JYU DISSERTATIONS 736

Huashuai Xu

# Harmonization of Multi-Site MRI Data

UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF INFORMATION TECHNOLOGY

JYU DISSERTATIONS 736

Huashuai Xu

# Harmonization of Multi-Site MRI Data

Esitetään Jyväskylän yliopiston informaatioteknologian tiedekunnan suostumuksella
julkisesti tarkastettavaksi Agoran auditoriossa 2
joulukuun 18. päivänä 2023 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Information Technology of the University of Jyväskylä,
in building Agora, auditorium 2, on December 18, 2023, at 12 o'clock.

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2023

Editors
Marja-Leena Rantalainen
Faculty of Information Technology, University of Jyväskylä
Päivi Vuorio
Open Science Centre, University of Jyväskylä



ISBN 978-951-39-9884-4 (PDF)
URN:ISBN:978-951-39-9884-4
ISSN 2489-9003

Permanent link to this publication: http://urn.fi/URN:ISBN:978-951-39-9884-4

# ABSTRACT

Xu, Huashuai
Harmonization of multi-site MRI data
Jyväskylä: University of Jyväskylä, 2023, 73 p. + original articles
(JYU Dissertations
ISSN 2489-9003; 736)
ISBN 978-951-39-9884-4 (PDF)

Combining magnetic resonance imaging (MRI) data from different sites is now common to improve research with larger, more varied groups, which makes studies more powerful and representative. However, this approach faces challenges due to differences in MRI scanners that can distort results. Two methods, independent component analysis (ICA) and general linear model (GLM), are used to correct these site effects, but they struggle to fully remove them without affecting the data's real signals, especially when these signals are related to the very scanner differences they aim to correct.

In this thesis, we introduced an effective noise-reduction method utilizing the dual-projection (DP) concept grounded on independent component analysis (ICA) to mitigate site-specific influences in combined data. This method can separate the signal effects from the identified site-related components and then remove site effects without losing signals of interest. To validate the method's effectiveness, we simulated two scenarios, one where the site and signal variables are correlated and another where they are not.

Structural and functional MRI data from the Autism Brain Imaging Data Exchange II and a traveling subject dataset from the Strategic Research Program for Brain Sciences were employed to test the ICA-DP methods for removing site effects and preserving signal effects.

We also proposed an innovative multimodal denoising approach that employs a dual projection (DP) methodology grounded on linked independent component analysis (LICA) to remove the site effects. Compared with unimodal studies, using LICA on multimodal MRI data offers a more precise estimation of site effects. Structural and functional MRI data from Autism Brain Imaging Data Exchange II validated the LICA-DP methods.

In conclusion, our approaches using ICA-DP and LICA-DP have demonstrated their efficacy in mitigating site-related influences while maintaining biological variation. Such a strategy can greatly boost the validity of neuroimaging studies, and we are confident it will be an indispensable resource for forthcoming research.

Keywords: multi-site, magnetic resonance imaging, site effects, biological variability, multimodal, dual-projection, independent component analysis

# TIIVISTELMÄ (ABSTRACT IN FINNISH)

Xu, Huashuai
Monen sivuston MRI-tietojen harmonisointi
Jyväskylä: Jyväskylän yliopisto, 2023, 73 s. + alkuperäiset artikkelit
(JYU Dissertations
ISSN 2489-9003; 736)
ISBN 978-951-39-9884-4 (PDF)

Magneettikuvauksen (MRI) tietojen yhdistäminen eri paikoista on nykyisin yleistä, jotta tutkittavaksi saadaan suurempia ja monimuotoisempia ryhmiä, mikä tekee tutkimuksista tehokkaampia ja edustavampia. Kuitenkin tämä lähestymistapa kohtaa haasteita johtuen MRI-laitteiden eroista, jotka voivat vääristää tuloksia. Näiden paikkakohtaisten vaikutusten korjaamiseen käytetään kahta menetelmää, riippumattomien komponenttien analyysia (ICA) ja yleistä lineaarista mallia (GLM), mutta niillä on vaikeuksia poistaa ne täysin vaikuttamatta datan todellisiin signaaleihin, erityisesti kun nämä signaalit liittyvät juuri niihin skannerieroavaisuuksiin, joita ne pyrkivät korjaamaan.

Tässä väitöskirjassa ehdotetaan tehokasta kohinanpoistomenetelmää, joka soveltaa kaksiprojektion (DP) teoriaa riippumattoman komponenttianalyysin (ICA) pohjalta poistaakseen paikkakohtaiset vaikutukset yhdistetystä datasta. Tämä menetelmä voi erottaa signaalivaikutukset tunnistetuista paikkakohtaisista komponenteista ja poistaa sitten paikkavaikutukset menettämättä kiinnostuksen kohteena olevia signaaleja. Validoidaksemme menetelmän tehokkuuden simuloimme kaksi eri skenaariota, joissa toisessa paikka- ja signaalimuuttuja korreloivat ja toisessa eivät.

ICA-DP-menetelmiä paikkavaikutusten poistamiseksi ja signaalivaikutusten säilyttämiseksi on testattu käyttäen useita erilaisia rakenteellisia ja toiminnallisia magneettikuvausaineistoja. Väitöskirjassa esitetään myös uudenlainen monimuotoinen kohinanpoistomenetelmä paikkavaikutusten poistamiseksi, jossa kaksiprojektiomenetelmä (DP) yhdistetään linkitetyn riippumattomien komponenttien analyysin (LICA) kanssa. Yksimuotoisiin tutkimuksiin verrattuna LICA:n käyttö monimuotoisissa MRI-tiedoissa tarjoaa tarkemman arvion paikkavaikutuksista. LICA-DP-menetelmän toimivuus todennettiin olemassa olevien rakenteellisten ja toiminnallisten MRI-aineistojen avulla. ICA-DP- ja LICA-DP-menetelmät osoittautuvat tehokkaiksi tavoiksi paikkavaikutusten poistamiseksi ja biologisen vaihtelun säilyttämiseksi. Tämä lähestymistapa voi merkittävästi parantaa neurokuvantamistutkimusten validiteettia, luoden arvokkaan työkalun myös tuleville tutkimuksille.

Avainsanat: monipaikkainen, magneettikuvaus, paikkavaikutukset, biologinen vaihtelu, multimodaalinen, kaksoisprojektio, riippumattomien komponenttien analyysi

**Author**

Huashuai Xu
Faculty of Information Technology
University of Jyväskylä
Finland
Email: huashuai.xu@foxmail.com
ORCID: 0009-0005-5749-7171

**Supervisors**

Professor Tommi Kärkkäinen
Faculty of Information Technology
University of Jyväskylä
Finland

Professor Tapani Ristaniemi
Faculty of Information Technology
University of Jyväskylä
Finland

Professor Fengyu Cong
Department of Biomedical Engineering
Dalian University of Technology
China

Professor Huanjie Li
Department of Biomedical Engineering
Dalian University of Technology
China

**Reviewers**

Xinian Zuo
State Key Lab of Cognitive Neuroscience and Learning
Beijing Normal University
China

Amit K. Shukla
School of Technology and Innovations
University of Vaasa
Finland

**Opponent**

Tarmo Lipping
Faculty of Information Technology and Communication Sciences
Tampere University
Finland

## ACKNOWLEDGEMENTS

I thoroughly enjoyed my four-year doctoral studies in Finland. I would like to express my sincere gratitude to all the individuals who have contributed to the successful completion of this thesis.

First and foremost, I am deeply grateful for the financial support provided by the China Scholarship Council, China, and the Grant from the Faculty of Information Technology, University of Jyväskylä, Finland. Their support has been instrumental in facilitating my research endeavors.

I would like to extend my heartfelt appreciation to my supervisor, Prof. Tapani Ristaniemi. Throughout my research journey, Prof. Tapani consistently offered valuable guidance and constructive suggestions during our seminars. His assistance went beyond academic matters; he even helped me print a poster for an international meeting and awarded me two credits. It was an incredible honor to work alongside Prof. Tapani in Finland, and I will forever cherish his wisdom and kindness. This thesis is dedicated to commemorating Prof. Tapani for his invaluable support.

Additionally, I would like to express my deepest thanks to my supervisor, Prof. Fengyu Cong, for providing me with the opportunity to study in Finland. Their encouragement and guidance have been instrumental in shaping my academic journey.

Huanjie Li has always been my supervisor since 06.2016. She is always patient, no matter how terrible my research work is. She always strives to help me overcome my lousy research habits, cultivate a positive research attitude, and continuously encourages me to improve. Since the passing of Tapani, Tommi Kärkkäinen became my second Finnish mentor. He is really a gentleman, always with a smile. Not only does he provide me with encouragement in my research work, but he also wholeheartedly assists me in paper revisions.

I am very grateful to have Prof. Xinian Zuo (State Key Lab of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, China) and Amit K. Shukla (School of Technology and Innovations, Computer Science, University of Vaasa) as the reviewers of this dissertation. Their constructive comments and suggestions helped to improve this thesis. I would also like to sincerely thank my opponent Prof. Tarmo Lipping from Tampere University of Technology.

I would also like to sincerely thank all my friends who have supported me throughout my research and enriched my daily life. I would like to express my heartfelt gratitude to Yuxing Hao for his collaboration, which was instrumental in completing my graduation project. Additionally, I would like to extend my appreciation to Dongyue Zhou, Yunge Zhang, Chenwei Yan, and other members of Professor Li's research group for their valuable assistance. A special mention goes to Dr. Deqing Wang, Dr. Xiulin Wang, Dr. Yongjie Zhu, Dr. Rui Yan, Dr. Jia Liu, Dr. Guanghui Zhang, Dr. Wenya Liu, Dr. Xiaoshuang Wang, Dr. Dongdong Zhou, Zhonghua Chen, Reza Mahini Sheikhhosseini, Xin Zuo, Dong Tang, Xiangyu Rong, Jiaqi Zheng, and Liting Song from our research group. I am grateful for their collaboration and camaraderie. I would also like to extend my

thanks to Anni and Miki from the footbag club, as well as Daniel, Oskari, Teemu, Veikka, Vadim from the table tennis group, and Kasperi, Veera, Weiyong Xu, and Xueqiao Li from the badminton group. Especially, thanks, Daniel, for inviting me to his public defense and the dinner party. Without Daniel, table tennis is just ping-pong; with Daniel, table tennis can be everything. Thanks, my brother Veikka, for teaching me Finnish traditions, helping me buy the Finnish student overall, and chatting heart-to-heart.

Lastly, I am eternally grateful to my entire family. I would like to express my profound appreciation to my parents for their unwavering love, support, and respect for all my decisions. Their constant encouragement has been a source of strength throughout my academic journey.

Jyväskylä, October, 2023
Huashuai Xu

## ACRONYMS

| | |
|---|---|
| ABIDE | Autism Brain Imaging Data Exchange |
| ALFF | Amplitude of low frequency fluctuations |
| ANOVA | Analysis of Variance |
| ARD | Automatic relevance determination |
| ASD | Autism Spectrum Disorder |
| ComBat | Combine batches |
| CSF | Cerebrospinal fluid |
| CT | Cortical Thickness |
| DP | Dual projection |
| fALFF | Fractional amplitude of low frequency fluctuations |
| fMRI | Functional magnetic resonance imaging |
| FWHM | Full Width at Half Maximum |
| GLM | General linear modal |
| GM | Gray matter |
| HC | Healthy controls |
| ICA | Independent component analysis |
| LICA | Linked independent component analysis |
| MRI | Magnetic resonance imaging |
| OLS | Ordinary least squares |
| PSA | Pial Surface Area |
| ReHo | Regional homogeneity |
| SP | Single projection |
| tSNE | t-distributed Stochastic Neighbor Embedding |
| WM | White matter |

## FIGURES


FIGURE 1 The steps involved in the ICA-DP noise-reduction technique.....26
FIGURE 2 A summary of the preprocessing pipelines and process of LICA (Cited from (Groves et al., 2012)). .....28
FIGURE 3 Denoising effects on the signal- and noise-related components when the signal variable shows no significant correlation with the noise variable.....38
FIGURE 4 Denoising effects on the two mixed components when the signal variable displays a significant correlation with the noise variable .....39
FIGURE 5 Dimension reduction visualization by t-SNE before and after denoising for ABIDE II and traveling subject datasets.....40
FIGURE 6 The group-level evaluation concerning site effects before and after noise reduction. .....41
FIGURE 7 Age effects before and after site-effect removal. .....41
FIGURE 8 Group-level analysis of GM maps for group difference (ASD/HC) before and after data denoising.....42
FIGURE 9 Visualization of dimension reduction using t-SNE prior to and following site effects removal (Sites).....44
FIGURE 10 Group-level analysis for site effects before and after denoising. .....44
FIGURE 11 Associations between age and ALFF with different denoising strategies.....45
FIGURE 12 Associations between age and ReHo with different denoising strategies.....45
FIGURE 13 Group differences between ASD and HC before and after denoising. .....46
FIGURE 14 Sex differences before and after denoising.....47
FIGURE 15 The influence of the number of subjects on recovering spatial maps and subject courses of LICA. .....50
FIGURE 16 The influence of the number of modalities on recovering spatial maps and subject loadings .....51
FIGURE 17 The influence of the number of modalities on the linking performance. .....51
FIGURE 18 Components from LICA based on a single modality (ALFF, fALFF) and two modalities.....52
FIGURE 19 Group-level analysis for site effects before and after denoising. .52
FIGURE 20 Site effects before and after harmonization.....54
FIGURE 21 Age effects before and after harmonization. .....55
FIGURE 22 Sex effects before and after harmonization. .....56
FIGURE 23 Group differences (ASD/HC) before and after harmonization. ..57
FIGURE 24 Site effects before and after harmonization (multimodal). .....58
FIGURE 25 Age effects before and after harmonization (multimodal). .....58
FIGURE 26 Sex effects before and after harmonization (multimodal).....59

FIGURE 27 Group differences (ASD/HC) before and after harmonization (multimodal). ....59

## TABLES

TABLE 1 Scanning parameters for functional MRI data from ABIDE II. ....31
TABLE 2 Scanning parameters for structural MRI data from ABIDE II. .....32
TABLE 3 Demographic information of the multi-site ABIDE II data. .....32
TABLE 4 Scanning parameters for functional MRI data. .....33
TABLE 5 Scanning parameters for structural MRI data. .....33
TABLE 6 Component loadings are linearly correlated with signal and noise variables, while the signal variable is not significantly correlated to the noise variable. .....37
TABLE 7 Component loadings are linearly correlated with signal and noise variables, while the signal variable is significantly correlated to the noise variable. .....37
TABLE 8 Correlation between signal and subject courses and correlation among subject courses from different modalities. .....49

