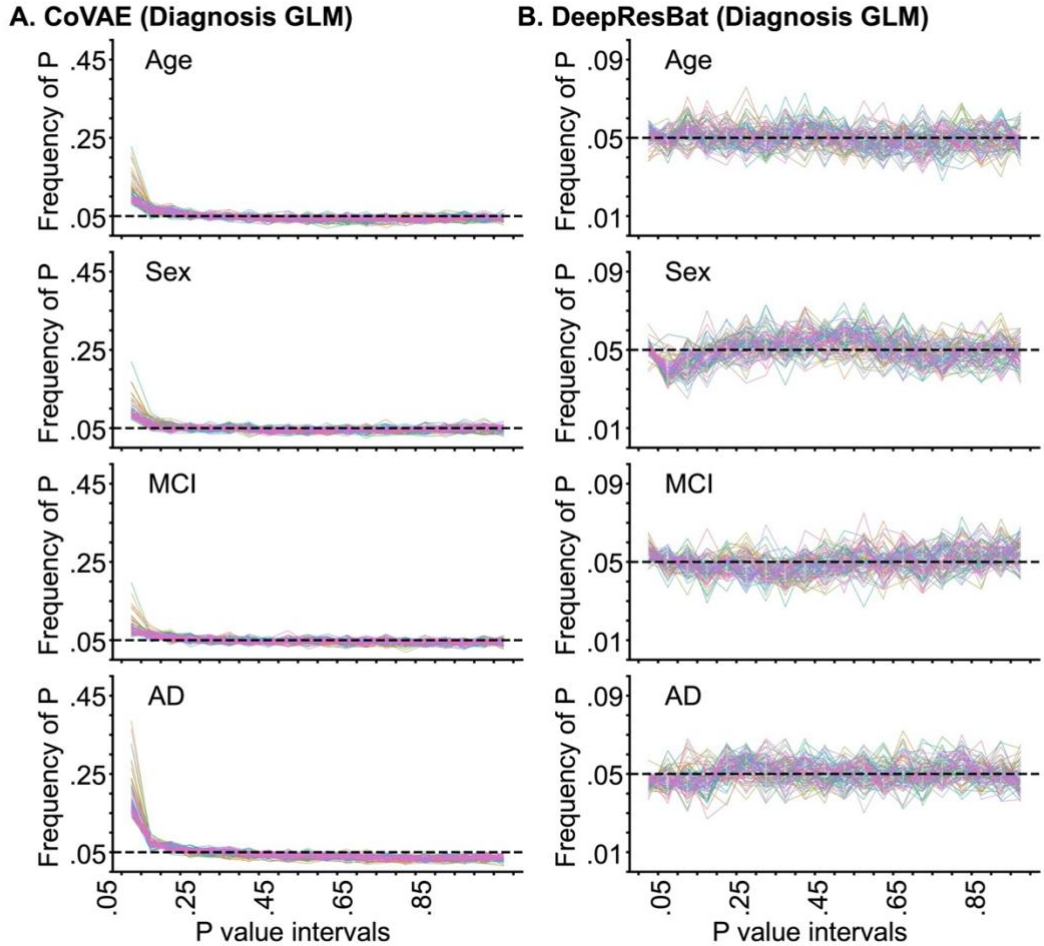

Figure 9. Frequency of p values of coVAE and DeepResBat for matched ADNI and MACC participants by GLM involving clinical diagnosis based on 1000 permutations. Each line corresponds to a single brain ROI. P values were binned in intervals of 0.05. Therefore, in the ideal scenario, the distributions of p values should follow a uniform distribution with a height of 0.05. (A) Frequency of p values for coVAE. (B) Frequency of p values for DeepResBat.

4 Discussion

Current deep learning approaches for harmonization do not explicitly account for covariate distribution differences across datasets. As discussed in Section 2.1, ignoring covariates can lead to theoretically worse harmonization outcomes. We then proposed two DNN-based harmonization approaches, coVAE and DeepResBat, which explicitly accounted for covariate distribution differences across datasets. We demonstrated that DeepResBat outperformed mixed effects and deep learning baselines across three evaluation experiments involving three large-scale MRI datasets.

More specifically, without any harmonization, XGBoost was able to predict almost perfectly which dataset a participant's MRI volumes came from (Figure 3). After harmonization with mixed effects models (ComBat and CovBat), dataset classification accuracies dropped significantly, suggesting that ComBat and CovBat were able to remove some dataset differences. DNN-based harmonization approaches further reduced the classification accuracies, suggesting even greater removal of dataset differences.

However, the removed dataset differences might contain important biological information, which should not be removed. Therefore, in the second experiment, we evaluated the strength of associations between the harmonized brain volumes and covariates (age, sex, MMSE and clinical diagnosis). Across both univariate GLM and multivariate MANOVA (Figures 4 to 6), we found that coVAE and DeepResBat yielded stronger associations between brain volumes and covariates. This suggests that coVAE and DeepResBat were retaining important biological information while removing undesirable dataset differences. Interestingly, DeepResBat was also more sensitive than coVAE, except for the association between brain volumes and sex in the matched ADNI and AIBL participants. Finally, cVAE exhibited weaker associations than ComBat, CovBat and even no harmonization (Figure 6), suggesting that cVAE was removing significant biological information in addition to unwanted dataset differences, thus providing empirical support for the theoretical discussion in Section 2.1.

Given the flexible nature of DNNs, we were concerned that explicitly accounting for covariates could lead to spurious associations between harmonized brain volumes and covariates when no association existed. Our permutation test (Figures 7 to 9) supported our concerns in the case of coVAE. Although coVAE provided a natural (and in our opinion, elegant) extension of cVAE, we found significant false positive rates for coVAE. On the

other hand, DeepResBat was able to exhibit an expected amount of false positives, consistent with less flexible mixed effects models (ComBat and CovBat).

Together, the three evaluation experiments suggest that DeepResBat is an effective deep learning alternative to ComBat. DeepResBat consisted of three steps: (1) regressed out the effects of covariates from the brain volumes, (2) followed by harmonizing the residuals, (3) and then adding the effects of covariates back to the harmonized residuals. Future research could investigate whether these three steps can be combined into a single optimization procedure by minimizing fitting Eq. (10) directly. However, we note that fitting Eq. (10) directly might lead to overfitting, yielding false positive issues, similar to coVAE.

While our current implementation of DeepResBat utilized XGBoost to estimate covariate effects, other nonlinear regression approaches can be used. Furthermore, instead of using cVAE in the harmonization step, cVAE can be replaced with other harmonization approaches, such as generative adversarial networks (Bashyam et al., 2021). One advantage of cVAE is that the approach readily works for more than two datasets by extending the one-hot encoding of sites. Therefore, although our experiments only harmonized pairs of datasets, DeepResBat can be readily applied to jointly harmonize three or more datasets.

A drawback of DeepResBat is that our current implementation operates on summary measures (e.g., volumes or thickness), rather than at the image level (Zuo et al., 2021; Cackowski et al., 2023). Therefore, the harmonization procedure needs to be repeated for different summary measures (e.g., using a different brain parcellation). However, this disadvantage also means that DeepResBat can be applied to harmonize not just imaging data, but also any tabular data (e.g., micro-array data), suggesting the broad applicability of DeepResBat to any field where instrumental harmonization is necessary.

5 Conclusion

In this study, we demonstrate the importance of incorporating covariates during harmonization. We propose two deep learning models, coVAE and DeepResBat, that account for covariate distribution differences across datasets. coVAE extends cVAE by concatenating covariates and site information with latent representations, while DeepResBat adopts a residual framework inspired by the classical ComBat framework. We found that coVAE introduces spurious associations between anatomical MRI and unrelated covariates, while DeepResBat effectively mitigates this false positive issue. Furthermore, DeepResBat outperformed ComBat, CovBat and cVAE in terms of removing dataset differences, while retaining biological effects of interest.

Acknowledgment

Our research is currently supported by the Singapore National Research Foundation (NRF) Fellowship (Class of 2017), the NUS Yong Loo Lin School of Medicine (NUHSRO/2020/124/TMR/LOA), the Singapore National Medical Research Council (NMRC) LCG (OFLCG19May-0035), NMRC STaR (STaR20nov-0003), and the USA NIH (R01MH120080). Our computational work was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Singapore NRF or the Singapore NMRC. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.;Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.;Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

CRediT authorship contribution statement

Lijun An: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Visualization, Writing – original draft, Writing – review & editing. **Chen Zhang**: Software, Validation, Visualization, Writing – review & editing. **Naren Wulan**: Investigation, Software, Validation, Visualization, Writing – review & editing. **Shaoshi Zhang**: Investigation, Software, Validation, Visualization, Writing – review & editing. **Pansheng Chen**: Software, Validation, Visualization, Writing – review & editing. **Fang Ji**: Resource, Writing – review & editing. **Kwun Kei Ng**: Investigation, Writing – review & editing. **Christopher Chen**: Resource, Writing – review & editing. **Juan Helen Zhou**: Resource, Writing – review & editing. **B.T. Thomas Yeo**: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Resource, Supervision, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

An, L., Chen, J., Chen, P., Zhang, C., He, T., Chen, C., Zhou, J. H., & Yeo, B. T. T. (2022). Goal-specific brain MRI harmonization. *NeuroImage*, 119570. https://doi.org/10.1016/j.neuroimage.2022.119570

Bashyam, V. M., Doshi, J., Erus, G., Srinivasan, D., Abdulkadir, A., Singh, A., Habes, M., Fan, Y., Masters, C. L., Maruff, P., Zhuo, C., Völzke, H., Johnson, S. C., Fripp, J., Koutsouleris, N., Satterthwaite, T. D., Wolf, D. H., Gur, R. E., Gur, R. C., … The iSTAGING and PHENOM consortia. (2021). Deep Generative Medical Image Harmonization for Improving Cross-Site Generalization in Deep Learning Predictors. *Journal of Magnetic Resonance Imaging*, jmri.27908. https://doi.org/10/gmzt7m

Beizaee, F., Desrosiers, C., Lodygensky, G. A., & Dolz, J. (2023). *Harmonizing Flows: Unsupervised MR harmonization based on normalizing flows* (arXiv:2301.11551). arXiv. http://arxiv.org/abs/2301.11551

Bethlehem, R. a. I., Seidlitz, J., White, S. R., Vogel, J. W., Anderson, K. M., Adamson, C., Adler, S., Alexopoulos, G. S., Anagnostou, E., Areces-Gonzalez, A., Astle, D. E., Auyeung, B., Ayub, M., Bae, J., Ball, G., Baron-Cohen, S., Beare, R., Bedford, S. A., Benegal, V., … Alexander-Bloch, A. F. (2022). Brain charts for the human lifespan. *Nature*, *604*(7906), Article 7906. https://doi.org/10.1038/s41586-022-04554-y

Blessed, G., Tomlinson, B. E., & Roth, M. (2018). The Association Between Quantitative Measures of Dementia and of Senile Change in the Cerebral Grey Matter of Elderly Subjects. *The British Journal of Psychiatry*, *114*(512), 797–811. https://doi.org/10.1192/bjp.114.512.797

Cackowski, S., Barbier, E. L., Dojat, M., & Christen, T. (2023). ImUnity: A generalizable VAE-GAN solution for multicenter MR image harmonization. *Medical Image Analysis*, *88*, 102799. https://doi.org/10.1016/j.media.2023.102799

Chen, A. A., Beer, J. C., Tustison, N. J., Cook, P. A., Shinohara, R. T., Shou, H., & Initiative, T. A. D. N. (2021). Mitigating site effects in covariance for machine learning in neuroimaging data. *Human Brain Mapping*, *43*(4). https://doi.org/10/gntvh2

Chen, J., Liu, J., Calhoun, V. D., Arias-Vasquez, A., Zwiers, M. P., Gupta, C. N., Franke, B., & Turner, J. A. (2014). Exploration of scanning effects in multi-site structural MRI studies. *Journal of Neuroscience Methods*, *230*, 37–50. https://doi.org/10.1016/j.jneumeth.2014.04.023

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chong, J. S. X., Liu, S., Loke, Y. M., Hilal, S., Ikram, M. K., Xu, X., Tan, B. Y., Venketasubramanian, N., Chen, C. L.-H., & Zhou, J. (2017). Influence of cerebrovascular disease on brain networks in prodromal and clinical Alzheimer's disease. *Brain*, *140*(11), 3012–3022. https://doi.org/10/gcj7wj

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

Dewey, B. E., Zhao, C., Reinhold, J. C., Carass, A., Fitzgerald, K. C., Sotirchos, E. S., Saidha, S., Oh, J., Pham, D. L., Calabresi, P. A., van Zijl, P. C. M., & Prince, J. L. (2019). DeepHarmony: A deep learning approach to contrast harmonization across scanner changes. *Magnetic Resonance Imaging*, *64*, 160–170. https://doi.org/10/ggbzsg

Dinsdale, N. K., Jenkinson, M., & Namburete, A. I. (2022). *FedHarmony: Unlearning Scanner Bias with Distributed Data* (arXiv:2205.15970). arXiv. http://arxiv.org/abs/2205.15970

Ellis, K. A., Bush, A. I., Darby, D., De Fazio, D., Foster, J., Hudson, P., Lautenschlager, N. T., Lenzo, N., Martins, R. N., Maruff, P., Masters, C., Milner, A., Pike, K., Rowe, C., Savage, G., Szoeke, C., Taddei, K., Villemagne, V., Woodward, M., … the AIBL Research Group. (2009). The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: Methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *International Psychogeriatrics*, *21*(4), 672–687. https://doi.org/10/dg74qm

Ellis, K. A., Rowe, C. C., Villemagne, V. L., Martins, R. N., Masters, C. L., Salvado, O., Szoeke, C., Ames, D., & Group, A. research. (2010). Addressing population aging and Alzheimer's disease through the Australian Imaging Biomarkers and Lifestyle study: Collaboration with the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's & Dementia*, *6*(3), 291–296. https://doi.org/10/btmfdj

Eriksson, D., Pearce, M., Gardner, J. R., Turner, R., & Poloczek, M. (2020). Scalable Global Optimization via Local Bayesian Optimization. *arXiv:1910.01739 [Cs, Stat]*. http://arxiv.org/abs/1910.01739

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., & Dale, A. M. (2002). Whole Brain Segmentation. *Neuron*, *33*(3), 341–355. https://doi.org/10.1016/s0896-6273(02)00569-x

Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., Adams, P., Cooper, C., Fava, M., McGrath, P. J., McInnis, M., Phillips, M. L., Trivedi, M. H., Weissman, M. M., & Shinohara, R. T. (2018). Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, *167*, 104–120. https://doi.org/10.1016/j.neuroimage.2017.11.024

Fortin, J.-P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E., Schultz, R. T., Verma, R., & Shinohara, R. T. (2017). Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*, *161*, 149–170. https://doi.org/10/gcmg6f

Garcia-Dias, R., Scarpazza, C., Baecker, L., Vieira, S., Pinaya, W. H. L., Corvin, A., Redolfi, A., Nelson, B., Crespo-Facorro, B., McDonald, C., Tordesillas-Gutiérrez, D., Cannon, D., Mothersill, D., Hernaus, D., Morris, D., Setien-Suero, E., Donohoe, G., Frisoni, G., Tronchin, G., … Mechelli, A. (2020). Neuroharmony: A new tool for harmonizing volumetric MRI data from unseen scanners. *NeuroImage*, *220*, 117127. https://doi.org/10.1016/j.neuroimage.2020.117127

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). *Why do tree-based models still outperform deep learning on tabular data?* (arXiv:2207.08815). arXiv. http://arxiv.org/abs/2207.08815

Hawco, C., Viviano, J. D., Chavez, S., Dickie, E. W., Calarco, N., Kochunov, P., Argyelan, M., Turner, J. A., Malhotra, A. K., Buchanan, R. W., & Voineskos, A. N. (2018). A longitudinal human phantom reliability study of multi-center T1-weighted, DTI, and resting state fMRI data. *Psychiatry Research: Neuroimaging*, *282*, 134–142. https://doi.org/10.1016/j.pscychresns.2018.06.004

He, T., An, L., Chen, P., Chen, J., Feng, J., Bzdok, D., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. T. (2022). Meta-matching as a simple framework to translate phenotypic predictive models from big to small data. *Nature Neuroscience*, *25*(6), Article 6. https://doi.org/10.1038/s41593-022-01059-9

Hilal, S., Chai, Y. L., Ikram, M. K., Elangovan, S., Yeow, T. B., Xin, X., Chong, J. Y., Venketasubramanian, N., Richards, A. M., Chong, J. P. C., Lai, M. K. P., & Chen, C. (2015). Markers of Cardiac Dysfunction in Cognitive Impairment and Dementia. *Medicine*, *94*(1), e297. https://doi.org/10/gpfkhg

Hilal, S., Tan, C. S., van Veluw, S. J., Xu, X., Vrooman, H., Tan, B. Y., Venketasubramanian, N., Biessels, G. J., & Chen, C. (2020). Cortical cerebral microinfarcts predict cognitive decline in memory clinic patients. *Journal of Cerebral Blood Flow & Metabolism*, *40*(1), 44–53. https://doi.org/10/gpfttm

Hu, F., Chen, A. A., Horng, H., Bashyam, V., Davatzikos, C., Alexander-Bloch, A., Li, M., Shou, H., Satterthwaite, T. D., Yu, M., & Shinohara, R. T. (2023). Image harmonization: A review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization. *NeuroImage*, *274*, 120125. https://doi.org/10.1016/j.neuroimage.2023.120125

Hua, X., Hibar, D. P., Lee, S., Toga, A. W., Jack, C. R., Weiner, M. W., & Thompson, P. M. (2010). Sex and age differences in atrophic rates: An ADNI study with n=1368 MRI scans. *Neurobiology of Aging*, *31*(8), 1463–1480. https://doi.org/10.1016/j.neurobiolaging.2010.04.033

Hutton, C., Draganski, B., Ashburner, J., & Weiskopf, N. (2009). A comparison between voxel-based cortical thickness and voxel-based morphometry in normal aging. *NeuroImage*, *48*(2), 371–380. https://doi.org/10.1016/j.neuroimage.2009.06.043

Ilievski, I., Akhtar, T., Feng, J., & Shoemaker, C. A. (2017). Efficient Hyperparameter Optimization of Deep Learning Algorithms Using Deterministic RBF Surrogates. *arXiv:1607.08316 [Cs, Stat]*. http://arxiv.org/abs/1607.08316

Iturbide, M., Bedia, J., Herrera, S., Baño-Medina, J., Fernández, J., Frías, M. D., Manzanas, R., San-Martín, D., Cimadevilla, E., Cofiño, A. S., & Gutiérrez, J. M. (2019). The R-based climate4R open framework for reproducible climate data access and post-processing. *Environmental Modelling & Software*, *111*, 42–54. https://doi.org/10.1016/j.envsoft.2018.09.009

Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L. Whitwell, J., Ward, C., Dale, A. M., Felmlee, J. P., Gunter, J. L., Hill, D. L. G., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., … ADNI Study. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, *27*(4), 685–691. https://doi.org/10/d6mhb9

Jack Jr., C. R., Bernstein, M. A., Borowski, B. J., Gunter, J. L., Fox, N. C., Thompson, P. M., Schuff, N., Krueger, G., Killiany, R. J., DeCarli, C. S., Dale, A. M., Carmichael, O. W., Tosun, D., Weiner, M. W., & Initiative, A. D. N. (2010). Update on the Magnetic Resonance Imaging core of the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's & Dementia*, *6*(3), 212–220. https://doi.org/10.1016/j.jalz.2010.03.004

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118–127. https://doi.org/10.1093/biostatistics/kxj037

Kingma, D. P., & Ba, J. (2017). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [Cs]*. http://arxiv.org/abs/1412.6980

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, *2*(1–2), 83–97. https://doi.org/10/b2k5tg

Liu, M., Zhu, A. H., Maiti, P., Thomopoulos, S. I., Gadewar, S., Chai, Y., Kim, H., Jahanshad, N., & Initiative, for the A. D. N. (2023). Style transfer generative adversarial networks to harmonize multisite MRI to a single reference image to avoid

overcorrection. *Human Brain Mapping*, *44*(14), 4875–4892. https://doi.org/10.1002/hbm.26422

Lu, B., Li, H.-X., Chang, Z.-K., Li, L., Chen, N.-X., Zhu, Z.-C., Zhou, H.-X., Li, X.-Y., Wang, Y.-W., Cui, S.-X., Deng, Z.-Y., Fan, Z., Yang, H., Chen, X., Thompson, P. M., Castellanos, F. X., & Yan, C.-G. (2022). A practical Alzheimer's disease classifier via brain imaging-based deep learning on 85,721 samples. *Journal of Big Data*, *9*(1), 101. https://doi.org/10.1186/s40537-022-00650-y

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. *In ICML Workshop on Deep Learning for Audio, Speech and Language Processing*.