# CONTENTS

ABSTRACT
TIIVISTELMÄ (ABSTRACT IN FINNISH)
ACKNOWLEDGEMENTS
ACRONYMS
FIGURES AND TABLES
CONTENTS
LIST OF INCLUDED ARTICLES

1 INTRODUCTION ..... 15
1.1 Magnetic resonance imaging ..... 16
1.2 Multi-site MRI studies: advantages and challenges ..... 17
1.3 Main contributions ..... 19
1.4 Structure of the dissertation ..... 20

2 METHODS ..... 21
2.1 Harmonization methods ..... 21
2.1.1 GLM and ComBat ..... 21
2.1.2 ICA and ICA-DP ..... 25
2.1.3 LICA and LICA-DP ..... 27
2.2 Dataset ..... 30
2.2.1 Simulated data ..... 30
2.2.2 ABIDE Ⅱ ..... 30
2.2.3 Traveling subjects ..... 33
2.3 Data preprocessing ..... 34
2.4 Denoising process ..... 34
2.5 Evaluate the denoising results ..... 35

3 OVERVIEW OF INCLUDED ARTICLES ..... 36
3.1 Article I: Removal of site effects and enhancement of signal using dual projection independent component analysis for pooling multi-site MRI data ..... 36
3.2 Article II: Harmonization of multi-site functional MRI data with dual-projection based ICA model ..... 43
3.3 Article III: Enhancing performance of linked independent component analysis: Investigating the influence of subjects and modalities ..... 48
3.4 Article IV: Harmonization of multi-site MRI data with dual-projection based Linked ICA model ..... 53

4 DISCUSSION ..... 61
4.1 Findings of multi-site MRI data harmonization methods ..... 61
4.1.1 ICA-DP ..... 61

4.1.2 LICA-DP ..... 62
4.2 Limitations ..... 64
4.3 Future directions ..... 64

5 CONCLUSION ..... 65

YHTEENVETO (SUMMARY IN FINNISH) ..... 66

REFERENCES ..... 67

ORIGINAL PAPERS

## LIST OF INCLUDED ARTICLES

I Hao, Yuxing, Huashuai Xu, Mingrui Xia, Chenwei Yan, Yunge Zhang, Dongyue Zhou, Tommi Kärkkäinen, Lisa D. Nickerson, Huanjie Li, and Fengyu Cong (2023). Removal of site effects and enhancement of signal using dual projection independent component analysis for pooling multi-site MRI data. European Journal of Neuroscience, 58(6): 3466-3487, https://doi.org/10.1111/ejn.16120

II Xu, Huashuai, Yuxing Hao, Yunge Zhang, Dongyue Zhou, Tommi Kärkkäinen, Lisa D. Nickerson, Huanjie Li, and Fengyu Cong (2023). Harmonization of multi-site functional MRI data with dual-projection based ICA model. Frontiers in Neuroscience, 17, 1225606, https://doi.org/10.3389/fnins.2023.1225606

III Huashuai Xu, Tommi Kärkkäinen, Huanjie Li, and Fengyu Cong (2023). Enhancing performance of linked independent component analysis: investigating the influence of subjects and modalities. In 2023 International Conference on Computers, Information Processing and Advanced Education (CIPAE), pp. 726-732. IEEE.

IV Huashuai Xu, Yuxing Hao, Yunge Zhang, Dongyue Zhou, Tommi Kärkkäinen, Lisa D. Nickerson, Huanjie Li, and Fengyu Cong. Harmonization of multi-site MRI data with dual-projection based Linked ICA model. To be submitted, 2023.

# 1 INTRODUCTION

Since its beginning in the early 1990s, magnetic resonance imaging (MRI) has become a popular tool for understanding the structure and function of the human brain and detecting brain diseases (Eklund et al., 2016). MRI encompasses various modalities, each with its unique capabilities. For example, functional magnetic resonance imaging (fMRI) measures brain activity by detecting changes in blood flow and oxygenation levels; structural MRI provides high-resolution images of the brain's structure and captures the differences in tissue properties, such as gray matter (GM), white matter (WM), cerebrospinal fluid (CSF), to create detailed, three-dimensional images of the brain's anatomy. With its non-invasive nature, MRI has revolutionized our understanding of the brain and continues advancing our cognitive science knowledge.

Most neuroimaging studies typically take place within a single research site, which often limits the ability to gather large datasets with ample sample sizes. This limitation, resulting from small sample sizes, may impede the detection of true differences (type II errors) and even increase the likelihood of detecting false differences (type I errors) due to lenient thresholds (Radua et al., 2020). The use of small sample sizes and the lack of harmonization across independent studies pose challenges in achieving satisfactory reliability and reproducibility in neuroimaging research (Nichols et al., 2017). Given this, multi-site studies have gained acceptance to overcome these limitations by providing increased statistical power to detect group differences and longitudinal changes, ultimately enhancing reliability and reproducibility.

In this chapter, we will first briefly introduce MRI data. Then, we will introduce multi-site data pooling, including its advantages and challenges. Finally, we will illustrate the motivation of the conducted research.

## 1.1 Magnetic resonance imaging

As a non-invasive imaging technique in recent years, MRI has played an increasingly important role in investigating brain structure, function, development, and pathologies, with the increasing flexibility and power to answer scientifically interesting and clinically relevant questions (Smith et al., 2004).

In the context of neuroimaging research, MRI has become an indispensable tool for studying the human brain and its complex mechanisms. It enables researchers to investigate various aspects of brain function, such as neural activity, brain connectivity, and anatomical abnormalities. By providing detailed images with exceptional spatial resolution, MRI allows for identifying and characterizing brain regions and their involvement in various cognitive processes and neurological disorders.

One of the critical advantages of MRI is its versatility. It offers multiple imaging modalities that can be employed to examine different aspects of brain structure and function. Structural MRI (sMRI) provides detailed anatomical information, allowing researchers to study brain morphology and detect structural changes associated with diseases or developmental processes. Functional MRI (fMRI) measures changes in blood flow and oxygenation levels, providing insights into temporally changing brain activity and connectivity during different tasks or resting states (Sui, Adali, et al., 2012; Sui, Yu, et al., 2012).

Over the past thirty years, numerous techniques and models have been developed to harness and interpret fMRI data. One commonly used method is the frequency-domain analysis. Y. F. Zang and colleagues introduced the concept of the amplitude of low-frequency fluctuations (ALFF) for a voxel's time series, which quantifies the overall signal power within the 0.01–0.1 Hz low-frequency range (Y. F. Zang et al., 2007). The fractional ALFF (fALFF), a subsequent refinement of ALFF, was presented by Zou et al. (2008), focusing on the proportion of power in the low-frequency band relative to the entire frequency spectrum. Additionally, regional homogeneity analysis, known as ReHo, has gained traction (Y. Zang et al., 2004). ReHo offers a voxel-centric assessment of brain activity by examining the synchronization of a voxel's time series with its immediate neighbors, utilizing Kendall's coefficient of concordance. This metric is instrumental in probing the localized synchronization of fMRI signals within the brain (Y. Zang et al., 2004).

Compared to other imaging techniques, Electroencephalography (EEG), and Magnetoencephalography (MEG), MRI offers several advantages: 1) It provides detailed anatomical visualization of the brain, allowing for the identification of structures and abnormalities with excellent spatial resolution. 2) MRI combines functional and structural imaging capabilities, integrating insights into brain activity and connectivity with anatomical details. With its non-invasiveness and safety, MRI offers a comprehensive approach to understanding the human brain.

However, despite its numerous advantages, there are challenges associated with MRI studies. The high initial investment, ongoing operational costs, and maintenance requirements are important considerations when utilizing MRI for medical imaging. The majority of neuroimaging research is carried out at individual sites, which often restricts the potential to gather extensive datasets. In Andy W. K. Yeung' survey (Yeung, 2018), in all 388 papers, 138 (35.6%) studies analyzed data from 25 or fewer subjects, and only 21 studies (5.4%) had 101 or more subjects. Studies in a single research site with small sample-size datasets have an increased risk of producing false positives or missing true effects (low statistical power), wasting research funding on studies with a low chance of achieving their objective (Szucs & Ioannidis, 2017). Thus, using small sample sizes can generate irreproducible results.

With the advancement of neuroimaging methods and information technology, leveraging big data has gained traction in neuroimaging research for many reasons. First, bigger sample sizes bolster statistical power and yield more dependable outcomes (Yu et al., 2018). Second, they better reflect broader populations, making the findings more universally relevant with more significant real-world impact (A. A. Chen et al., 2022). Lastly, samples ranging from several to tens of thousands in emerging research domains are frequently essential to pinpoint subtle effects (Takao et al., 2014; Han et al., 2023).

Using data from multiple sites is often inevitable to obtain data with a large sample size for neuroimaging studies. Consequently, multi-site MRI research is on the rise, enhancing the robustness of statistical evaluations to discern group variations, track long-term changes, and also bolster the consistency and replicability of neuroimaging investigations.

## 1.2 Multi-site MRI studies: advantages and challenges

To address these challenges from single-site studies, multi-site studies have gained prominence, wherein data from different research sites are pooled together to increase sample sizes and enhance statistical power.

In recent years, there has been a significant emphasis on collecting multi-site MRI data in neuroimaging. Several notable studies have emerged that aim to leverage the advantages of pooling data from multiple research sites to gain insights into the human brain. We list here some studies:

- The Adolescent Brain Cognitive Development (ABCD) study (https://abcdstudy.org/): a large-scale, longitudinal, multi-site investigation aimed at understanding the impact of biological and environmental factors on developmental outcomes.
- Enhancing NeuroImaging Genetics Through Meta-Analysis (ENIGMA) study (https://enigma.ini.usc.edu/): a collaborative network of researchers working together on a range of large-scale studies that

integrate data from 70 institutions all over the world (Thompson et al., 2014).

- Autism Brain Imaging Data Exchange I & II (ABIDE I & II) dataset (https://fcon_1000.projects.nitrc.org/indi/abide/): a consortium openly sharing 2156 datasets, including 1112 (536 individuals with Autism Spectrum Disorder (ASD) and573 age-matched controls) from ABIDE I, and 1044 (487 individuals with ASD and 557 controls) from ABIDE II (Di Martino et al., 2014, 2017)
- Depression Imaging REsearch Consor-Tium (DIRECT) dataset (https://rfmri.org/REST-meta-MDD): 2428 functional brain images processed with a standardized pipeline across all participating sites (X. Chen et al., 2022; Yan et al., 2019).

The ABCD study focuses on brain structure and function throughout childhood and adolescence, utilizing multimodal magnetic resonance imaging (MRI) data acquired from 29 different scanners at 21 locations across the United States (Casey et al., 2018). One significant advantage of ENIGMA consortium is the implementation of standardized protocols for preprocessing MRI data, which effectively reduces the heterogeneity across sites due to methodological factors. Consistent preprocessing pipelines are applied across all sites to obtain multiple MRI modalities (Radua et al., 2020).

Joint efforts across multiple sites provide a golden chance to gather broader and more varied participant groups, amplifying the research's statistical strength and making the findings more reflective of the broader population. By pooling data from multiple sites, researchers can access a broader range of participants, including individuals from different demographic backgrounds, geographical locations, and clinical populations. This increased sample size and diversity strengthens the statistical analyses and allows for more robust generalizations and a better understanding of the underlying phenomena being investigated.

Nevertheless, integrating MRI data from multiple sites brings the challenge of non-biological sources of variability known as site effects. These site effects arise from the inherent differences in MRI scanners, including field strength, manufacturer specifications (e.g., Siemens vs. GE), models (e.g., Philips Ingenia vs. Philips Achieva), software versions, and various imaging parameters (Dudley et al., 2023). Even when MRI data collection parameters are ideally matched across locations, a feat that's often difficult or unattainable, site-related discrepancies remain an inherent aspect of multi-site investigations. (Dudley et al., 2023; Fortin et al., 2017, 2018).

Site effects can confound the interpretation of results and mask or distort the actual underlying biological effects for nearly all MRI modalities, including

- T1-Weighted Imaging (T1WI) (J. Chen et al., 2014; Fortin et al., 2018; Maikusa et al., 2021; Parekh et al., 2022; Radua et al., 2020);
- Diffusion-Weighted Imaging (DWI) (Fortin et al., 2017);
- functional MRI (fMRI) (Biswal et al., 2010; Groves et al., 2011; Li et al., 2020);

- Magnetic Resonance Spectroscopy (MRS) (Bell et al., 2022).

These site effects can introduce unwanted biases into the data, potentially overshadowing genuine biological effects and leading to spurious findings (Onicas et al., 2022). They have also been shown to be significantly more straightforward to detect than biological effects by statistical analysis and machine learning methods (Hu et al., 2023).

## 1.3 Main contributions

Consequently, careful consideration and mitigation strategies are essential to account for and minimize the impact of site effects when pooling data across different sites. Due to the complex nature of site effects, traditional statistical methods (inclusion of site in a lineal modal as a mean effect) for adjusting the confounders are inadequate to account for site effects sufficiently (Hu et al., 2023). Two main techniques, independent component analysis (ICA) and general linear modal (GLM), for removing site effects face challenges when completely eradicating site effects and preserving signal effects when signals of interest (e.g., age, sex) correlate with site-related variables. In order to better remove site effects and protect signal effects, the motivations of our studies can be concluded from two aspects.

When dealing with single-modality MRI data, we often encounter a hurdle wherein the conventional independent component analysis (ICA) methods are insufficient to mitigate the site effects completely. This variability may lead to confounding results in our analysis if not accounted for. We propose a novel approach involving a dual projection (DP) method to address this issue more effectively (Hao et al., 2023; Xu, Hao, et al., 2023). This method can separate the signal effects correlated with site variables from the identified site effects for removal without losing signals of interest. Simulated, vivo structural and functional MRI data from Autism Brain Imaging Data Exchange II and a traveling subject dataset from the Strategic Research Program for Brain Sciences, were used to test the ICA-DP methods. This method is designed to provide a more robust and comprehensive removal of these site effects, thereby enhancing the reliability and reproducibility of MRI data analysis.

When dealing with multimodal MRI data, We implement DP in linked independent component analysis (LICA) and denoise site effects from multi-modality data. The advantage of multimodal fusion is that it can capitalize on the strength of each modality in a joint analysis compared with a separate analysis of each (Xu, Li, et al., 2023). Firstly, LICA is used to identify more site-related noise components from multimodal MRI data. Then we use DP method to separate the signal effects from the identified site effects. In this way, we can remove more site effects without losing signals of interest. A dataset from Autism Brain Imaging Data Exchange II and a traveling subject dataset from the Strategic

Research Program for Brain Sciences were used to test the proposed LICA-DP denoising method.