Magnotta, V. A., Matsui, J. T., Liu, D., Johnson, H. J., Long, J. D., Bolster, B. D., Mueller, B. A., Lim, K., Mori, S., Helmer, K. G., Turner, J. A., Reading, S., Lowe, M. J., Aylward, E., Flashman, L. A., Bonett, G., & Paulsen, J. S. (2012). MultiCenter Reliability of Diffusion Tensor Imaging. *Brain Connectivity*, *2*(6), 345–355. https://doi.org/10/gk82p6

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., … Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, *603*(7902), Article 7902. https://doi.org/10.1038/s41586-022-04492-9

Modanwal, G., Vellal, A., Buda, M., & Mazurowski, M. A. (2020). MRI image harmonization using cycle-consistent generative adversarial network. In H. K. Hahn & M. A. Mazurowski (Eds.), *Medical Imaging 2020: Computer-Aided Diagnosis* (p. 36). SPIE. https://doi.org/10/gmzt6h

Moyer, D., Ver Steeg, G., Tax, C. M. W., & Thompson, P. M. (2020). Scanner invariant representations for diffusion MRI harmonization. *Magnetic Resonance in Medicine*, *84*(4), 2174–2189. https://doi.org/10.1002/mrm.28243

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in PyTorch. *Undefined*. https://www.semanticscholar.org/paper/Automatic-differentiation-in-PyTorch-Paszke-Gross/b36a5bb1707bb9c70025294b3a310138aae8327a

Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., Bashyam, V., Nasrallah, I. M., Satterthwaite, T. D., Fan, Y., Launer, L. J., Masters, C. L., Maruff, P., Zhuo, C., Völzke, H., Johnson, S. C., Fripp, J., Koutsouleris, N., Wolf, D. H., … Davatzikos, C. (2020). Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, *208*, 116450. https://doi.org/10.1016/j.neuroimage.2019.116450

Regis, R. G., & Shoemaker, C. A. (2013). Combining radial basis function surrogates and dynamic coordinate search in high-dimensional expensive black-box optimization. *Engineering Optimization*, *45*(5), 529–555. https://doi.org/10.1080/0305215x.2012.687731

Russkikh, N., Antonets, D., Shtokalo, D., Makarov, A., Vyatkin, Y., Zakharov, A., & Terentyev, E. (2020). Style transfer with variational autoencoders is a promising approach to RNA-Seq data harmonization and analysis. *Bioinformatics*, *36*(20), 5076–5085. https://doi.org/10.1093/bioinformatics/btaa624

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, *81*, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011

Sohn, K., Lee, H., & Yan, X. (2015). Learning Structured Output Representation using Deep Conditional Generative Models. *Advances in Neural Information Processing Systems*,

28. https://proceedings.neurips.cc/paper/2015/hash/8d55a249e6baa5c06772297520da2051-Abstract.html

Tachet, R., Zhao, H., Wang, Y.-X., & Gordon, G. (2020). Domain Adaptation with Conditional Distribution Matching and Generalized Label Shift. *arXiv:2003.04475 [Cs, Stat]*. http://arxiv.org/abs/2003.04475

Thompson, P. M., Andreassen, O. A., Arias-Vasquez, A., Bearden, C. E., Boedhoe, P. S., Brouwer, R. M., Buckner, R. L., Buitelaar, J. K., Bulayeva, K. B., Cannon, D. M., Cohen, R. A., Conrod, P. J., Dale, A. M., Deary, I. J., Dennis, E. L., de Reus, M. A., Desrivieres, S., Dima, D., Donohoe, G., … Ye, J. (2017). ENIGMA and the individual: Predicting factors that affect the brain in 35 countries worldwide. *NeuroImage*, *145*, 389–408. https://doi.org/10.1016/j.neuroimage.2015.11.057

Tian, Y. E., Cropley, V., Maier, A. B., Lautenschlager, N. T., Breakspear, M., & Zalesky, A. (2023). Heterogeneous aging across multiple organ systems and prediction of chronic disease and mortality. *Nature Medicine*, *29*(5), Article 5. https://doi.org/10.1038/s41591-023-02296-6

van de Mortel, L. A., Thomas, R. M., van Wingen, G. A., & Initiative, for the A. D. N. (2021). Grey Matter Loss at Different Stages of Cognitive Decline: A Role for the Thalamus in Developing Alzheimer's Disease. *Journal of Alzheimer's Disease*, *83*(2), 705–720. https://doi.org/10.3233/JAD-210173

Vogel, J. W., Young, A. L., Oxtoby, N. P., Smith, R., Ossenkoppele, R., Strandberg, O. T., La Joie, R., Aksman, L. M., Grothe, M. J., Iturria-Medina, Y., Pontecorvo, M. J., Devous, M. D., Rabinovici, G. D., Alexander, D. C., Lyoo, C. H., Evans, A. C., & Hansson, O. (2021). Four distinct trajectories of tau deposition identified in Alzheimer's disease. *Nature Medicine*, *27*(5), Article 5. https://doi.org/10.1038/s41591-021-01309-6

Volkow, N. D., Koob, G. F., Croyle, R. T., Bianchi, D. W., Gordon, J. A., Koroshetz, W. J., Pérez-Stable, E. J., Riley, W. T., Bloch, M. H., Conway, K., Deeds, B. G., Dowling, G. J., Grant, S., Howlett, K. D., Matochik, J. A., Morgan, G. D., Murray, M. M., Noronha, A., Spong, C. Y., … Weiss, S. R. B. (2018). The conception of the ABCD study: From substance use to a broad NIH collaboration. *Developmental Cognitive Neuroscience*, *32*, 4–7. https://doi.org/10.1016/j.dcn.2017.10.002

Wachinger, C., Rieckmann, A., & Pölsterl, S. (2021). Detect and correct bias in multi-site neuroimaging datasets. *Medical Image Analysis*, *67*, 101879. https://doi.org/10/gh5vwj

Wang, R., Chaudhari, P., & Davatzikos, C. (2021). Harmonization with Flow-Based Causal Inference. In M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, & C. Essert (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* (Vol. 12903, pp. 181–190). Springer International Publishing. https://doi.org/10.1007/978-3-030-87199-4_17

Yu, M., Linn, K. A., Cook, P. A., Phillips, M. L., McInnis, M., Fava, M., Trivedi, M. H., Weissman, M. M., Shinohara, R. T., & Sheline, Y. I. (2018). Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. *Human Brain Mapping*, *39*(11), 4213–4227. https://doi.org/10/gff7m4

Zhao, F., Wu, Z., Wang, L., Lin, W., Xia, S., the UNC/UMN Baby Connectome Project Consortium, Zhao, F., Wu, Z., Wang, L., Lin, W., Xia, S., Shen, D., & Li, G. (2019). Harmonization of Infant Cortical Thickness Using Surface-to-Surface Cycle-Consistent Adversarial Networks. In S. Zhou, P.-T. Yap, & A. Khan (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* (Vol. 11767,

pp. 475–483). Springer International Publishing. https://doi.org/10.1007/978-3-030-32251-9_52

Zuo, L., Dewey, B. E., Liu, Y., He, Y., Newsome, S. D., Mowry, E. M., Resnick, S. M., Prince, J. L., & Carass, A. (2021). Unsupervised MR harmonization by learning disentangled representations using information bottleneck theory. *NeuroImage*, *243*, 118569. https://doi.org/10/gmzt6k

Supplementary Material

Vendor	Scanner Model	Field Strength	Number of scans
GE	Discovery	3T	595
	Genesis Signa	3T	273
	Signa Excite	1.5T	838
		3T	30
	Signa HDx	1.5T	464
		3T	42
	Signa HDxt	1.5T	212
		3T	405
Philips	Achieva	1.5T	67
		3T	481
	Gemini	3T	32
	Gyroscan Intera	1.5T	12
	Gyroscan NT	1.5T	2
	Ingenia	3T	84
	Ingenuity	3T	18
	Intera	1.5T	319
		3T	216
	Intera Achieva	1.5T	6
		3T	1
Siemens	Allegra	3T	48
	Avanto	1.5T	385
	Biograph	3T	12
	Espree	1.5T	22
	NUMARIS/4	1.5T	2
	Prisma	3T	2
	Prisma_fit	3T	3
	Skyra	3T	274
	Sonata	1.5T	371
	SonataVision	1.5T	25
	Symphony	1.5T	547
	SymphonyTim	1.5T	88
	Trio	3T	107
	TrioTim	3T	1371
	Verio	3T	601

Table S1. Scanner information for 7955 scans in ADNI dataset.

Vendor	Scanner Model	Field Strength	Number of scans
Siemens	Avanto	1.5T	241
	TrioTim	3T	558
	Verio	3T	134

Table S2. Scanner information for 933 scans in AIBL dataset.

	Timepoint	ADNI value	AIBL value	P value
AGE	1	71.0±5.5	70.8±5.3	0.96
	2	72.5±5.5	72.6±5.5	0.98
	3	74.2±5.5	73.9±5.6	0.93
	4	75.7±5.5	75.6±5.5	0.99
MMSE	1	29.3±0.9	29.2±0.9	1.00
	2	29.5±0.5	29.5±0.5	1.00
	3	29.7±0.5	29.7±0.5	1.00
	4	29.5±0.8	29.5±0.8	1.00
AD diagnosis	1	100%-0%-0%	100%-0%-0%	1.00
	2	100%-0%-0%	100%-0%-0%	1.00
	3	100%-0%-0%	100%-0%-0%	1.00
	4	100%-0%-0%	100%-0%-0%	1.00
Sex	-	50%	50%	1.00

Table S3. ADNI-AIBL matching results for participants having 4 time points (scans). For clinical diagnosis in the table, the percentage is showed as CN%-MCI%-AD%. For sex in the table, the portion is the ratio of male subjects. For Age/MMSE, the p value was calculated from a two-sample t-test. For Sex/AD diagnosis, the p value was calculated from the chi-square goodness of fit test.

	Timepoint	ADNI value	AIBL value	P value
AGE	1	73.3±3.3	73.1±3.3	0.96
	2	74.8±3.3	75.2±3.3	0.94
	3	76.3±3.3	76.1±3.3	0.97
MMSE	1	29.0±0.0	20.0±0.0	1.00
	2	30.0±0.0	30.0±0.0	1.00
	3	30.0±0.0	30.0±0.0	1.00
AD diagnosis	1	100%-0%-0%	100%-0%-0%	1.00
	2	100%-0%-0%	100%-0%-0%	1.00
	3	100%-0%-0%	100%-0%-0%	1.00
Sex	-	50%	50%	1.00

Table S4. ADNI-AIBL matching results for participants having 3 time points (scans). For clinical diagnosis in the table, the percentage is showed as CN%-MCI%-AD%. For sex in the table, the portion is the ratio of male subjects. For Age/MMSE, the p value was calculated from a two-sample t-test. For Sex/AD diagnosis, the p value was calculated from the chi-square goodness of fit test.

	Timepoint	ADNI value	AIBL value	P value
AGE	1	74.4±9.8	74.5±9.8	0.99
	2	76.1±9.8	76.1±9.9	0.99
MMSE	1	27.9±2.8	27.9±2.8	1.00
	2	27.8±2.8	27.8±2.8	1.00
AD diagnosis	1	57%-43%-0%	57%-43%-0%	1.00
	2	57%-43%-0%	57%-43%-0%	1.00
Sex	-	88%	88%	1.00

Table S5. ADNI-AIBL matching results for participants having 2 time points (scans). For clinical diagnosis in the table, the percentage is showed as CN%-MCI%-AD%. For sex in the table, the portion is the ratio of male subjects. For Age/MMSE, the p value was calculated from a two-sample t-test. For Sex/AD diagnosis, the p value was calculated from the chi-square goodness of fit test.

	Timepoint	ADNI value	AIBL value	P value
AGE	1	74.8±5.9	74.8±5.9	1.00
MMSE	1	27.3±3.9	27.3±3.9	0.98
AD diagnosis	1	68%-19%-13%	68%-19%-13%	1.00
Sex	-	43%	43%	1.00

Table S6. ADNI-AIBL matching results for participants having 1 time point (scan). For clinical diagnosis in the table, the percentage is showed as CN%-MCI%-AD%. For sex in the table, the portion is the ratio of male subjects. For Age/MMSE, the p value was calculated from a two-sample t-test. For Sex/AD diagnosis, the p value was calculated from the chi-square goodness of fit test.

	Timepoint	ADNI value	MACC value	P value
AGE	1	71.5±6.8	72.3±6.7	0.67
	2	73.5±6.8	73.8±6.8	0.91
	3	75.9±6.9	75.5±6.6	0.81
MMSE	1	26.9±3.7	27.0±3.5	0.94
	2	26.1±4.5	26.1±4.5	0.98
	3	24.9±6.3	25.2±6.3	0.87
AD diagnosis	1	39%-46%-15%	36%-54%-10%	0.72
	2	43%-36%-21%	46%-36%-18%	0.88
	3	43%-36%-21%	46%-32%-22%	0.91
Sex	-	57%	57%	1.00

Table S7. ADNI-MACC matching results for participants having 3 time points (scans). For clinical diagnosis in the table, the percentage is showed as CN%-MCI%-AD%. For sex in the table, the portion is the ratio of male subjects. For Age/MMSE, the p value was calculated from a two-sample t-test. For Sex/AD diagnosis, the p value was calculated from the chi-square goodness of fit test.

	Timepoint	ADNI value	MACC value	P value
AGE	1	73.6±5.7	73.9±5.6	0.78
	2	75.8±5.6	75.5±5.6	0.71
MMSE	1	24.7±4.9	24.8±4.6	0.86
	2	23.4±6.9	23.5±6.6	0.91
AD diagnosis	1	35%-38%-27%	35%-40%-25%	0.80
	2	37%-30%-33%	37%-35%-28%	0.49
Sex	-	51%	58%	0.20

Table S8. ADNI-MACC matching results for participants having 2 time points (scans). For clinical diagnosis in the table, the percentage is showed as CN%-MCI%-AD%. For sex in the table, the portion is the ratio of male subjects. For Age/MMSE, the p value was calculated from a two-sample t-test. For Sex/AD diagnosis, the p value was calculated from the chi-square goodness of fit test

	Timepoint	ADNI value	MACC value	P value
AGE	1	75.7±6.7	75.7±6.7	0.97
MMSE	1	21.0±5.9	21.0±5.9	0.94
AD diagnosis	1	14%-34%-52%	14%-38%-48%	0.64
Sex	-	52%	56%	0.34

Table S9. ADNI-MACC matching results for participants having 1 time points (scans). For clinical diagnosis in the table, the percentage is showed as CN%-MCI%-AD%. For sex in the table, the portion is the ratio of male subjects. For Age/MMSE, the p value was calculated from a two-sample t-test. For Sex/AD diagnosis, the p value was calculated from the chi-square goodness of fit test.

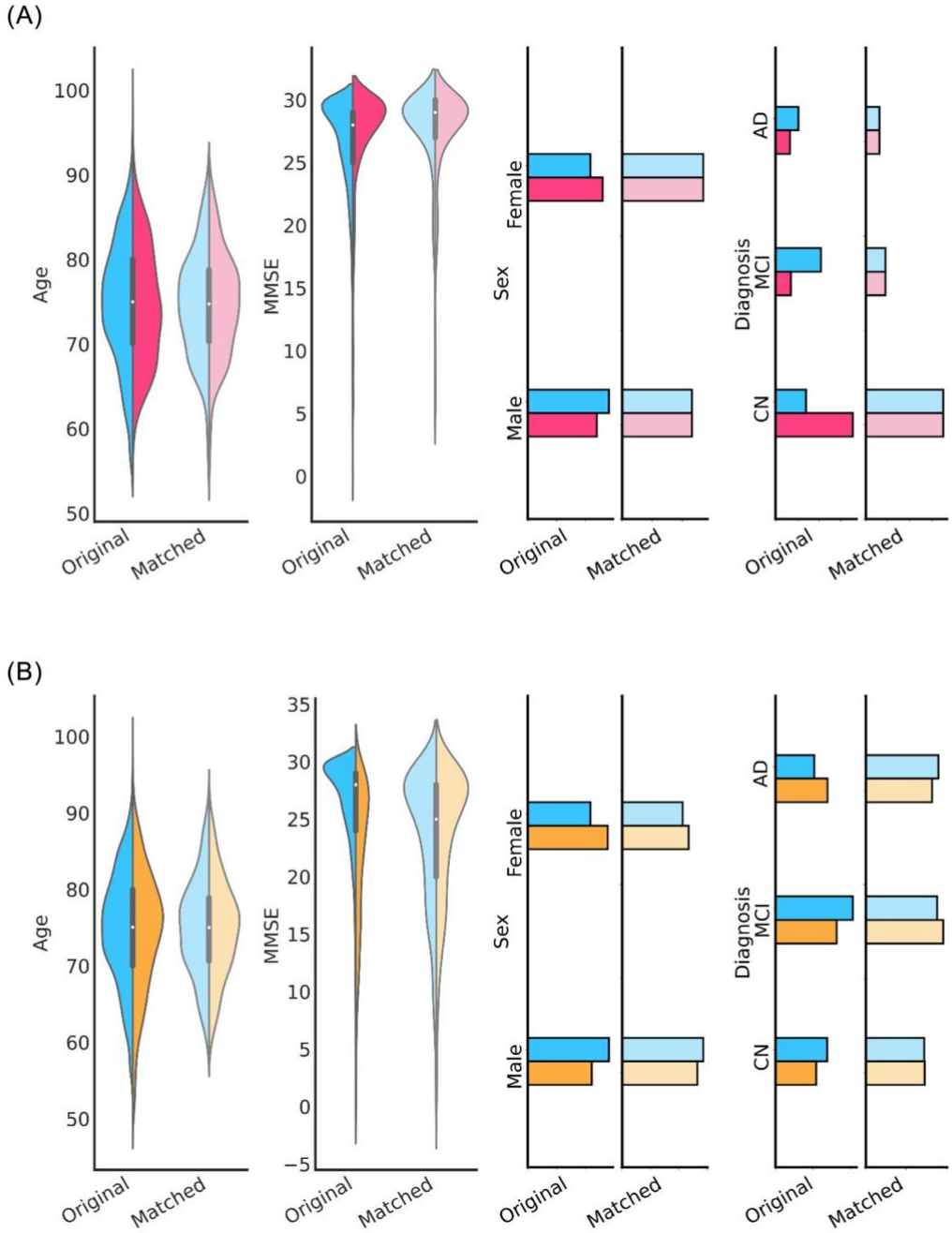

Figure S1. Age, MMSE, sex and clinical diagnosis distributions before and after matching. (A) Distributions of age, sex, MMSE and clinical diagnosis for ADNI (blue) and AIBL (red). Differences in the attributes between ADNI and AIBL were not significant after matching. (B) Distributions of age, sex, MMSE and clinical diagnosis for ADNI (blue) and MACC (yellow). Differences in the attributes between ADNI and MACC were not significant after matching. P values showing the quality of the matching procedure are found in Tables S3 to S9.