To evaluate the efficiency of the harmonization methods, we employ a range of methodologies to both visualize and quantify site-related effects prior to and following the denoising process. Furthermore, we appraise the performance of the denoising techniques based on their capacity to retain the integrity of the signal effects.

## 1.4 Structure of the dissertation

The structure of this dissertation is listed as follows: Chapter 1 introduces the advantages and challenges of multi-site MRI studies. In Chapter 2, we introduce the basic multi-site MRI harmonization methods. Chapter 3 briefly summarizes the included articles and lists the contributions of the authors to the articles. Chapter 4 presents the discussion and conclusion of this dissertation, as well as the research limitations and future directions. Chapter 5 presents the conclusion of this dissertation.

# 2 METHODS

In this chapter, we first introduce various multi-site data harmonization methods, including the approach we proposed. Subsequently, we describe the datasets utilized to validate the effectiveness of these methods, the preprocessing of the data, and the multi-site harmonization process. Finally, we delineate the criteria employed to assess the efficacy of the multi-site data harmonization methods.

## 2.1 Harmonization methods

Many statistical and machine learning methods have been developed to model, attenuate, or eliminate site effects (Descoteaux et al., 2022; Hu et al., 2023). Machine learning models can be challenging to interpret. They often act as 'black boxes', providing little insight into how they are making their decisions. This can be problematic when trying to understand the significance of the findings. So, we only introduce and compare different statistical methods here, including their foundations and critical advantages and disadvantages.

### 2.1.1 GLM and ComBat

The first attempts utilize sites as variates in a general linear model (GLM), also referred to as residual harmonization. GLM method adjusts the images for site effects via linear regression without considering the confounding between site and signal variables (e.g., age, sex) (Fortin et al., 2018).

**Statistical foundation**

The GLM model can be written as follows:

$$Y_{\text{non-denoised}} = X_{\text{sites}}\beta_{\text{sites}} + \varepsilon\,, \tag{1}$$

where $Y_{\text{non-denoised}}$ is the original data, $X_{\text{sites}}$ is the design matrix for the site effects, the corresponding regression coefficient $\beta_{\text{sites}}$ can be obtained using ordinary

least squares (OLS), $\varepsilon$ is the residual. The modification is carried out by deducting the site-associated term:

$$Y_{\text{denoised}}^{\text{GLM}} = Y_{\text{non-denoised}} - X_{\text{sites}}\beta_{\text{sites}} . \quad (2)$$

The adjusted GLM (AdGLM) method utilizes sites as covariates in a general linear model.

$$Y_{\text{non-denoised}} = X_{\text{signal}}\beta_{\text{signal}} + X_{\text{sites}}\beta_{\text{sites}} + \varepsilon , \quad (3)$$

where $X_{signal}$ and $\beta_{signal}$ are the design matrix and corresponding regression coefficient of the biological signal. By removing the site effects, the denoised data by AdGLM is:

$$Y_{\text{denoised}}^{\text{AdGLM}} = Y_{\text{non-denoised}} - X_{\text{sites}}\beta_{\text{sites}} . \quad (4)$$

ComBat (Fortin et al., 2017, 2018; Johnson et al., 2007) is a GLM-derived method based on the empirical Bayes approach. The ComBat approach was initially proposed for microarray gene expression data (Johnson et al., 2007); its main objective was to refine the location/scale model for limited data sets. The technique assumes that the data can be modeled as a linear combination of signal variables and site influences, encompassing both additive and multiplicative factors:

$$Y_{\text{non-denoised}} = \alpha + X_{\text{signal}}\beta_{\text{signal}} + \gamma + \delta\varepsilon , \quad (5)$$

where $\alpha$ is the average value, $X_{\text{signal}}$ is the design matrix for the signal variables and $\beta_{\text{signal}}$ is the corresponding regression coefficient, $\gamma$ and $\delta$ are the additive and multiplicative factors, respectively. Subsequently, ComBat standardizes the data by eliminating the impacts of mean and signal variables:

$$\text{Y}_{\text{normalized}} = Y_{\text{non-denoised}} - \alpha - X_{\text{signal}}\beta_{\text{signal}} . \quad (6)$$

In the end, ComBat employs an empirical Bayes (EB) methodology to obtain a refined estimation of the site-specific adjustment factor $\gamma^*$ and site-specific scaling factor $\delta^*$. In detail, this Bayesian approach generalizes the AdGLM approach described above, incorporating empirical priors over the site-specific means and variances. This integration results in partial pooling across the features (Descoteaux et al., 2022). After eliminating these site-specific influences and reintegrating the effects of the average and signal variables, we ultimately obtain the denoised data via ComBat:

$$Y_{\text{denoised}}^{\text{ComBat}} = \frac{Y_{\text{non-denoised}} - \alpha - X_{\text{signal}}\beta_{\text{signal}} - \gamma^*}{\delta^*} + \alpha + X_{\text{signal}}\beta_{\text{signal}} . \quad (7)$$

**Advantages and disadvantages**

GLM is simple and easy to implement. However, its major limitation is that it neglects to incorporate biological signals of interest. This oversight inevitably results in losing biologically meaningless and meaningful biases while regressing out site effects. Consequently, its application is generally not recommended.

The adjusted GLM can be readily applied in any statistical framework, facilitating various regression evaluations. As demonstrated in Equation 4, if the

design matrix of the site effects $X_{\text{sites}}$ and biological signal design matrix $X_{\text{signal}}$ are orthogonal and uncorrelated with each other, estimating and interpreting site effects is relatively straightforward. Under such circumstances, the site merely functions as an additive effect, which can be readily estimated and eliminated without influencing the signal effects (J. Chen et al., 2014).

Nevertheless, neuroimaging studies incorporating data from multiple sites often unmet this ideal condition. For instance, many individual neuroimaging samples are confined to a specific age range, which results in a correlation between age and site effects. In such contexts, removing an estimate of the site effects may inadvertently eliminate variations of biological signals. To be more specific, consider a hypothetical study across two sites exploring mild cognitive impairment in senior individuals. In this scenario, Site A typically recruits participants who are older than those at Site B, and Scanner A has a propensity to gauge cortical thickness higher than Scanner B. Without accounting for scanner effects, the impact of age would likely be misestimated. This underestimation arises as the variable of interest (i.e., age) is associated with the image feature (i.e., cortical thickness), and a site/scanner effect is apparent in both. When employing the AdGLM on the data that incorporates both scanner and age variables, the collinearity of these variables may induce instability in estimating the regression coefficients, the signal effects $X_{\text{signal}}$ 'compete' with the site effects $X_{\text{sites}}$ and diminish their association with the dependent information. As a result, the model might either overstate or understate the influence of these variables. Although meticulous study planning and participant recruitment can mitigate the issue of multicollinearity, it is often challenging, if not impossible, to entirely eradicate selection bias in multi-site studies (Descoteaux et al., 2022; Dudley et al., 2023). Nygaard et al. (2016) showed that the AdGLM method resulted in overconfident group outcomes with uneven samples.

In summary, the AdGLM technique may decrease statistical power and obscure genuine impacts when a significant variable within the study group is not uniformly spread across scanning devices, leading to multicollinearity in the general linear model(Dudley et al., 2023).

ComBat first demonstrated its practical utility when applied to voxel-level fractional anisotropy (FA) values derived from two diffusion MRI datasets (Fortin et al., 2017). Subsequent research corroborated ComBat's effectiveness on various neuroimaging features, including cortical thickness (Fortin et al., 2018) and functional connectivity (Yu et al., 2018). Since its initial introduction and validation, ComBat has achieved extensive recognition and utilization in the field of MRI imaging (Beer et al., 2020; Bell et al., 2022; Da-ano et al., 2020; Eshaghzadeh Torbati et al., 2021; Horng et al., 2022; Meyers et al., 2022; Orlhac et al., 2021; Radua et al., 2020). From 2017 onwards, more than 50 imaging research projects have cited ComBat as their preferred method for site effect rectification (Orlhac et al., 2021). ComBat demands minimal computational resources. Moreover, its implementations are available in R, Matlab, and Python, catering to higher-dimensional data types like voxel-based morphometry. This underscores its widespread adoption by the neuroimaging community

(Descoteaux et al., 2022). ComBat enhances the reliability of estimated parameters in smaller samples using a Bayesian approach. It is more effective in mitigating site effects while preserving the biologically relevant effects than AdGLM.

While ComBat has emerged as a promising tool for harmonization across multi-site neuroimaging studies, it is essential to note that it is also with its limitations. First, ComBat inherently presumes that the signal characteristics of various features originate from a uniform distribution, having a consistent mean and variance for every site. However, this assumption may not be tenable if the target distribution exhibits heteroscedasticity. Heteroscedasticity refers to the condition where the standard deviation of the predicted variable is inconsistent. For instance, the variable of cortical thickness may display a higher degree of variance in older individuals than in younger ones.

Second, sites with smaller data sets are subjected to more extensive regularization compared to those with larger data sets. This can result in unbalanced adjustments, particularly when there are significant discrepancies in sample sizes between sites. For example, The site-specific shift factor $\gamma^*$ in a data set with two sites with N1= 10000 samples and N2= 100 will be different from the $\gamma^*$ of two sites with N1= 500 and N2= 500 (Descoteaux et al., 2022).

Further, like AdGLM, ComBat can not deal with the collinear problem. ComBat operates on the assumption that site effects are independent of the signal, and it does not account for site-by-signal interactions. This implies that ComBat is ideally suited for scenarios where the biological effects presumed to influence the variables of interest are uniformly distributed across sites, thus enabling their estimation across all subjects. However, issues arise when there is strong collinearity between biological signals and a site. For instance, consider a scenario where one site predominantly contains a cohort of young subjects while one site contains older ones. ComBat's model's assumptions are violated in such situations, leading to potentially inaccurate and biased adjustments.

### 2.1.2 ICA and ICA-DP

Independent component analysis (ICA) is a data-driven computational method that decomposes the data matrix into a set of statistically independent non-Gaussian maps and associated courses (e.g., time, subject). The method is often applied to digital images, in signal processing, and in many other areas. ICA is a type of blind source separation (BSS) technique. 'Blind' because it operates without a model of the source signals, and "source separation" because the aim is to extract the original source signals from a set of mixed signals.

**Statistical foundation**

Firstly, the ICA method decomposes the data into independent spatial maps and their corresponding loadings.

$$Y_{\text{non-denoised}} = A * S\,, \tag{8}$$

where *S* is the spatial map, and A is the corresponding loadings. For juxtaposition against our ICA-DP technique, we refer to the conventional ICA as ICA-SP (single projection). To maintain the signal effects, the ICA-SP approach solely discards components purely associated with site effects and retains the mixed components untouched.

$$Y_{\text{denoised}}^{\text{ICA-SP}} = Y_{\text{non-denoised}} - A_{\text{sites}} pinv(A_{\text{sites}}) Y_{\text{non-denoised}}\,. \tag{9}$$

where $A_{\text{sites}}$ is the course of pure site-related components, the pinv() function refers to the computation of the Moore-Penrose pseudoinverse of a matrix. The pseudoinverse is a generalization of the matrix inverse for matrices that may not be square or invertible.

The algorithm of ICA consists of these general steps: 1) Centering: The observed signals are centralized by subtracting their means; 2) Whitening: This step transforms the data so that potential correlations are removed, and the variance for each component is equalized. A popular whitening method is principal component analysis (PCA); 3) Algorithm Iteration: In this stage, an iterative algorithm is used to maximize the statistical independence of the estimated components.

To eradicate the site effects, we introduced the ICA-DP technique (Hao et al., 2023). Initially, ICA-DP distinguishes the signal effects from the mixed components (during the first projection phase):

$$A'_{\text{sites}} = A_{\text{mixed}} - Var_{\text{signal}} pinv\left(Var_{\text{signal}}\right) A_{\text{mixed}}\,, \tag{10}$$

where $A_{\text{mixed}}$ is the course of mixed components and $Var_{\text{signal}}$ is the signal variable. Then [ $A_{\text{sites}}$ $A'_{\text{sites}}$ ] is utilized as the whole site effects to be regressed out (second projection procedure).

$$Y_{\text{denoised}}^{\text{ICA-DP}} = Y_{\text{non-denoised}} - [\, A_{\text{sites}}\, A'_{\text{sites}}\, ] pinv([\, A_{\text{sites}}\, A'_{\text{sites}}\, ]) Y_{\text{non-denoised}}. \tag{11}$$

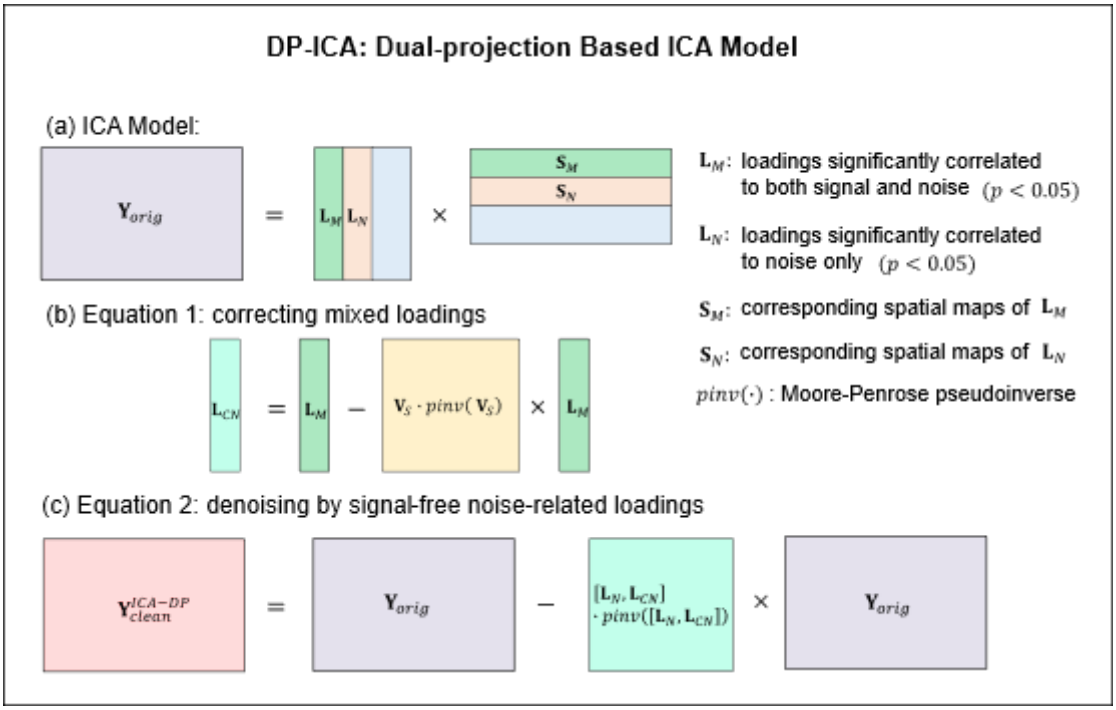


FIGURE 1 The steps involved in the ICA-DP noise-reduction technique (a) Identifying the loadings that are associated with noise variables retrieved by ICA. This includes mixed loadings that show significant correlation to noise and signal and those that are solely significantly linked to noise. (b) Refining the mixed loadings to only those correlated with noise ($\mathbf{L}_{CN}$) by projecting out the signal-related data. (c) Securing cleaned data by excluding the comprehensive noise-associated components ($\mathbf{L}_N, \mathbf{L}_{CN}$)

**Advantages and disadvantages**

ICA has been widely used to identify and remove structured noise components from fMRI signals, such as head motion-related (McKeown et al., 2003), physiological (McKeown et al., 1998), and scanner-induced noise (J. Chen et al., 2014; Feis et al., 2015).