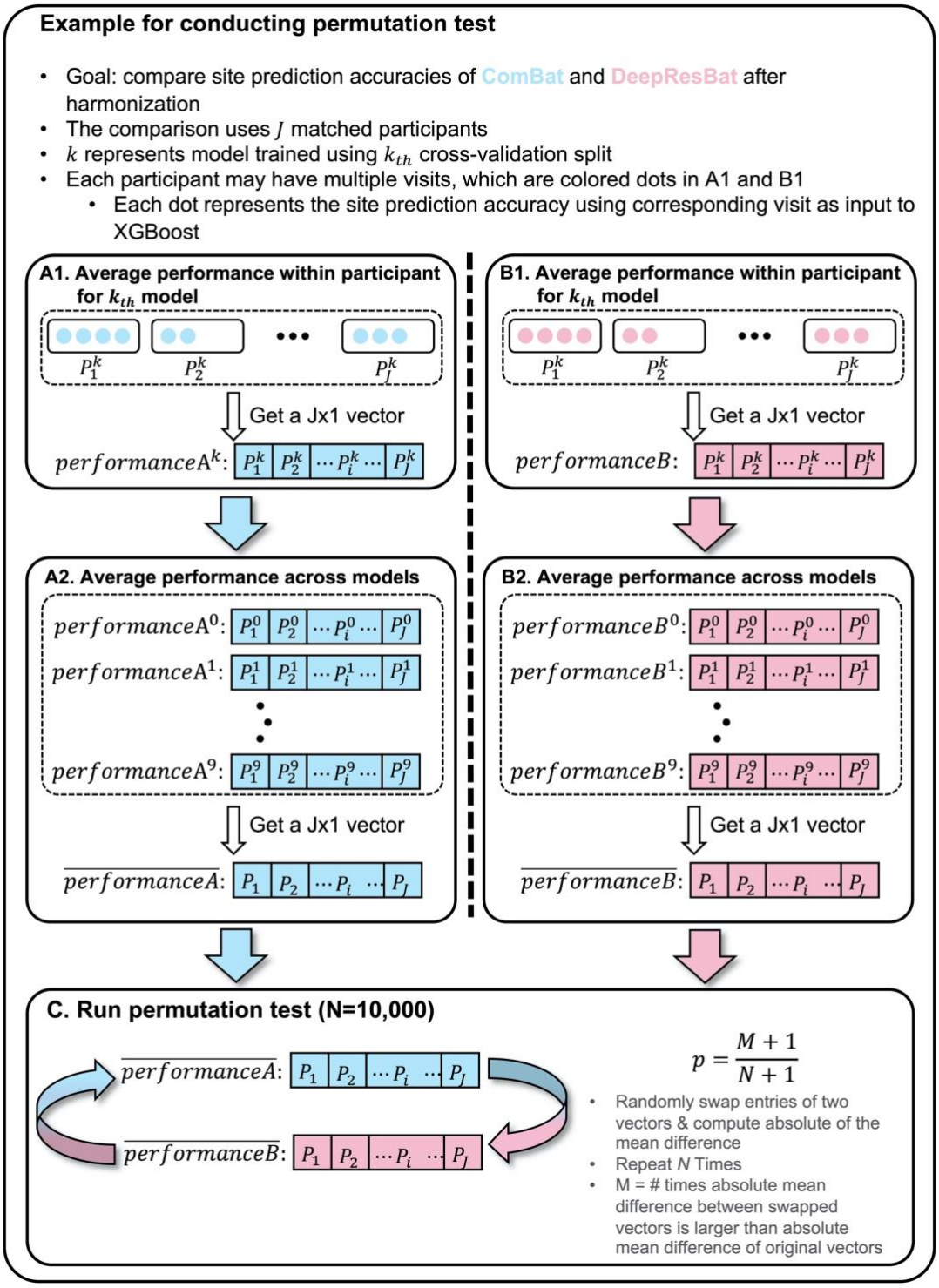

Figure S2. Illustration of permutation test for comparing site prediction accuracies of ComBat and DeepResBat. (A1) For a given model, we averaged the site prediction accuracies within each participant for ComBat. (B1) Same as A1 but for DeepResBat. (A2) Averaging the site prediction accuracies across the 10 models within each participant. (B1 & B2) Same as A1 and A2 but for DeepResBat. (C) Permute 10,000 times to obtain p value.

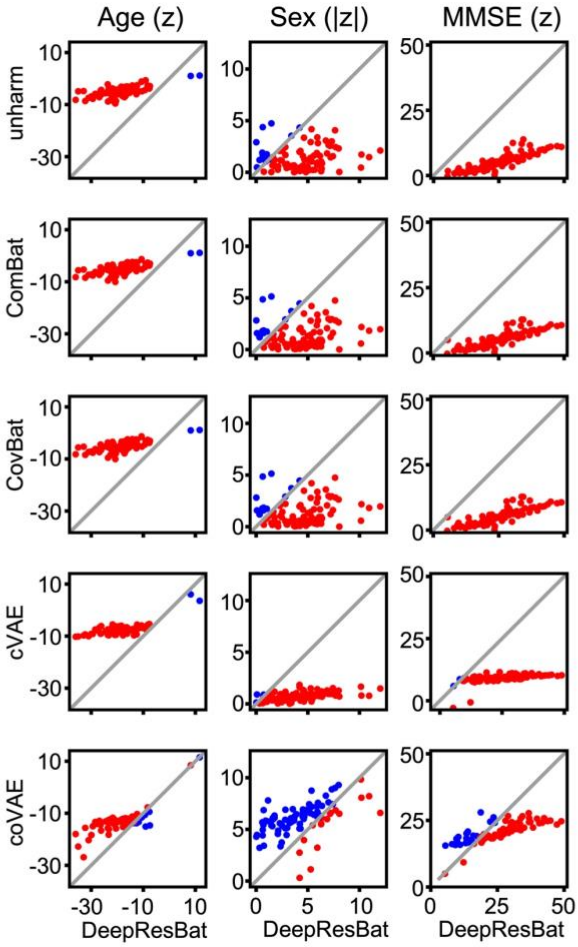

Figure S3. Comparison of z statistics from GLM involving MMSE for DeepResBat and baselines on matched ADNI and AIBL participants. Each row compares DeepResBat and one baseline approach: no harmonization (row 1), ComBat (row 2), CovBat (row 3), cVAE (row 4) and coVAE (row 5). Each column represents one covariate: age (column 1), sex (column 2), and MMSE (column 3). Each subplot compares z statistics of DeepResBat against another baseline for a given covariate across 87 grey matter ROIs. Each dot represents one grey matter ROI. Red dots indicate better performance by DeepResBat. Blue dots indicate worse performance by DeepResBat.

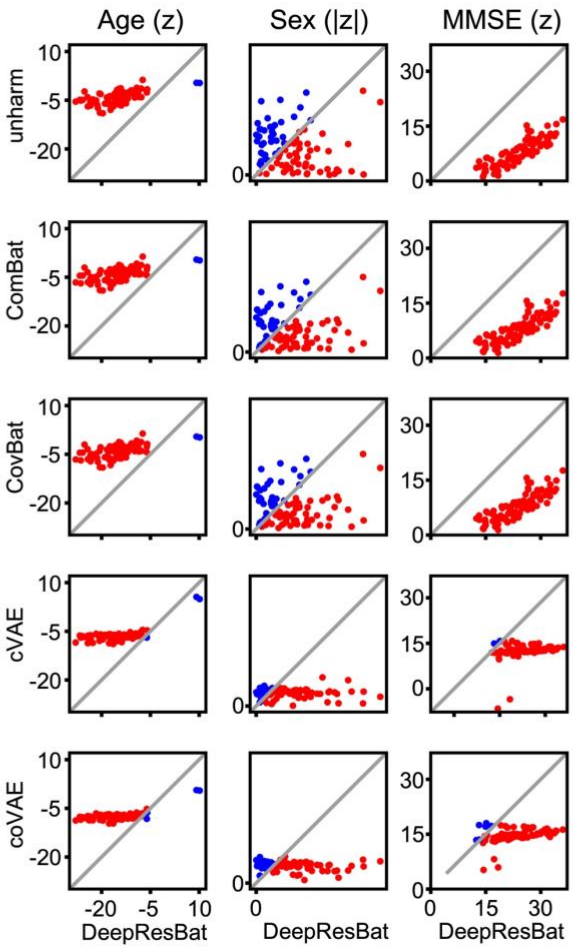

Figure S4. Comparison of z statistics from GLM involving MMSE for DeepResBat and baselines on matched ADNI and MACC participants. Each row compares DeepResBat and one baseline approach: no harmonization (row 1), ComBat (row 2), CovBat (row 3), cVAE (row 4) and coVAE (row 5). Each column represents one covariate: age (column 1), sex (column 2), and MMSE (column 3). Each subplot compares z statistics of DeepResBat against another baseline for a given covariate across 87 grey matter ROIs. Each dot represents one grey matter ROI. Red dots indicate better performance by DeepResBat. Blue dots indicate worse performance by DeepResBat.

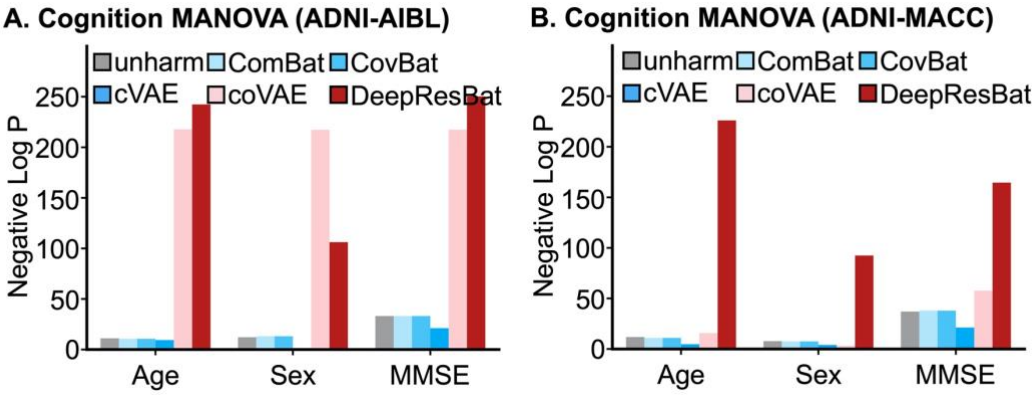

Figure S5. Significance bar plot by MANOVA involving MMSE. A larger negative of log p value indicates a stronger association, and thus better performance. (A) Bar plot for matched ADNI and AIBL participants. (B) Bar plot for matched ADNI and MACC participants.

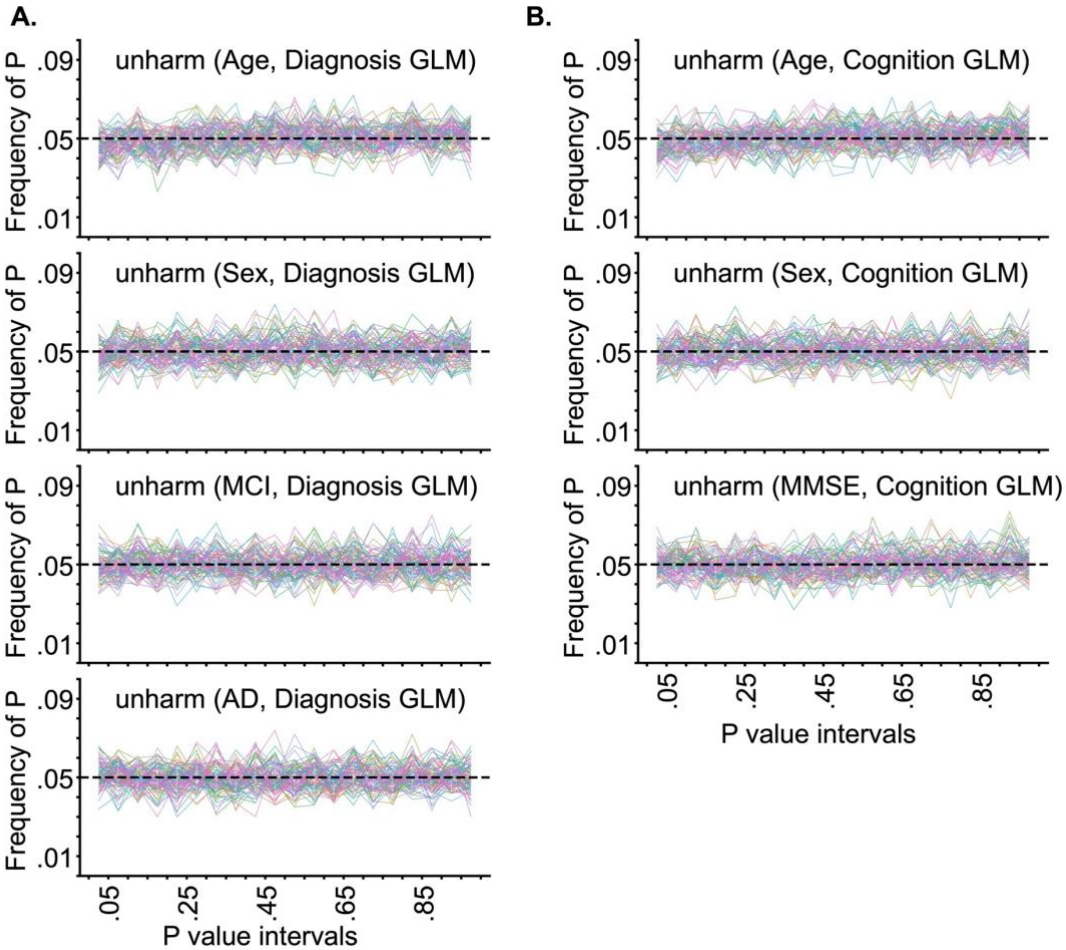

Figure S6. Frequency of p values of unharmonized data for matched ADNI and AIBL participants by GLMs involving clinical diagnosis and MMSE based on 1000 permutations. Each line corresponds to a single brain ROI. P values were binned in intervals of 0.05. Therefore, in the ideal scenario, the distributions of p values should follow a uniform distribution with a height of 0.05. (A) Frequency of p values by GLMs involving clinical diagnosis. (B) Frequency of p values by GLMs involving MMSE.

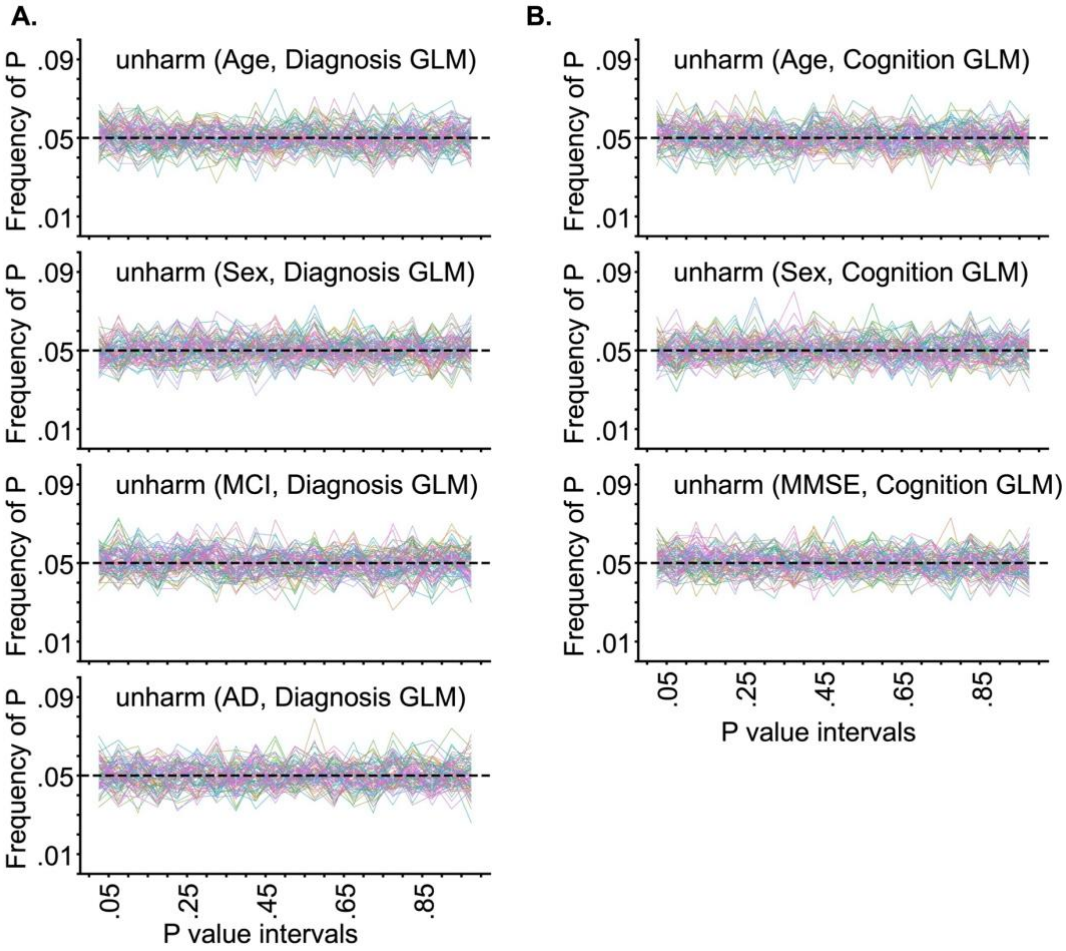

Figure S7. Frequency of p values of unharmonized data for matched ADNI and MACC participants by GLMs involving clinical diagnosis and MMSE based on 1000 permutations. Each line corresponds to a single brain ROI. P values were binned in intervals of 0.05. Therefore, in the ideal scenario, the distributions of p values should follow a uniform distribution with a height of 0.05. (A) Frequency of p values by GLMs involving clinical diagnosis. (B) Frequency of p values by GLMs involving MMSE.

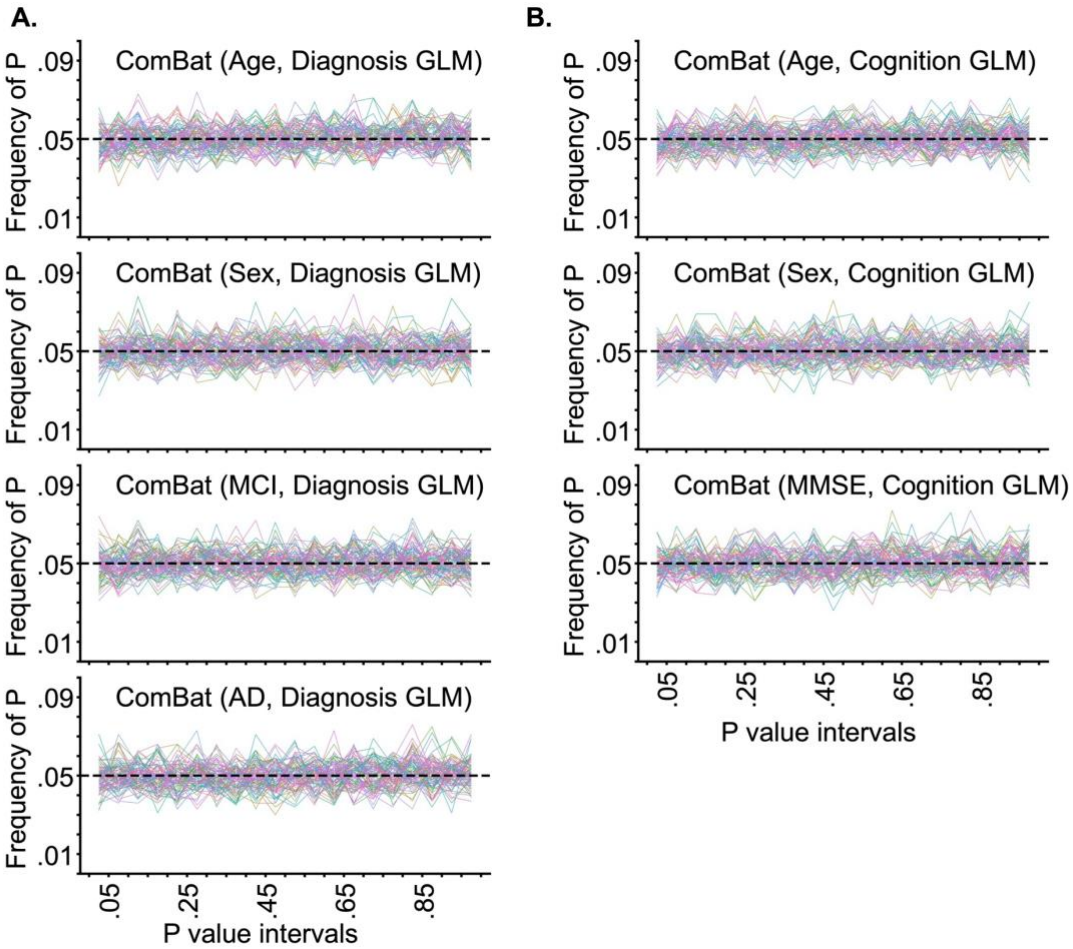

Figure S8. Frequency of p values of ComBat for matched ADNI and AIBL participants by GLMs involving clinical diagnosis and MMSE based on 1000 permutations. Each line corresponds to a single brain ROI. P values were binned in intervals of 0.05. Therefore, in the ideal scenario, the distributions of p values should follow a uniform distribution with a height of 0.05. (A) Frequency of p values by GLMs involving clinical diagnosis. (B) Frequency of p values by GLMs involving MMSE.