As a data-driven approach, 1) ICA does not assume a specific statistical model and, therefore, can be more flexible in handling complex and unknown distributions of the data; 2) ICA can separate the MRI data into independent spatial components. This ability to decompose the data into its essential elements can be beneficial in identifying and removing site effects; 3) ICA has the ability to separate mixed signals into their original sources without requiring prior knowledge of the sources or the mixing process. This is particularly useful when the cause of the site effects is unknown.

Compared with GLM-based methods, which assume a constant effect for each site and ignore within-site day-to-day variations in these effects, ICA facilitates the determination and removal of site effects via data-driven instead of formulating covariates to represent the site effects grounded in robust presumptions.

However, there are also assumptions of ICA: 1) The sources are statistically independent of each other; 2) The sources are non-Gaussian; 3) There are at least as many observations (mixed signals) as there are sources.

The greatest challenge in using ICA to remove site effects is that the components obtained from ICA are often correlated with both the site and the biological signal (named mixed components). Generally, we do not modify these mixed components to preserve biological variabilities. Consequently, this leads to incomplete removal of site effects.

To address this issue, we introduce the ICA-DP method to mitigate site effects. Within ICA-DP, mixed components from ICA are divided into segments exclusively linked to signal and those solely connected to site discrepancies, using a projection process. Site effects drawn from mixed components through this projection phase are amalgamated with other distinct site-related components. These are then eliminated from the data via a second projection step.

### 2.1.3 LICA and LICA-DP

In 2011, Adrian R. Groves proposed the Linked independent component analysis (LICA) method (Groves et al., 2011). LICA is a multivariate data analysis method that allows researchers to identify common features across multiple modalities. LICA was developed to extend the capabilities of standard ICA and allow for data integration from multiple modalities.

**Statistical foundation**

The main highlight of LICA is that it treats multiple datasets simultaneously rather than considering them individually. These datasets can be different imaging modalities (such as structural and functional MRI) from the same subjects or the same imaging modality acquired under different conditions. LICA can find shared or linked components across different datasets, which can help identify common underlying biological or physiological effects.

LICA is designed to capture covariance trends across various modalities. It facilitates simultaneous ICA decompositions on diverse modalities while ensuring the subject weights remain the same across them. In this sense, LICA provides a kind of "joint" ICA solution. This approach can lead to more interpretable results, as it takes advantage of the relationships among the datasets to find common patterns of variability. Moreover, LICA balances the information content from different modalities, even permitting a modality to be excluded from a specific component. A distinct feature of a LICA component includes its spatial maps (one for each modality) and the subject loadings that are shared across modalities (shown in Figure 2) and may involve multiple (named multimodal components) or only modality (unimodal components). Whereas the spatial maps indicate the spatial variability at the group level, the subject loadings shed light on a subject's particular influence on the component (Doan et al., 2017).

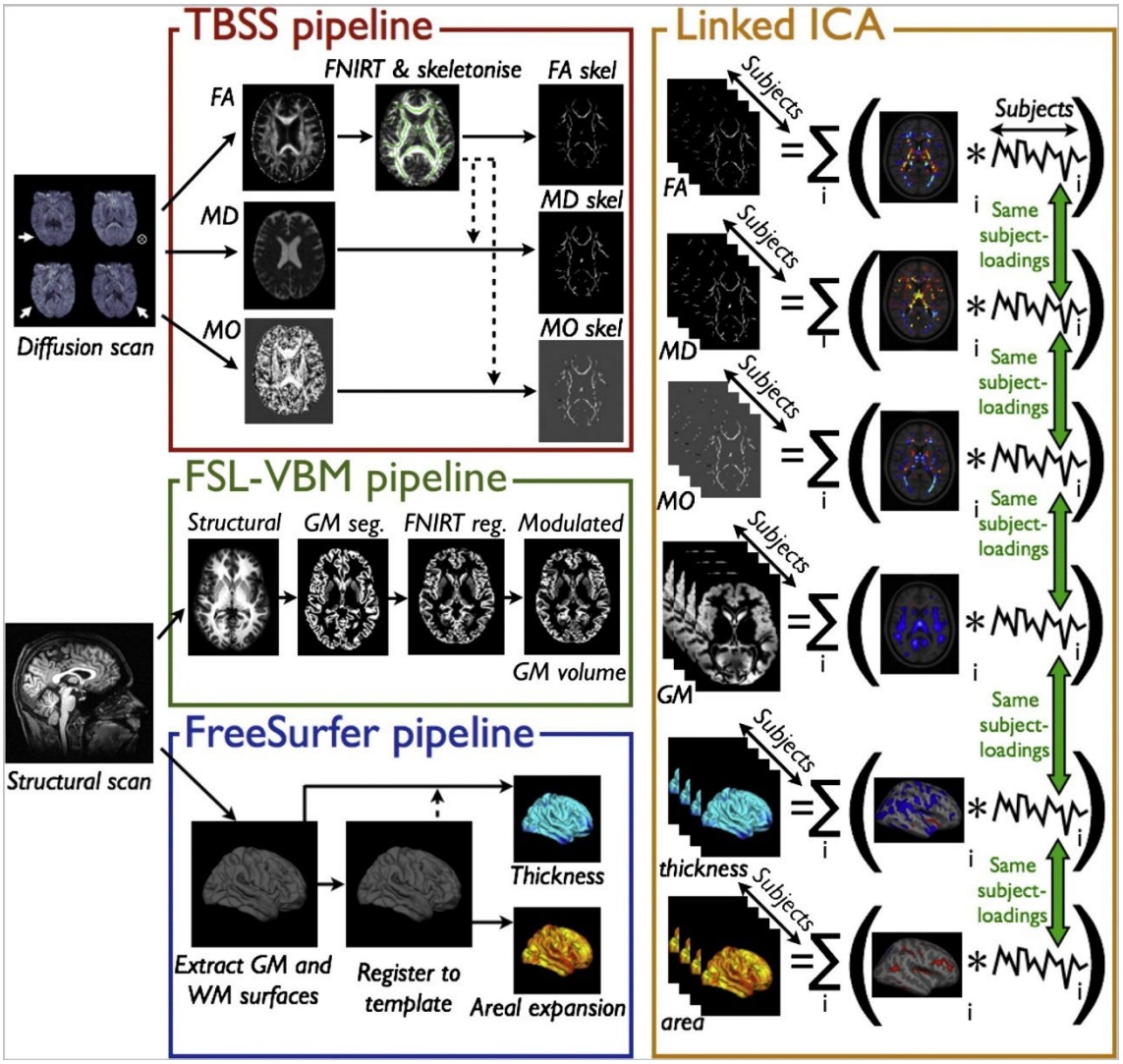


FIGURE 2 A summary of the preprocessing pipelines and process of LICA (Cited from (Groves et al., 2012)).

Since its introduction, LICA has found extensive applications in neuroimaging research (Sui, Adali, et al., 2012; Sui, Yu, et al., 2012). Adrian R. Groves utilized LICA to scrutinize age-related effects across multiple imaging modalities (Groves et al., 2012).

Li et al. (2020) proposed a denoising technique for multimodal imaging metrics that utilized LICA as a novel approach to mitigate site influences from multi-study data, and the novel method has demonstrated superior performance over the standard GLM and conventional single-modality ICA denoising methods in mitigating site effects in multimodal MRI data.

To retain the signal effects, the traditional LICA method (we rename the traditional LICA as LICA-SP for comparison with the LICA-DP method below) selectively eliminates only those components exclusively related to site effects and leaves mixed components (those correlated with both the site and the signal) untouched.

$$Y_{\text{denoised}}^{\text{LICA-SP}} = Y_{\text{non-denoised}} - A_{\text{sites}} pinv(A_{\text{sites}}) Y_{\text{non-denoised}} . \quad (12)$$

To remove the site effects more thoroughly, we have introduced the LICA-DP method, which aims to implement DP in LICA and denoise site effects from multi-modality data. Firstly, LICA is used to identify more site-related noise components from multimodal MRI data. Then we use DP method to separate the signal effects from the identified site effects.

$$Y_{\text{denoised}}^{\text{LICA-DP}} = Y_{\text{non-denoised}} - [\, A_{\text{sites}} \, A'_{\text{sites}} \,] pinv([\, A_{\text{sites}} \, A'_{\text{sites}} \,]) Y_{\text{non-denoised}}. \quad (13)$$

**Advantages and disadvantages**

The advantages of the LICA method stem from the benefits of multimodal fusion. Multimodal fusion refers to using a common symmetric model that explains different sorts of data. Different modalities can provide different types of information, giving a more comprehensive view of the subject matter. For instance, in neuroimaging, structural MRI can provide detailed anatomical information, functional MRI can provide information about brain activity, and diffusion MRI can provide information about white matter connectivity. A primary motivation for multimodal fusion is to take advantage of the cross-information provided by diverse imaging methods, which in turn can be helpful for uncovering patterns of related changes across various modalities when they exist.

Relative to traditional GLM and single-modality ICA denoising techniques, denoising methods rooted in LICA are superior in eliminating site-associated influences (Li et al., 2020). The superior performance of LICA is attributable to its distinctive linkage function. The linkage function in linked ICA is a way to create a shared association between different modalities. The idea is to assume that a common latent variable or factor links different types of data. The linkage function quantifies this relationship by constraining the subject weights to be the same across modalities. It ensures that the independent components (ICs) derived from the different modalities are meaningfully related or 'linked' to each other. In simpler terms, imagine two different brain images (one showing brain structure and another showing brain function) for the same subjects. These are two different modalities. If certain structural changes in the brain are associated with specific functional changes, the LICA's linkage function helps capture and represent that relationship. So, instead of treating these two types of data as entirely separate, the linkage function allows LICA to analyze them connected, revealing insights that might be missed if each modality were analyzed independently. In this way, LICA can identify components more related to site difference (Li et al., 2020; Xu, Li, et al., 2023), thus more effectively modeling and eliminating site effects.

In addition, LICA is a Bayesian ICA method, differing from traditional ICA methods like FastICA (Hyvärinen & Oja, 2000). LICA directly integrates dimensionality reduction into the ICA methodology by applying automatic relevance determination (ARD) priors on the components (Bishop, 1999; Roberts, 2001). The eliminated components (or part-components) are removed from the

model during the iteration. This step precludes any further inference on these spatial maps bearing zero weight, making the process more efficient (Groves et al., 2011).

A primary challenge in employing LICA to eliminate site effects is the frequent correlation of the derived components with both the site and the biological signal, resulting in what we refer to as mixed components. Typically, to retain biological variability, we refrain from modifying these mixed components, which unfortunately leads to incomplete removal of site effects.

To address this issue, we introduce the LICA-DP method designed to remove site effects effectively. In the LICA-DP framework, mixed components generated by LICA are dissected into parts solely related to the signal and exclusively to site differences, achieved by applying a projection process. Site effects isolated from the mixed components via this projection step are added with other pure site-related components. These are then cleared from the data using a second projection procedure.

## 2.2 Dataset

Our research utilized simulated, structural, and functional MRI data to test our proposed ICA-DP and LICA-DP methods.

### 2.2.1 Simulated data

Our research incorporated simulated data in two aspects: 1) Firstly, we evaluated the impact of varying degrees of correlation between signal and noise variables on the ICA-DP method. Two distinct types of relationships were simulated between subject loadings and signal/noise variables: i) The signal variable was not significantly correlated with the noise variable; however, subject loadings were linearly correlated with either the signal, noise variables, or both. ii) The signal variable was significantly correlated with the noise variable, with subject loadings being linearly correlated with either the signal, noise variables, or both. 2) Secondly, we assessed the influence of the number of subjects and modalities on the performance of LICA.

The spatial components used in the simulations were derived from standard brain templates (ThomasYeoLab), and details on subject courses can be found in Chapter 3.

### 2.2.2 ABIDE Ⅱ

The Autism Brain Imaging Data Exchange (ABIDE II) is an initiative to broaden the horizons of brain connectomics research (http://fcon_1000.projects.nitrc.org/indi/abide/abide_II.html) in Autism Spectrum Disorder (ASD). As a continuation of the initial ABIDE effort (ABIDE I), which released 1112 datasets in 2012 (Di Martino et al., 2014), this expanded

multi-site open-data resource comprises resting-state functional magnetic resonance imaging (MRI), accompanying structural MRI, and phenotypic datasets. Images manifesting conspicuous artifacts, extensive head movement (exceeding one voxel size), or incomplete scanning of the whole brain were excluded from our study. After strict quality control, we obtained Cortical Thickness (CT) and Pial Surface Area (PSA) data for 952 subjects (comprising 428 Autism Spectrum Disorder (ASD) patients and 524 Healthy Controls (HC)), Grey Matter (GM) data for 913 subjects (ASD: 402, HC: 511), and functional MRI (fMRI) data for 795 subjects (ASD: 341, HC: 454). Data were gleaned from 18 different sites, encompassing various manufacturers such as Siemens, Philips, and GE. The specific acquisition parameters are detailed in Tables 1 and 2, while demographic information is presented in Table 3.