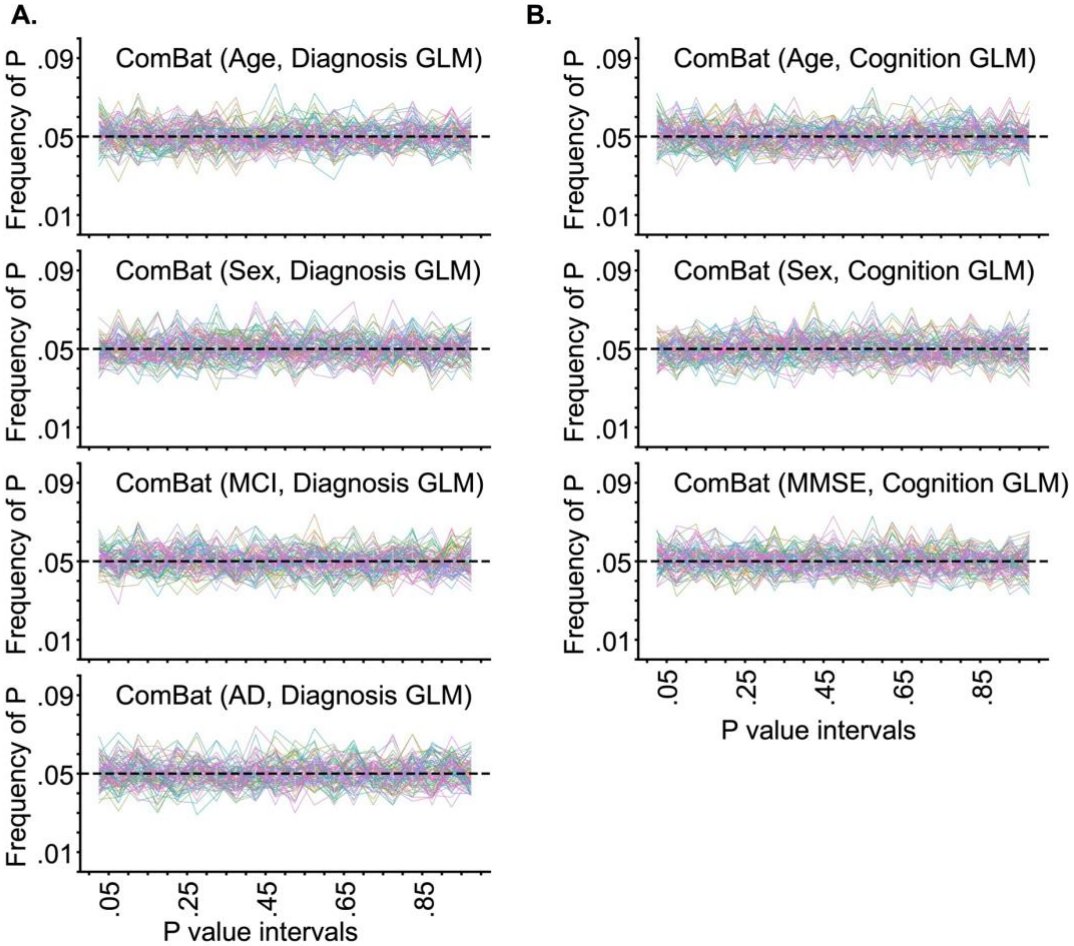

Figure S9. Frequency of p values of ComBat for matched ADNI and MACC participants by GLMs involving clinical diagnosis and MMSE based on 1000 permutations. Each line corresponds to a single brain ROI. P values were binned in intervals of 0.05. Therefore, in the ideal scenario, the distributions of p values should follow a uniform distribution with a height of 0.05. (A) Frequency of p values by GLMs involving clinical diagnosis. (B) Frequency of p values by GLMs involving MMSE.

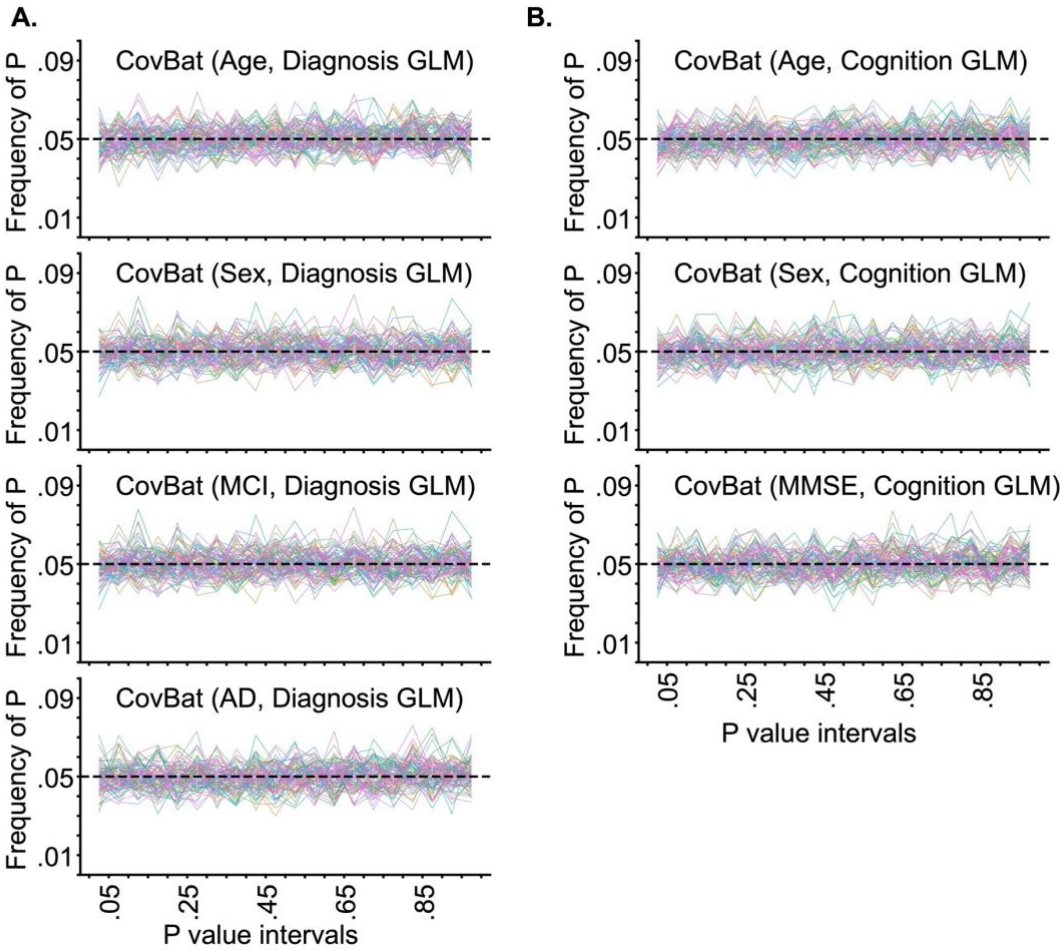

Figure S10. Frequency of p values of CovBat for matched ADNI and AIBL participants by GLMs involving clinical diagnosis and MMSE based on 1000 permutations. Each line corresponds to a single brain ROI. P values were binned in intervals of 0.05. Therefore, in the ideal scenario, the distributions of p values should follow a uniform distribution with a height of 0.05. (A) Frequency of p values by GLMs involving clinical diagnosis. (B) Frequency of p values by GLMs involving MMSE.

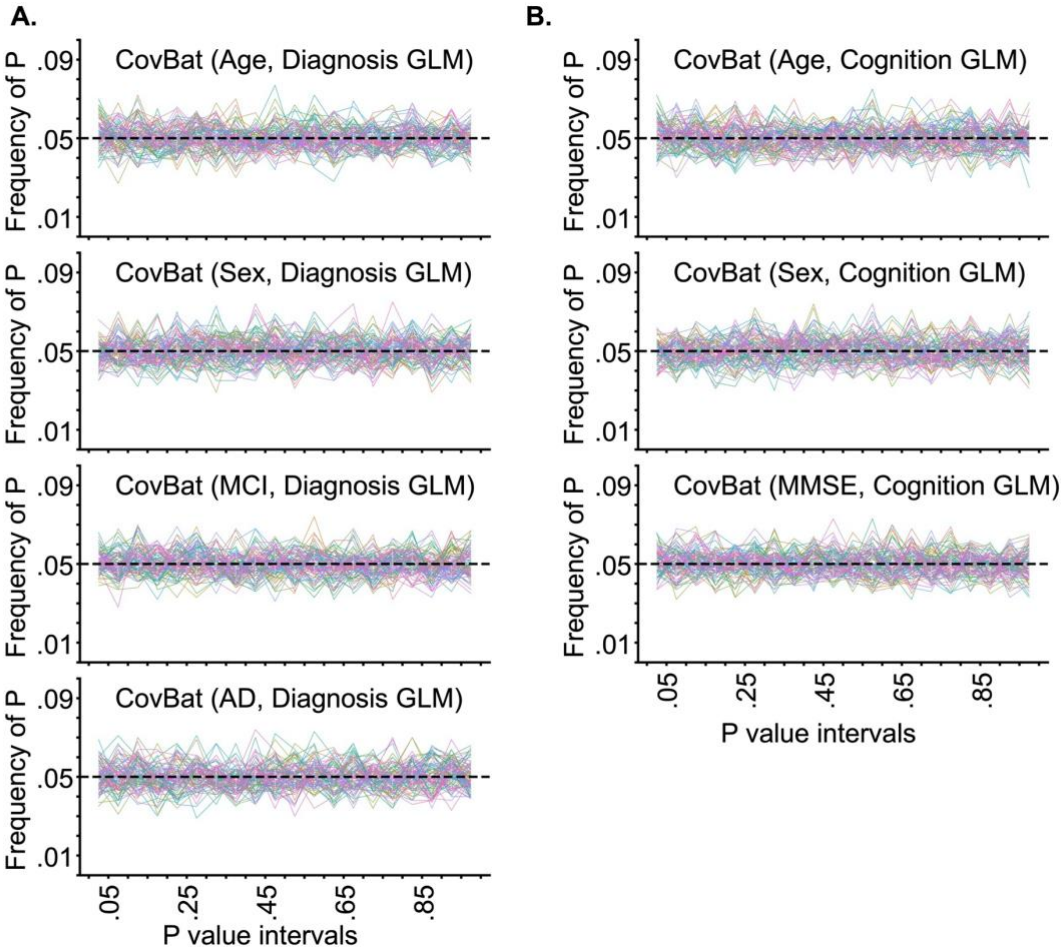

Figure S11. Frequency of p values of CovBat for matched ADNI and MACC participants by GLMs involving clinical diagnosis and MMSE based on 1000 permutations. Each line corresponds to a single brain ROI. P values were binned in intervals of 0.05. Therefore, in the ideal scenario, the distributions of p values should follow a uniform distribution with a height of 0.05. (A) Frequency of p values by GLMs involving clinical diagnosis. (B) Frequency of p values by GLMs involving MMSE.

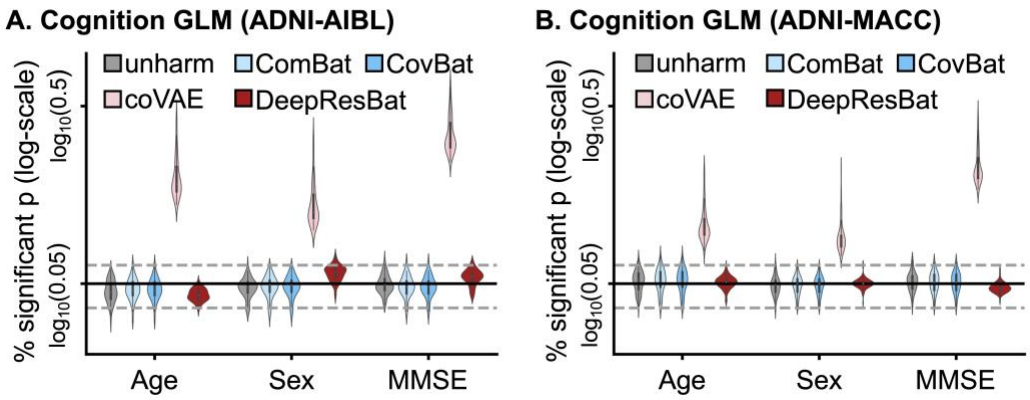

Figure S12. Percentage of nominally significant p values (i.e., p < 0.05) from GLM with MMSE after 1000 permutations of covariates. More specifically, each data point in the violin plot represents a brain ROI volume. Percentage is calculated based on the number of permutations in which p value of corresponding covariate was nominally significant (i.e., p < 0.05) divided by 1000 permutations. Percentage (vertical axis) is shown on a log scale. The black solid line is the expected percentage (which is 0.05), while the grey dashed lines indicated 95% confidence intervals. (A) GLM analysis involving MMSE for matched ADNI and AIBL participants. (B) GLM analysis involving MMSE for matched ADNI and MACC participants.