TABLE 1 Scanning parameters for functional MRI data from ABIDE II. The data were collected from 18 different sites: Erasmus University Medical Center (EMC), ETH Zürich (ETH), Georgetown University (GU), Indiana University (IU), Kennedy Krieger Institute (KKI), Katholieke Universiteit Leuven (KUL), Oregon Health and Science University (OHSU), Olin Neuropsychiatry Research Center (ONRC), Stanford University (SU), University of California Davis (UCD), University of California Los Angeles (UCLA), University of Miami (UM), University of Utah School of Medicine (USM), Barrow Neurological Institute (BNI), Institut Pasteur and Robert Debré Hospital (IP), NYU Langone Medical Center (NYU), San Diego State University (SDSU), Trinity Centre for Health Sciences (TCD).

| Sites | Scanners | TR/TE (ms) | FA (degree) | Voxel Size |
|---|---|---|---|---|
| EMC | GE MR750 | 2000/30 | 85 | 3.6 × 3.6 × 4.0 |
| ETH | Philips Achieva | 2000/25 | 90 | 3 × 3 × 3 |
| GU | Siemens TriTim | 2000/30 | 90 | 3 × 3 × 3 |
| IU | Siemens TriTim | 813/28 | 60 | 3.4 × 3.4 × 3.4 |
| KKI | Philips Achieva | 2500/30 | 75 | 3 × 3 × 3 |
| KUL | PhilipsAchieva | 2500/30 | 90 | 1.6 × 1.6 × 3.1 |
| OHSU | Siemens TriTim | 475/30 | 60 | 3 × 3 × 3 |
| ONRC | Siemens Skyra | 2500/30 | 90 | 3.8 × 3.8 × 3.8 |
| SU | GE SIGNA | 2000/30 | 80 | 3.4 × 3.4 × 3.5 |
| UCD | Siemens TriTim | 2000/24 | 90 | 3.5 × 3.5 × 3.5 |
| UCLA | Siemens TriTim | 3000/28 | 90 | 3 × 3 × 4 |
| UM | GE Healthcare | 2000/30 | 75 | 3.4 × 3.4 × 3.4 |
| USM | Siemens TriTim | 2000/28 | 90 | 3.1 × 3.1 × 4 |
| BNI | Philips Ingenia | 3000/25 | 80 | 3.8 × 3.8 × 4 |
| IP | Philips Achieva | 2700/45 | 90 | 3.6 × 3.7 × 4 |
| NYU | Siemens Allegra | 2000/15 | 90 | 3 × 3 × 4 |
| SDSU | GE MR750 | 2000/30 | 90 | 3.4 × 3.4 × 3.4 |
| TCD | Philips Achieva | 2000/27 | 90 | 3 × 3 × 3.2 |

TABLE 2 Scanning parameters for structural MRI data from ABIDE II.

| Sites | Scanners | TR/TE (ms) | FA (degree) | Voxel Size |
|---|---|---|---|---|
| EMC | GE MR750 | 1664/4.24 | 16 | 0.9 × 0.9 × 0.9 |
| ETH | Philips Achieva | 3000/3.9 | 8 | 0.9 × 0.9 × 0.9 |
| GU | Siemens TriTim | 2530/3.5 | 7 | 1 × 1 × 1 |
| IU | Siemens TriTim | 2400/2.3 | 8 | 0.7 × 0.7 × 0.7 |
| KKI | Philips Achieva | 3500/3.7 | 8 | 1 × 1 × 1 |
| KUL | PhilipsAchieva | 2000/4.6 | 8 | 1 × 1 × 1.2 |
| OHSU | Siemens TriTim | 2300/3.58 | 10 | 1 × 1 × 1.1 |
| ONRC | Siemens Skyra | 2200/2.88 | 13 | 0.8 × 0.8 × 0.8 |
| SU | GE SIGNA | 5.9/1.8 | 11 | 1 × 1 × 1 |
| UCD | Siemens TriTim | 2000/3.16 | 8 | – |
| UCLA | Siemens TriTim | 2300/2.86 | 9 | 1 × 1 × 1.2 |
| UM | GE Healthcare | -/- | 12 | 1 × 1 × 1 |
| USM | Siemens TriTim | 2300/2.91 | 9 | 1 × 1 × 1.2 |
| BNI | Philips Ingenia | 2500/3.1 | 9 | 1.1 × 1.1 × 1.1 |
| IP | Philips Achieva | 2500/5.6 | 30 | 1 × 1 × 1 |
| NYU | Siemens Allegra | 2530/3.25 | 7 | 1.3 × 1.3 × 1.3 |
| SDSU | GE MR750 | 2683/3.17 | 8 | 1 × 1 × 1 |
| TCD | Philips Achieva | 3000/3.9 | 8 | 0.9 × 0.9 × 0.9 |

TABLE 3 Demographic information of the multi-site ABIDE II data.

| Sites | Structural (ASD/HC) | CT | PSA | fMRI |
|---|---|---|---|---|
| EMC | 38(18/20) | 38(18/20) | 38(18/20) | 27(14/13) |
| ETH | 25(8/17) | 32(9/23) | 32(9/23) | 29(7/22) |
| GU | 76(33/43) | 77(33/44) | 77(33/44) | 68(27/41) |
| IU | 36(18/18) | 37(18/19) | 37(18/19) | 37(18/19) |
| KKI | 165(32/133) | 165(32/133) | 165(32/133) | 148(25/123) |
| KUL | 7(7/0) | 27(27/0) | 27(27/0) | 25(25/0) |
| OHSU | 84(33/51) | 88(35/53) | 88(35/53) | 84(33/51) |
| ONRC | 45(16/29) | 45(16/29) | 45(16/29) | 41(15/26) |
| SU | 32(15/17) | 32(15/17) | 32(15/17) | 31(14/17) |
| UCD | 26(13/13) | 26(13/13) | 26(13/13) | -- |
| UCLA | 24(12/12) | 24(12/12) | 24(12/12) | 24(12/12) |
| UM | 19(7/12) | 19(7/12) | 19(7/12) | -- |
| USM | 29(13/16) | 32(16/16) | 32(16/16) | 21(9/12) |
| BNI | 55(29/26) | 58(29/29) | 58(29/29) | 57(29/28) |
| IP | 53(22/31) | 53(22/31) | 53(22/31) | 34(13/21) |
| NYU | 104(75/29) | 104(75/29) | 124(75/29) | 89(61/28) |
| SDSU | 57(32/25) | 57(32/25) | 57(32/25) | 54(30/24) |
| TCD | 38(19/19) | 38(19/19) | 38(19/19) | 26(9/17) |
| Total | 913(402/511) | 952(428/524) | 952(428/524) | 795(341/454) |

### 2.2.3 Traveling subjects

We utilized a traveling subject dataset from the DecNef Project Brain Data Repository (https://bicr-resource.atr.jp/srpbsts/), compiled by the Strategic Research Program for the Promotion of Brain Science (SRPBS) (Tanaka et al., 2021; Yamashita et al., 2019). This dataset included nine healthy male participants, ranging in age from 24 to 32 years. Each participant underwent T1-weighted MRI scans across 12 distinct centers. These centers employed 3T scanners from various manufacturers: Siemens, GE, and Philips.

The advantage of the traveling subject dataset is that the subjects in each site are the same, excluding the influence of other variables (e.g., age, gender). Three sites (ATT, UTO, and YC2) were excluded because of the duplicate data.

TABLE 4 Scanning parameters for functional MRI data.

| Sites | Scanners | TR/TE (ms) | FA (degree) | Voxel Size |
|---|---|---|---|---|
| ATT | SiemensTimTrio | 2500/30 | 80 | 3.3 × 3.3 × 3.2 |
| ATV | Siemens Verio | 2500/30 | 80 | 3.3 × 3.3 × 3.2 |
| COI | Siemens Verio | 2500/30 | 80 | 3.3 × 3.3 × 3.2 |
| HUH | GE Signa HDxt | 2500/30 | 80 | 3.3 × 3.3 × 3.2 |
| HKH | Siemens Spectra | 2500/30 | 80 | 3.3 × 3.3 × 3.2 |
| KPM | Philips Achieva | 2500/30 | 80 | 3.3 × 3.3 × 3.2 |
| SWA | Siemens Verio | 2500/30 | 80 | 3.3 × 3.3 × 3.2 |
| KUT | SiemensTimTrio | 2500/30 | 80 | 3.3 × 3.3 × 3.2 |
| KUS | Siemens Skyra | 2500/30 | 80 | 3.3 × 3.3 × 3.2 |
| UTO | GE MR750W | 2500/30 | 80 | 3.3 × 3.3 × 3.2 |
| YC1 | Philips Achieva | 2500/30 | 80 | 3.3 × 3.3 × 3.2 |
| YC2 | Philips Achieva | 2500/30 | 80 | 3.3 × 3.3 × 3.2 |

TABLE 5 Scanning parameters for structural MRI data.

| Sites | Scanners | TR/TE (ms) | FA (degree) | Voxel Size |
|---|---|---|---|---|
| ATT | SiemensTimTrio | 2300/2.98 | 9 | 1 × 1 × 1 |
| ATV | Siemens Verio | 2300/2.98 | 9 | 1 × 1 × 1 |
| COI | Siemens Verio | 2300/2.98 | 9 | 1 × 1 × 1 |
| HUH | GE Signa HDxt | 1900/2.38 | 10 | 0.8 × 0.75 × 0.75 |
| HKH | Siemens Spectra | 6788/1.928 | 20 | 1 × 1 × 1 |
| KPM | Philips Achieva | 7.1/3.31 | 10 | 1 × 1 × 1 |
| SWA | Siemens Verio | 2300/2.98 | 9 | 1 × 1 × 1 |
| KUT | SiemensTimTrio | 2000/3.4 | 8 | 0.9375 × 0.9375 × 1 |
| KUS | Siemens Skyra | 2300/2.98 | 9 | 1 × 1 × 1 |
| UTO | GE MR750W | 7.7/3.1 | 11 | 1 × 1.0156 × 1.0156 |
| YC1 | Philips Achieva | 6.99/3.176 | 9 | 1 × 1 × 1 |
| YC2 | Philips Achieva | 7.01/3.155 | 9 | 1 × 1 × 1 |

## 2.3 Data preprocessing

The preprocessing workflow for structural and functional MRI data employed widely accepted steps.

**Structural MRI:** Grey Matter (GM) images were created from high-spatial resolution structural MR images using FSL-VBM (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSLVBM). The process began with removing non-brain tissue, followed by GM segmentation. GM images were then non-linearly registered to the MNI 152 standard space. The resulting images were concatenated and averaged to form a study-specific grey matter template. Subsequently, all native GM images were non-linearly registered to this study-specific template. The modulated GM images were then smoothed using an isotropic Gaussian kernel (sigma = 3mm).

**fMRI:** The raw fMRI data underwent preprocessing with FSL FEAT (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FEAT), which included removing the first six volumes, motion correction, and spatial normalization to the standard MNI space. Two functional modalities, ALFF and ReHo, were generated from the preprocessed fMRI data using DPABI (Yan et al., 2016). For ReHo, spatial smoothing (with Full Width at Half Maximum (FWHM) of 6 mm) was performed after the ReHo calculation. In contrast, for ALFF, spatial smoothing was conducted prior to the calculation (Jia et al., 2019).

## 2.4 Denoising process

We utilized the Matlab version of ComBat (https://github.com/Jfortin1/ComBatHarmonization) in our study, and the input $X_{signal}$ was set as group differences (ASD/HC), age, and sex. We utilized FSL MELODIC (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/MELODIC) for the ICA-related analysis. First, ICA was utilized on the original data to categorize components into pure noise, pure signal, and mixed categories. We used the Pearson correlation coefficient and Analysis of Variance (ANOVA) to identify signal, noise, and mixed components. For numerical variables such as group differences (ASD/HC), age, and sex, we determined the nature of the components by calculating the Pearson correlation coefficient between subject loadings and these variables. Since we cannot calculate the Pearson correlation from categorical variables, we used ANOVA to calculate the correlation between loadings and site differences. To be specific, we grouped the loadings that came from the same site together. Then, we used a one-way ANOVA with multiple (the number of sites) levels to determine the correlation between loadings and site differences. Components only associated with the signal variable ($p < 0.05$, Bonferroni correction) were identified as pure signal components, while those only related to the noise variable were designated as pure noise components.

Components related to signal and noise variables were categorized as mixed components. Only pure noise components were utilized for the ICA-SP method to regress out the site effects. On the other hand, for the ICA-DP method, all components linked to noise, encompassing the mixed ones, were used to mitigate the site influences. The noise impacts derived from both the mixed and purely noise components are deemed the combined site-associated noise effects targeted for elimination by the ICA-DP technique.

We downloaded the LICA tool from the FSL website (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLICA), and after unzipping, we placed it within the Matlab directory (https://uk.mathworks.com/). As LICA can autonomously determine the optimal number of components needed to describe the data, we preset a large number of components, allowing LICA to downweight and discard the weaker components gradually. The identification of site-related components is the same as that used in ICA-related methods.

## 2.5 Evaluate the denoising results

We employed a series of analyses to compare our DP-related denoising methods in terms of their ability to eliminate site effects and retain signal effects.

For the visualization of site effects, we employed t-distributed Stochastic Neighbor Embedding (tSNE) (Van Der Maaten & Hinton, 2008) to examine the distribution of data points, assessing whether they tended to cluster by site. tSNE is a machine learning algorithm designed for visualization of high-dimensional data by projecting it into a two-dimensional space, while preserving the pairwise similarities of the original data points as closely as possible. It is particularly well-suited for the visualization of complex datasets in fields like bioinformatics or speech analysis, where the data can have hundreds or even thousands of dimensions.

Group-level F-test was also implemented to identify regions significantly varied due to site differences. The evaluation utilized a generalized linear model execution of a one-way ANOVA, with the site as the factor and age, sex, and group distinction (ASD/HC) as covariates.

Demonstrating that a harmonization technique effectively eliminates site effects is crucial, and it is equally important to prove the preservation of biological variability in the data. A method that eradicates both site effects and biological effects is scientifically unproductive (Fortin et al., 2018). We utilized age, sex, and group differences (ASD/HC) as variables of interest to assess the preservation of biological variability in the various harmonization methods discussed in this study. Besides t-SNE and group-level tests, the Pearson correlation coefficient between median image measures and age of all the subjects was used to show the relationship.

All results presented in the study utilized Matlab (https://uk.mathworks.com/), BrainNet (Xia et al., 2013), and FSLeyes (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSLeyes) for visualization.

# 3 OVERVIEW OF INCLUDED ARTICLES

This chapter presents the overview of each study, including methods, main results, and the author's contributions.

## 3.1 Article I: Removal of site effects and enhancement of signal using dual projection independent component analysis for pooling multi-site MRI data

Hao, Yuxing, Huashuai Xu, Mingrui Xia, Chenwei Yan, Yunge Zhang, Dongyue Zhou, Tommi Kärkkäinen, Lisa D. Nickerson, Huanjie Li, and Fengyu Cong (2023). Removal of site effects and enhancement of signal using dual projection independent component analysis for pooling multi-site MRI data. European Journal of Neuroscience, 58(6): 3466-3487, https://doi.org/10.1111/ejn.16120

**Methods**

ICA tends to identify components that represent a blend of signal and noise instead of separating them into distinct components. In the past work, only those components linked with scanner effects but not correlated with signals of interest were removed to maintain relevant signals. So, traditional ICA is regarded as a conservative denoising method since mixed components are retained.

To overcome this challenge in ICA-based methods, we propose a new technique called ICA with dual-projection (ICA-DP) technique for mitigating site effects. ICA-DP separates mixed components derived from ICA into segments solely associated with the signal and those solely linked to noise using a projection process. The noise effects isolated from the mixed components are combined with the other ICA components that reflect site variance and removed from the data with a second projection step.

We evaluate this new technique with simulated MRI data and in vivo multi-site datasets, and compare the performance of ICA-DP against traditional ICA and ComBat denoising methods. For the simulated data, the relationships

between subject courses and signal/noise variables are shown in Tables 6 and 7. For the structural MRI data, we utilized data from the first 13 sites within the ABIDE database and 9 sites in the traveling-subject dataset.

TABLE 6 Component loadings are linearly correlated with signal and noise variables, while the signal variable is not significantly correlated to the noise variable.

| #Component | Signal Variable (r/p) | Noise Variable (r/p) |
|---|---|---|
| 1 | 0.9425(0) | 0.3999(3.8e-5) |
| 2 | 0.2999(2.4e-3) | 0.9728(0) |
| 3 | -- | 0.5999(4.2e-11) |
| 4 | 0.5999(4.2e-11) | -- |

TABLE 7 Component loadings are linearly correlated with signal and noise variables, while the signal variable is significantly correlated to the noise variable.

| Correlation between signal and noise | #Component | Signal | Noise |
|---|---|---|---|
| 0.2999 (2.4e-3) | 1 | 0.7946(0) | 0.2412(1.6e-2) |
| | 2 | 0.2590(9.3e-3) | 0.7959(0) |
| 0.4999 (1.2e-7) | 1 | 0.7962(0) | 0.4279(8.9e-6) |
| | 2 | 0.3447(4.5e-4) | 0.7859(0) |
| 0.6999 (5.6e-16) | 1 | 0.7957(0) | 0.5932(7.8e-11) |
| | 2 | 0.4993(1.2e-7) | 0.7761(0) |

**Results**

Figure 3 presents the components tied to signal and noise isolated by ICA from hypothetical data before and after noise reduction, given the lack of significant correlation between signal and noise variables. Figure 3(a) depicts outcomes for spatially independent data, whereas Figure 3(b) outlines results when the initial two components spatially intersect. In the absence of a correlation between signal and noise variables, noise reduction outcomes are analogous for both spatially independent and dependent datasets. Every denoising technique proficiently eliminates the standalone noise component #3 and retains the standalone signal component #4. Yet, the ICA-SP approach falls short in purging noise influences from the mixed components #1 (more attuned to the original data's signal) and #2 (more attuned to the original data's noise). In such contexts, ICA-DP, GLM, and ComBat exhibit similar efficiencies. The mixed components #1 and #2 witness a purging of noise influences, with an amplification of the signal effect by enhancing its correlations with the signal variable post-denoising through ICA-DP, GLM, and ComBat. These two mixed components are consolidated into one, predominantly tied to the signal variable. Areas linked to noise are also eradicated post-denoising with ICA-DP, GLM, and ComBat.

Figure 4 showcases the denoising effects on the two mixed components when a significant correlation exists between the signal and noise variables. Three distinct correlation intensities between these variables are modeled. Figure 4(a) outlines the outcomes for spatially independent data, whereas Figure 4(b)

displays the results when the first two components are spatially overlapped. Of all denoising methodologies, solely ICA-DP adeptly diminishes noise while amplifying the signal effects, regardless of the correlation between signal and noise variables in both modeled datasets.

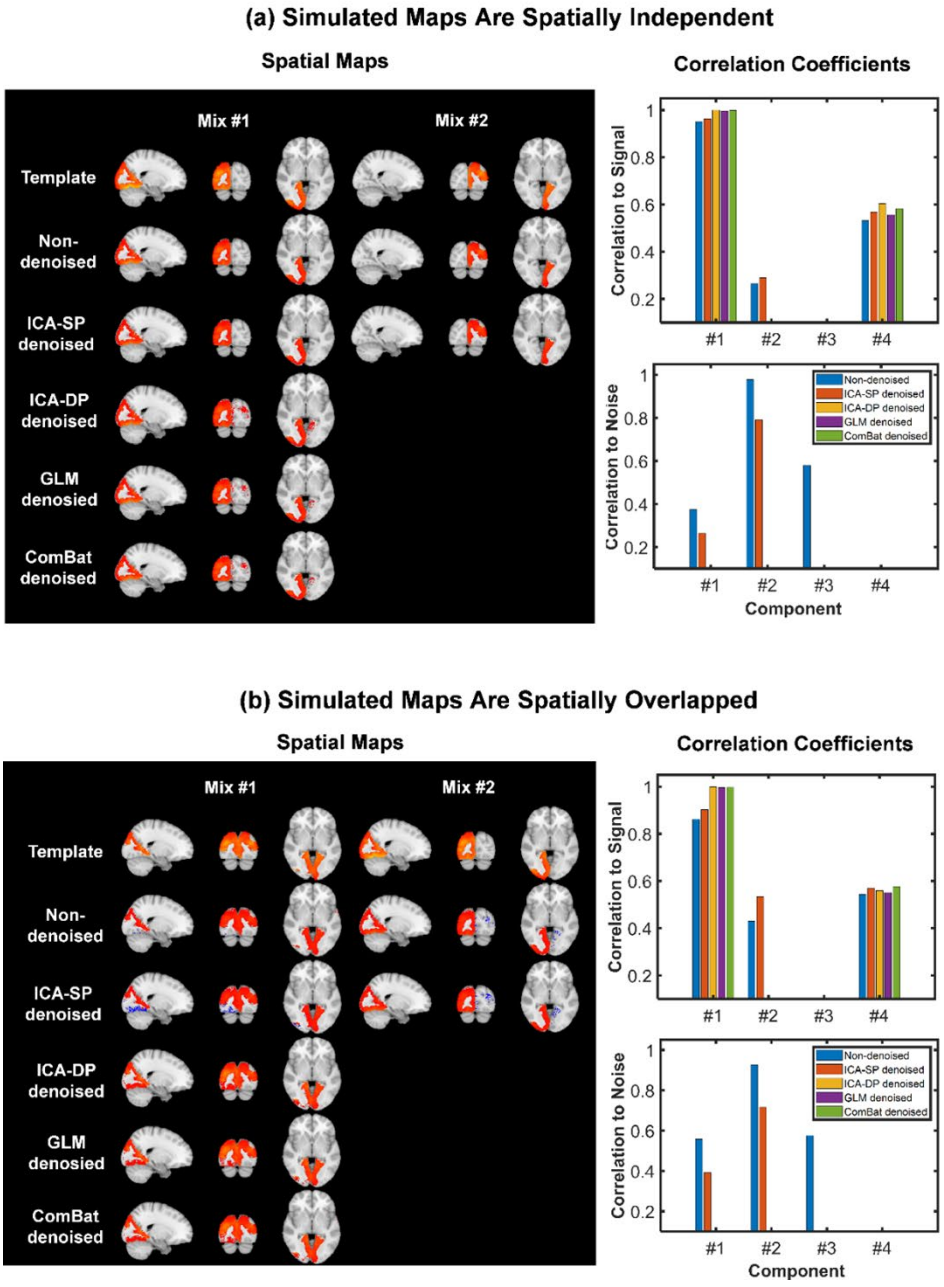


FIGURE 3 Denoising effects on the signal- and noise-related components when the signal variable shows no significant correlation with the noise variable. (a) When the spatial maps of all 10 components are spatially independent, (b) Similar patterns are observed when the spatial maps of the first two components are spatially overlapped.


FIGURE 4 Denoising effects on the two mixed components when the signal variable displays a significant correlation with the noise variable. Mix #1 is more associated with the signal variable, while Mix #2 is more related to the noise variable in the non-denoised data. (a) In the scenario where the spatial maps of all 10 components are spatially independent. (b) Similar denoising outcomes are observed when the spatial maps of the first two components are spatially overlapped.

Figures 5 and 6 show the site effects before and after harmonization. The site effects for both datasets globally affect the non-denoised GM data. While the ICA-SP approach has diminished the site effects, its removal has not been adequate. Once processed by both ICA-DP and ComBat, there are no notable regions linked to site variables in either dataset.

Figures 7–8 present the group-level analyses for signal effects, encompassing age and group differences (ASD/HC). The ICA-DP method amplified the signal effects by identifying a greater number of regions with significant signal differences. In contrast, both ComBat and ICA-SP diminished the signal effects, leading to fewer regions or regions with diminished significance.

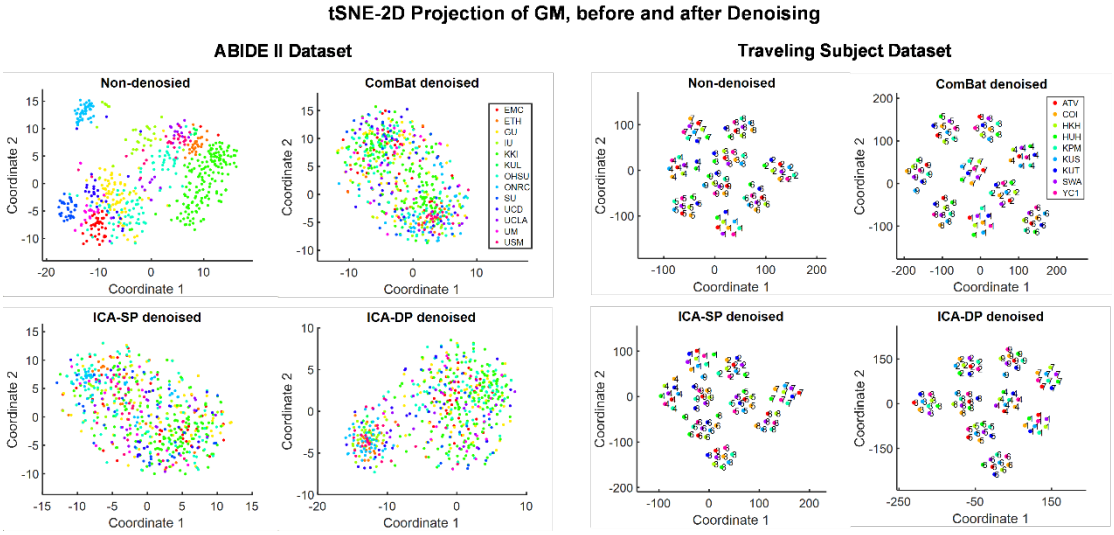


FIGURE 5 Dimension reduction visualization by t-SNE before and after denoising for ABIDE II and traveling subject datasets. The site-cluster distribution of ABIDE II dataset before denoising indicates the site effects, and it decreases when the data points are randomly distributed after denoising. For the traveling subject dataset, the subject-cluster distribution indicated the dominance of subject heterogeneity, as the subjects from this dataset are the same ones scanned at different centers (subject numbers labeled the data points). No significant difference existed before and after denoising, and subject heterogeneity was well preserved.

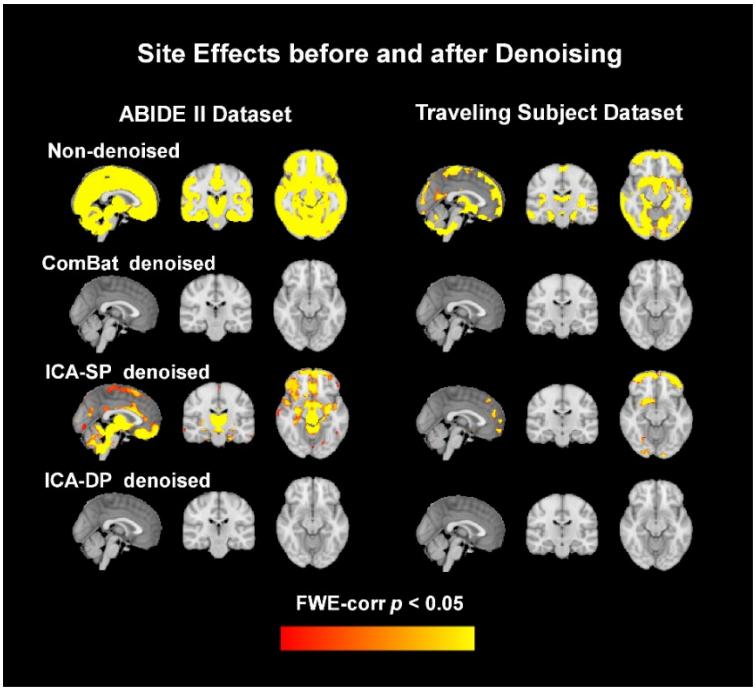


FIGURE 6 The group-level evaluation concerning site effects before and after noise reduction. Both ComBat and ICA-DP completely eliminated site effects. However, while ICA-SP managed to decrease these effects, noticeable regions impacted by site effects remain discernible.