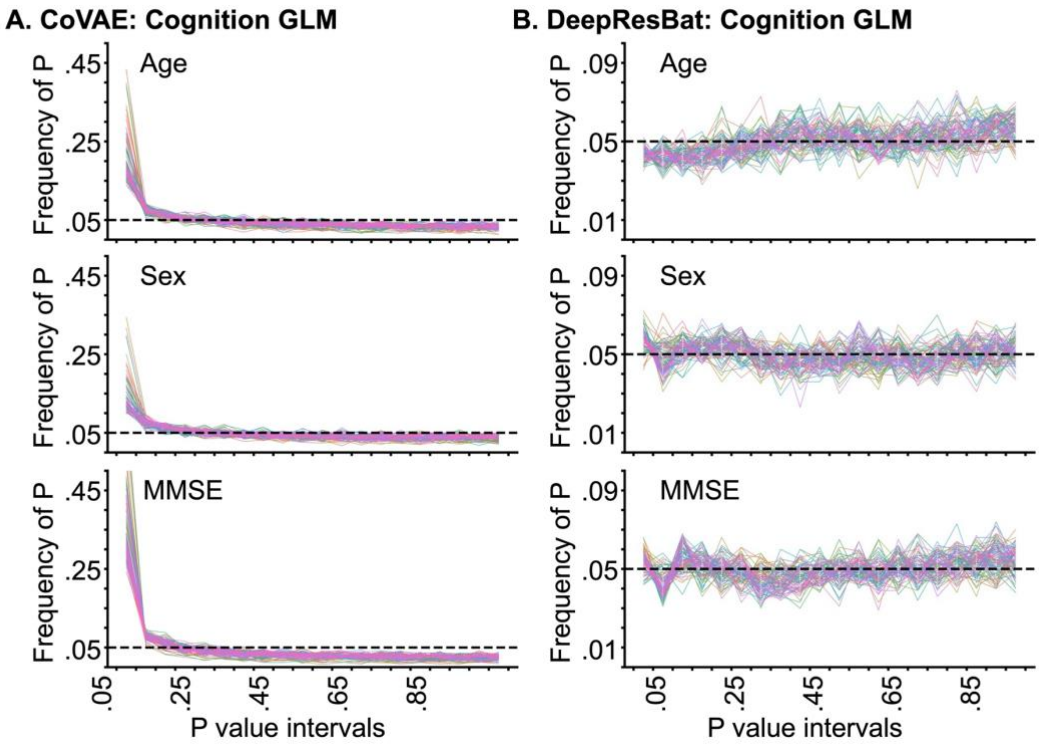

Figure S13. Frequency of p values of coVAE and DeepResBat for matched ADNI and AIBL participants by GLM involving MMSE based on 1000 permutations. Each line corresponds to a single brain ROI. P values were binned in intervals of 0.05. Therefore, in the ideal scenario, the distributions of p values should follow a uniform distribution with a height of 0.05. (A) Frequency of p values for coVAE. (B) Frequency of p values for DeepResBat.

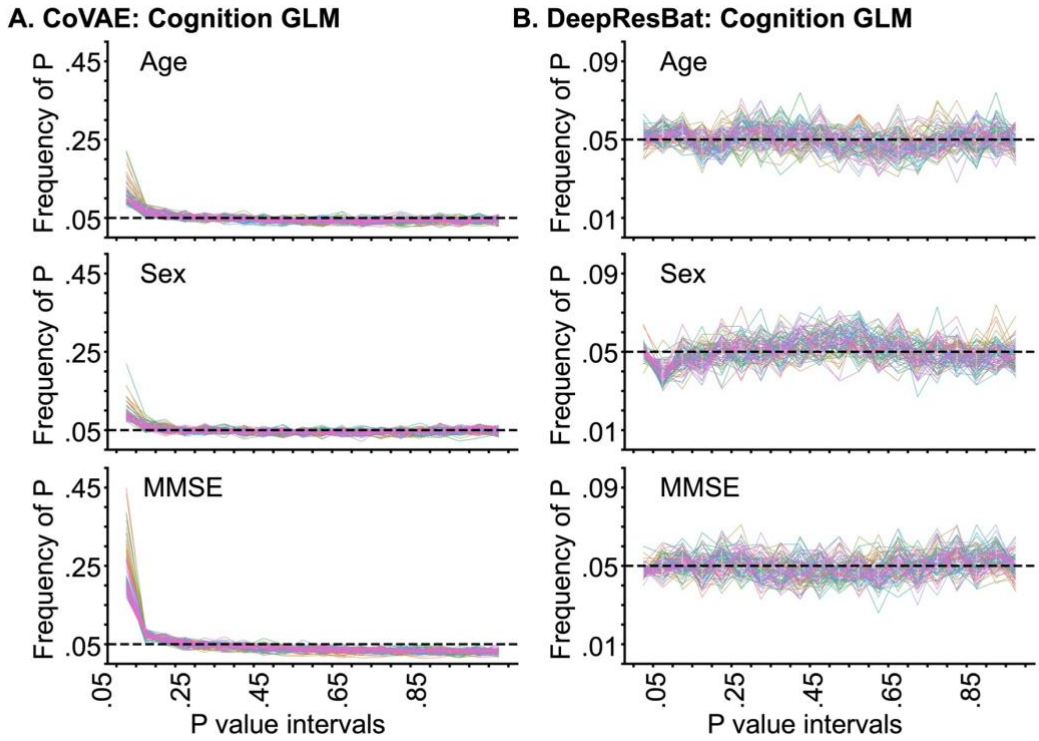

Figure S14. Frequency of p values of coVAE and DeepResBat for matched ADNI and MACC participants by GLM involving MMSE based on 1000 permutations. Each line corresponds to a single brain ROI. P values were binned in intervals of 0.05. Therefore, in the ideal scenario, the distributions of p values should follow a uniform distribution with a height of 0.05. (A) Frequency of p values for coVAE. (B) Frequency of p values for DeepResBat.

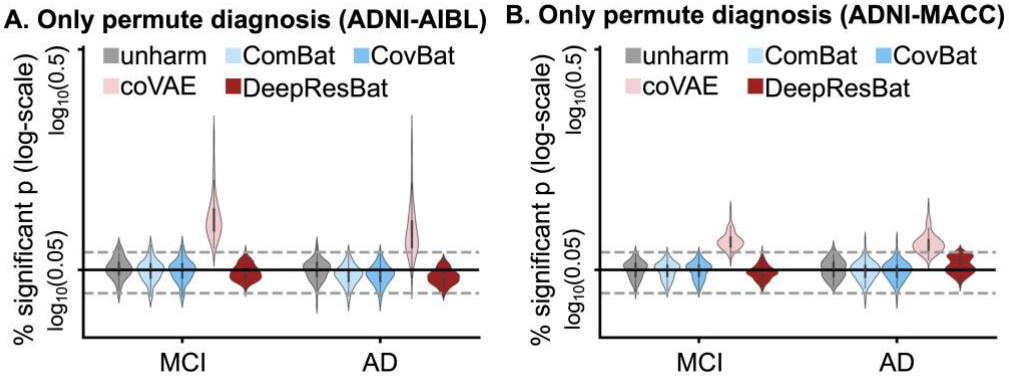

Figure S15. Percentage of nominally significant p values (i.e., p < 0.05) from GLM with clinical diagnosis after 1000 permutations of clinical diagnosis only. More specifically, each data point in the violin plot represents a brain ROI volume. Percentage is calculated based on the number of permutations in which p value of corresponding covariate was nominally significant (i.e., p < 0.05) divided by 1000 permutations. Percentage (vertical axis) is shown on a log scale. The black solid line is the expected percentage (which is 0.05), while the grey dashed lines indicated 95% confidence intervals. (A) GLM analysis involving clinical diagnosis for matched ADNI and AIBL participants. (B) GLM analysis involving clinical diagnosis for matched ADNI and MACC participants.

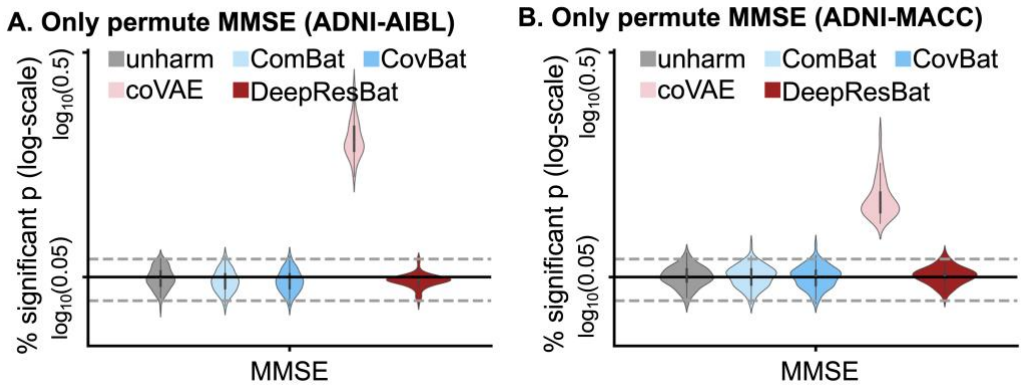

Figure S16. Percentage of nominally significant p values (i.e., p < 0.05) from GLM with MMSE after 1000 permutations of MMSE only. More specifically, each data point in the violin plot represents a brain ROI volume. Percentage is calculated based on the number of permutations in which p value of corresponding covariate was nominally significant (i.e., p < 0.05) divided by 1000 permutations. Percentage (vertical axis) is shown on a log scale. The black solid line is the expected percentage (which is 0.05), while the grey dashed lines indicated 95% confidence intervals. (A) GLM analysis involving MMSE for matched ADNI and AIBL participants. (B) GLM analysis involving MMSE for matched ADNI and MACC participants.