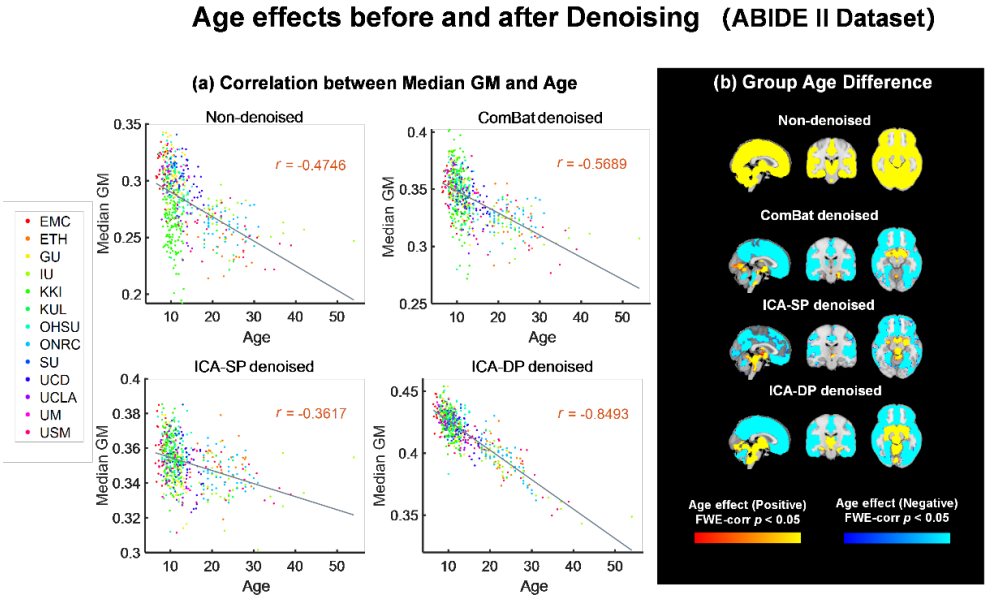


FIGURE 7 Age effects before and after site-effect removal. (a) The relationship between age and median GM value before and after site-effect removal. The Pearson correlation values were as follows: -0.4746 (Non-denoised), -0.5689 (ComBat denoised), -0.3617 (ICA-SP denoised) and -0.8493 (ICA-DP denoised). (b) A group-level analysis of GM maps regarding age effects before and after data denoising. The negative correlations with age become more pronounced after the denoising process, even though such effects were not apparent in the original, non-denoised data.

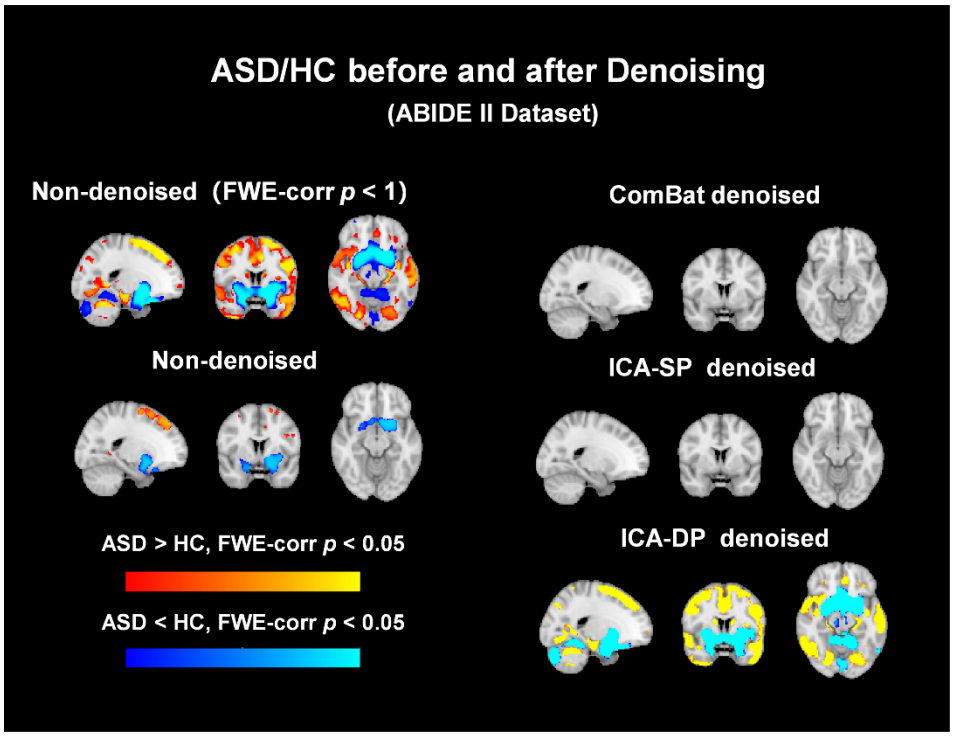


FIGURE 8 Group-level analysis of GM maps for group difference (ASD/HC) before and after data denoising. No significant regions were found from the data denoised by ComBat and ICA-SP, while ICA-DP could increase the significance of the regions related to ASD/HC.

### Research contributions

Our study introduced the ICA-DP method to address the issue that traditional ICA methods cannot completely remove site effects. This method thoroughly eliminated site effects and retained the biological signals of interest. Simulated, vivo structural MRI data from Autism Brain Imaging Data Exchange II and a traveling subject dataset from the Strategic Research Program for Brain Sciences, were used to test the ICA-DP methods.

### Authors' contributions

Huashuai Xu and Yuxing Hao contributed equally to this study, proposing the ideas of the whole study, analyzing the data, and writing and revising the manuscript. Yunge Zhang and Donagyue Zhou downloaded the data and preprocessed them. Chenwei Yan helped me modify some pictures in this article. Mingrui Xia, Tommi Kärkkäinen, Lisa D. Nickerson, Huanjie Li, and Fengyu Cong supervised the whole study and revised the manuscript.

## 3.2 Article II: Harmonization of multi-site functional MRI data with dual-projection based ICA model

Xu, Huashuai, Yuxing Hao, Yunge Zhang, Dongyue Zhou, Tommi Kärkkäinen, Lisa D. Nickerson, Huanjie Li, and Fengyu Cong (2023). Harmonization of multi-site functional MRI data with dual-projection based ICA model. Frontiers in Neuroscience, 17, 1225606, https://doi.org/10.3389/fnins.2023.1225606

### Methods

This research delves into the application of our ICA-DP denoising technique in harmonizing fMRI data sourced from the ABIDE II. Through frequency-domain and regional homogeneity examinations, two modalities, namely amplitude of low frequency fluctuation (ALFF) and regional homogeneity (ReHo), are employed to benchmark our method against two well-regarded harmonization techniques: ICA and ComBat.

In determining the efficacy of these harmonization techniques, we adopt an array of visualization and quantification methods to analyze site effects both pre and post-denoising. Moreover, we gauge each denoising method's capability to retain signal effects.

### Results

Figures 9 and 10 show the site effects before and after harmonization. Both non-denoised modalities were universally impacted by site effects. While attempting to manage these site effects, the ICA-SP method only achieved limited success, leaving residual site-related biases in the data. Following denoising with ICA-DP and ComBat, brain regions exhibited site-related differences after their application, showcasing their robustness.

Figures 11-14 display the group-level analyses for signal effects, including age, sex, and group differences(ASD/HC). ICA-DP preserved and even increased the signal effects by detecting more significantly different regions related to signals, while ComBat and ICA-SP decreased the signal effects with fewer or less significant regions, suggesting a potential suppression of real signal effects.

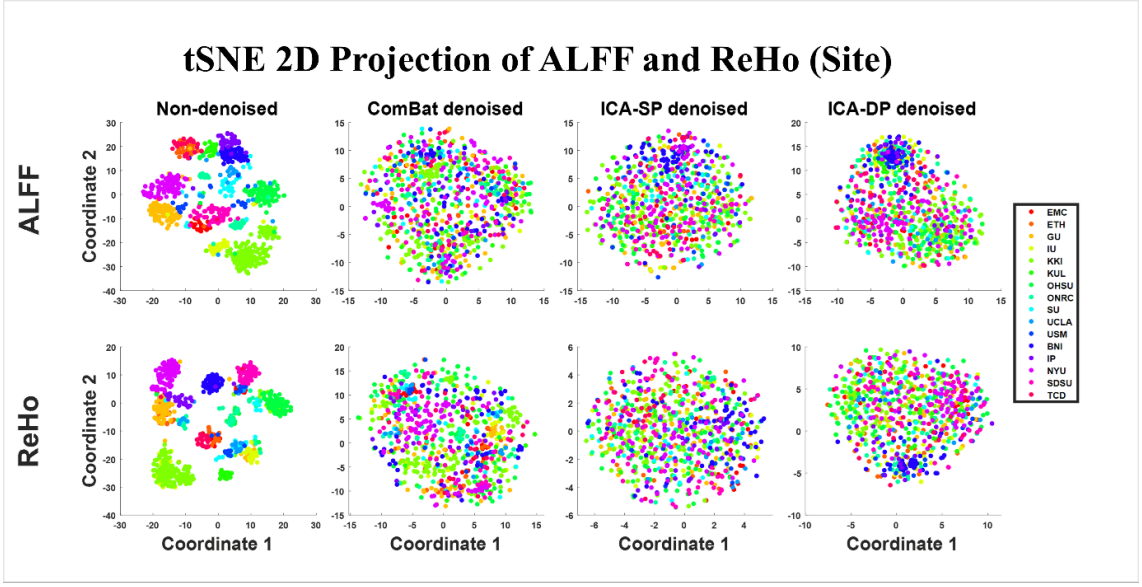


FIGURE 9 Visualization of dimension reduction using t-SNE prior to and following site effects removal (Sites). The distribution of sites grouped together before denoising showcased the site effects. This effect diminished as the data points became more scattered after denoising.

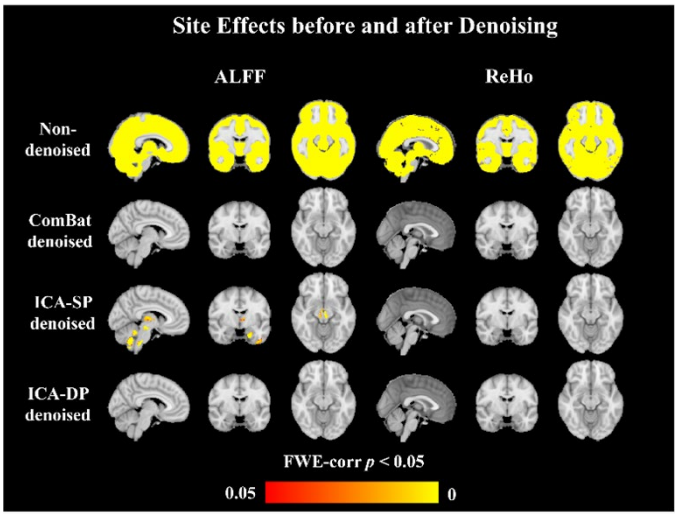


FIGURE 10 Group-level analysis for site effects before and after denoising. The site impacts were fully mitigated by ComBat and ICA-DP. While ICA-SP lessened these effects, certain notable areas remained evident.

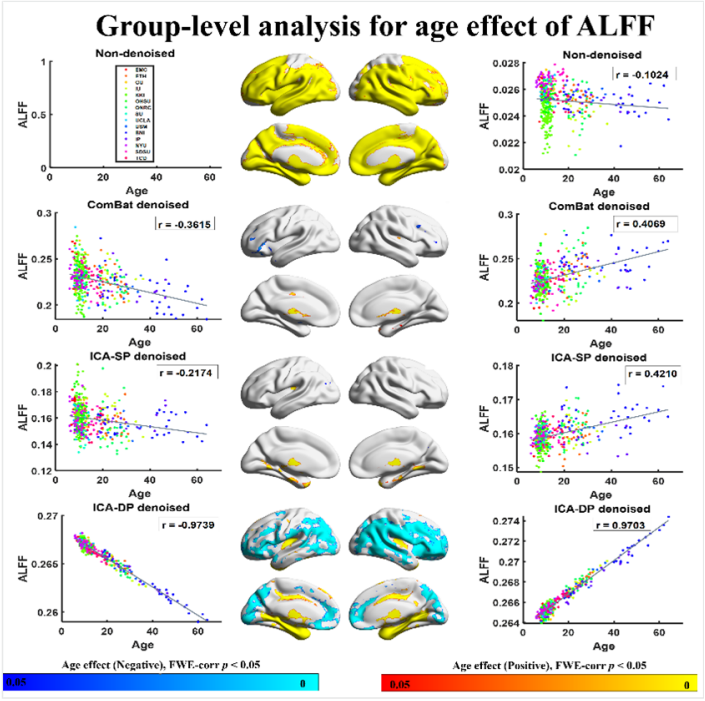


FIGURE 11 Associations between age and ALFF with different denoising strategies. "Positive" association indicates increasing amplitude with increasing age, whereas "Negative" refers to decreasing amplitude with increasing age. Associations with age are enhanced by ICA-DP and weakened by ICA-SP and ComBat.

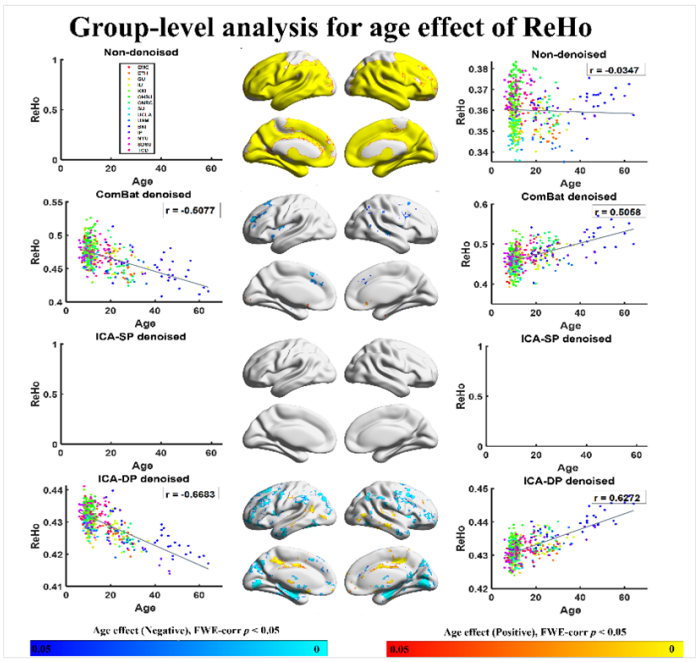


FIGURE 12 Associations between age and ReHo with different denoising strategies. "Positive" refers to significantly increasing amplitude with increasing age, whereas "Negative" refers to significantly increasing amplitude with decreasing age. The age effects are enhanced by ICA-DP, while weakened by ICA-SP and ComBat.

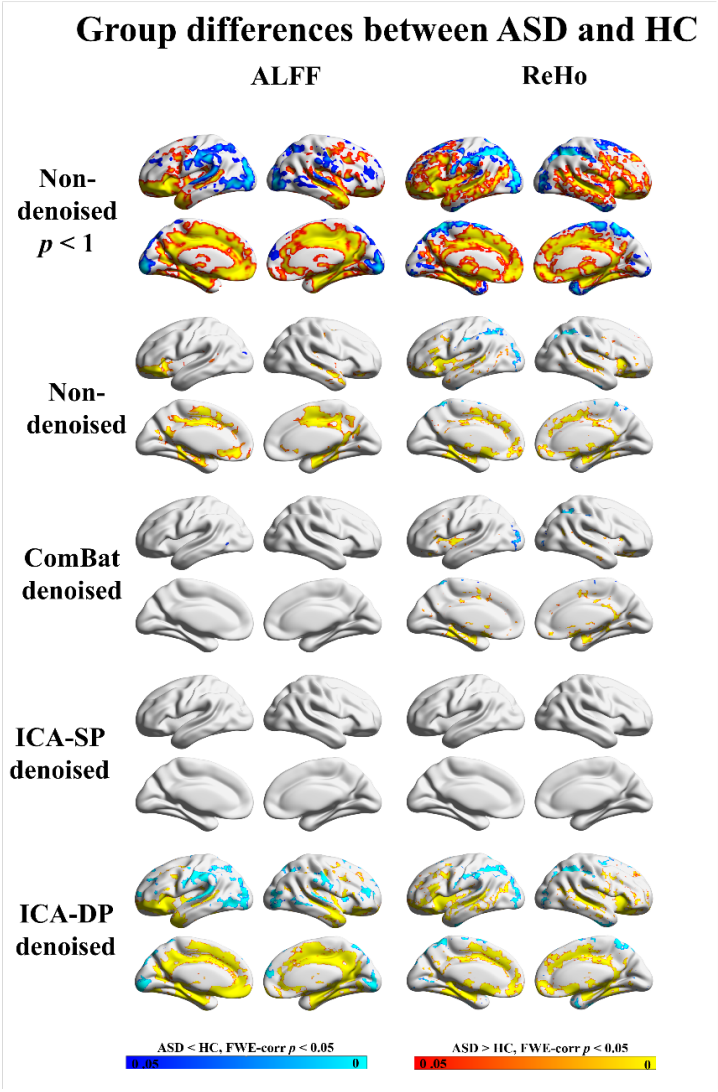


FIGURE 13 Group differences between ASD and HC before and after denoising. From the data denoised with ICA-SP, no notable regional differences are observed. Data denoised with ComBat revealed reduced regions, whereas ICA-DP enhanced the prominence of regions associated with ASD/HC. A FWE-corr $p$-value of less than 1 for non-cleaned data suggested that the tested regions from ICA-DP denoised data were not reintroduced artifacts.

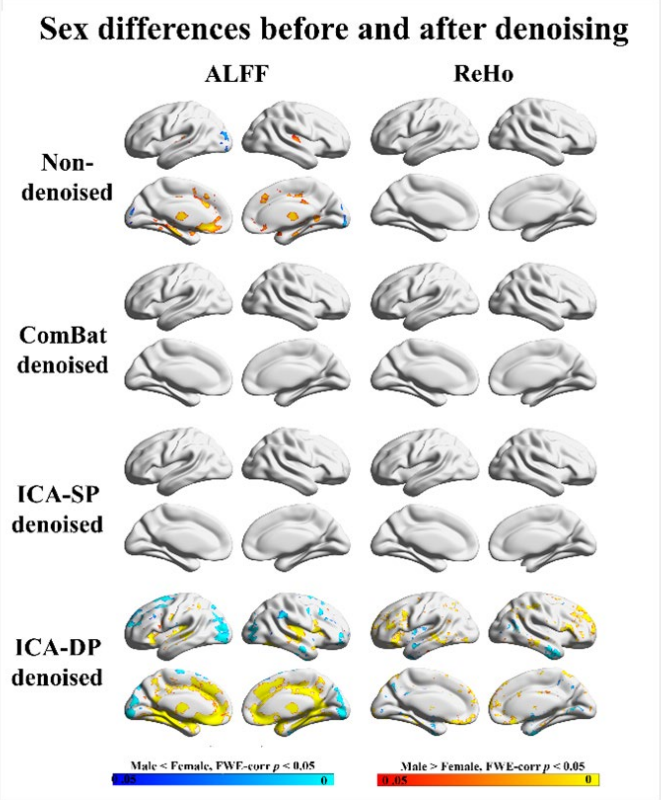


FIGURE 14 Sex differences before and after denoising. "Male < Female" refers to significantly greater amplitude in females, whereas "Male > Female" refers to significantly greater amplitude in males. ICA-DP amplified the effects related to sex, whereas ICA-SP and ComBat diminished them.

### Research contributions

This study extended the ICA-DP method to functional magnetic resonance data, further validating its effectiveness in both eliminating site effects and preserving biological signals.

### Authors' contributions

Huashuai Xu proposed the ideas of the whole study, analyzed the data, and wrote and revised the manuscript. Yunge Zhang and Donagyue Zhou downloaded the data and preprocessed them. Yuxing Hao contributed to the guidance of methods. Tommi Kärkkäinen, Lisa D. Nickerson, Huanjie Li, and Fengyu Cong supervised the whole study and revised the manuscript.

## 3.3 Article III: Enhancing performance of linked independent component analysis: Investigating the influence of subjects and modalities

Huashuai Xu, Tommi Kärkkäinen, Huanjie Li, and Fengyu Cong (2023). Enhancing performance of linked independent component analysis: investigating the influence of subjects and modalities. In 2023 International Conference on Computers, Information Processing and Advanced Education (CIPAE), pp. 726-732. IEEE.

**Methods**

In this research, we assess the influence of the number of subjects and modalities on the performance of LICA, employing both simulated multimodal MRI data and real multimodal MRI datasets from ABIDE II. The use of simulated data enabled us to evaluate the impact of variations in the number of subjects and modalities. Real multi-site MRI data were deployed to underscore the benefits of multimodal fusion in identifying site-related components and mitigating site effects.

To evaluate the impact of the number of subjects on the LICA results, we opted for three distinct component quantities: 10, 20, and 30. For each chosen component number (with all simulated components being spatially independent), the number of subjects ranged from 40 to 200 (inclusive of 40, 50, 60, 70, 80, 90, 100, 150, 200). The effectiveness of LICA was determined by calculating the correlation between the spatial maps and subject courses generated by LICA and those implemented in the simulation.

We explored the impact of the number of modalities on LICA results from two perspectives of view: 1) We repeatedly input the aforementioned simulated data (considered as Pseudo multimodal data) as different modalities into LICA to ascertain the influence of modality count; 2) We initially established an interesting multimodal component by defining a signal variable related to one component in each modality at varying levels (Table 8), and considered the remaining nine components as non-interest components. This multimodal component, derived from LICA, was employed to examine the effects of the number of subjects and modalities on LICA. The number of subjects used in this context was 100.

In this study, to match the number of subjects in each site, we employed ALFF, fALFF, and ReHo to assess the performance of LICA in the context of multimodal data fusion.

TABLE 8 Correlation between signal and subject courses and correlation among subject courses from different modalities. There is one component related to the signal variable in each modality, and the correlation coefficient ranges from 0.4 to 0.9, with 0.1 intervals. The corresponding correlation coefficients among subject courses from different modalities can be seen in the right column.

| Correlation between signal and subject courses | Correlation among subject courses from different modalities |
|---|---|
| 0.4 | 0.16 |
| 0.5 | 0.26 |
| 0.6 | 0.36 |
| 0.7 | 0.49 |
| 0.8 | 0.64 |
| 0.9 | 0.81 |

**Results**

Our simulation findings revealed that enhancing the number of modalities and subjects can improve outcomes when LICA fails to accurately recover spatial maps or subject courses. The correlation among subject courses from diverse modalities, the number of modalities, and the selection of components for decomposition all influence LICA's linking performance. Furthermore, our results derived from real-world datasets illustrated the benefits of multimodal fusion via LICA, which includes

- the identification of an increased number of site-related components and
- the removal of a greater amount of site effects.

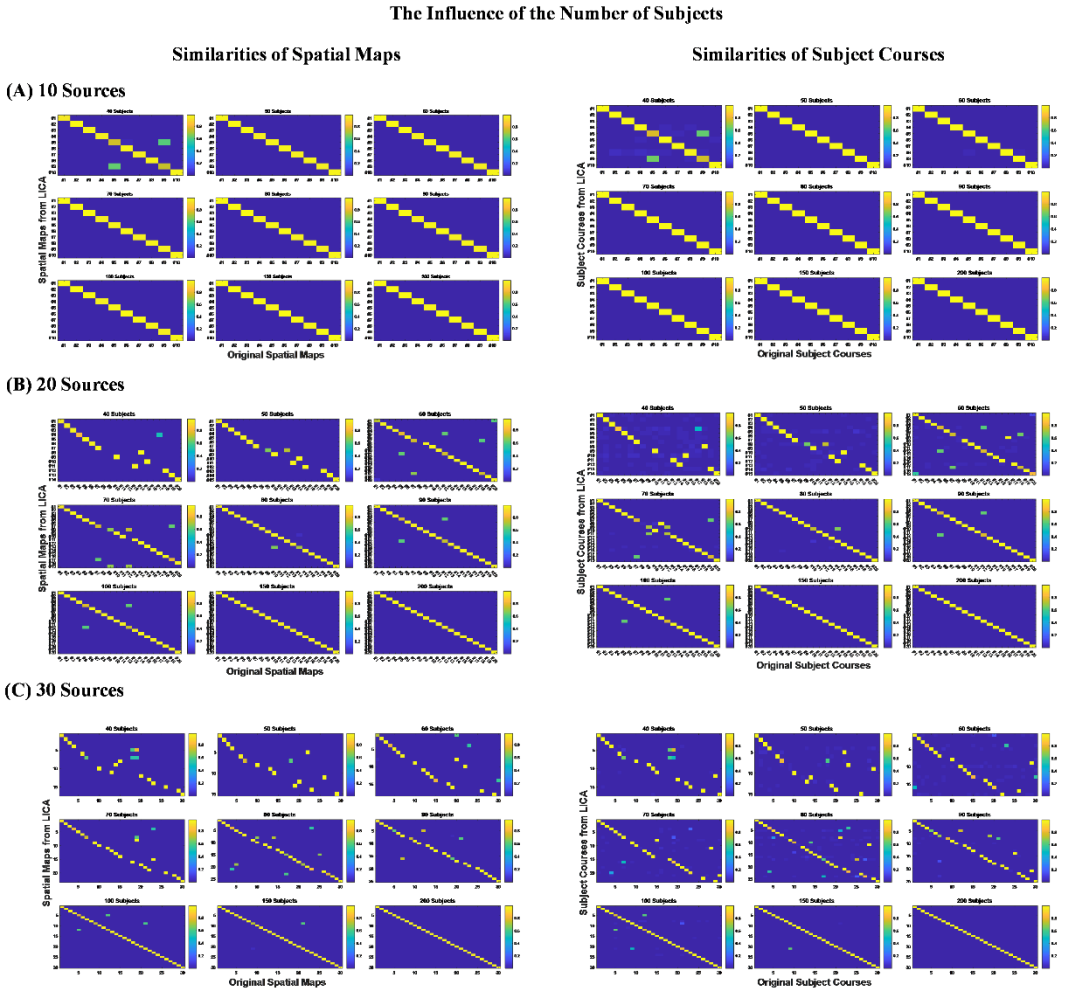


FIGURE 15 The influence of the number of subjects on recovering spatial maps and subject courses of LICA. (A) Given 10 independent components, LICA can successfully recover all components from the data with over 40 subjects. However, there are two correlated components that ought to be independent when the subject count is exactly 40; (B) In the presence of 20 independent components, LICA can extract all components from the data with more than 80 subjects and achieve full component independence when the subject count exceeds 150; (C) For scenarios with 30 independent components, LICA can retrieve all components from the data with over 100 subjects and ensure complete component independence when the number of subjects is greater than 200.

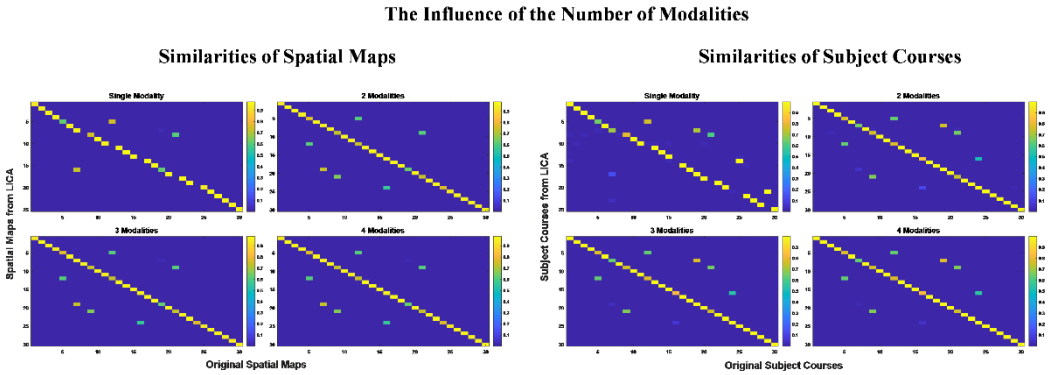


FIGURE 16 The influence of the number of modalities on recovering spatial maps and subject loadings when the number of sources is 30, and the number of subjects is 90. LICA can retrieve a larger number of simulated multimodal components when multiple modalities are used compared to a single modality.

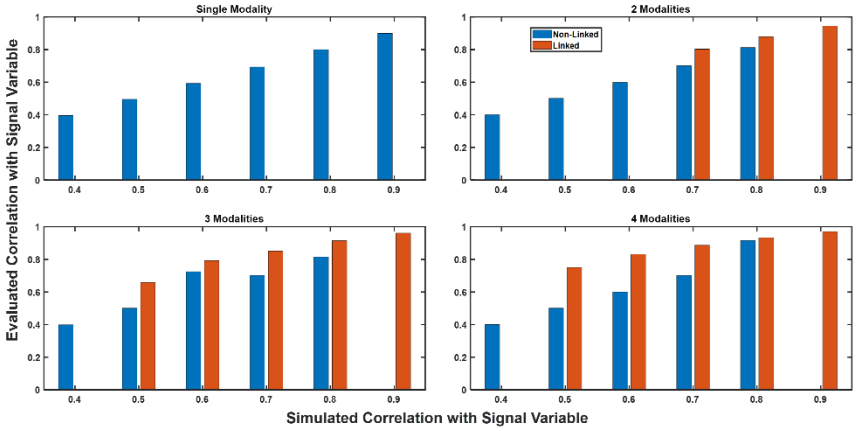


FIGURE 17 The influence of the number of modalities on the linking performance. The signal-related components from multiple modalities can not be linked when they are weakly correlated (less than 0.16) and can start to be linked with the increase of the correlation (larger than 0.25). As the number of modalities increases, LICA begins to link the related components earlier, and the correlation between the corresponding components and signal variable also becomes stronger.

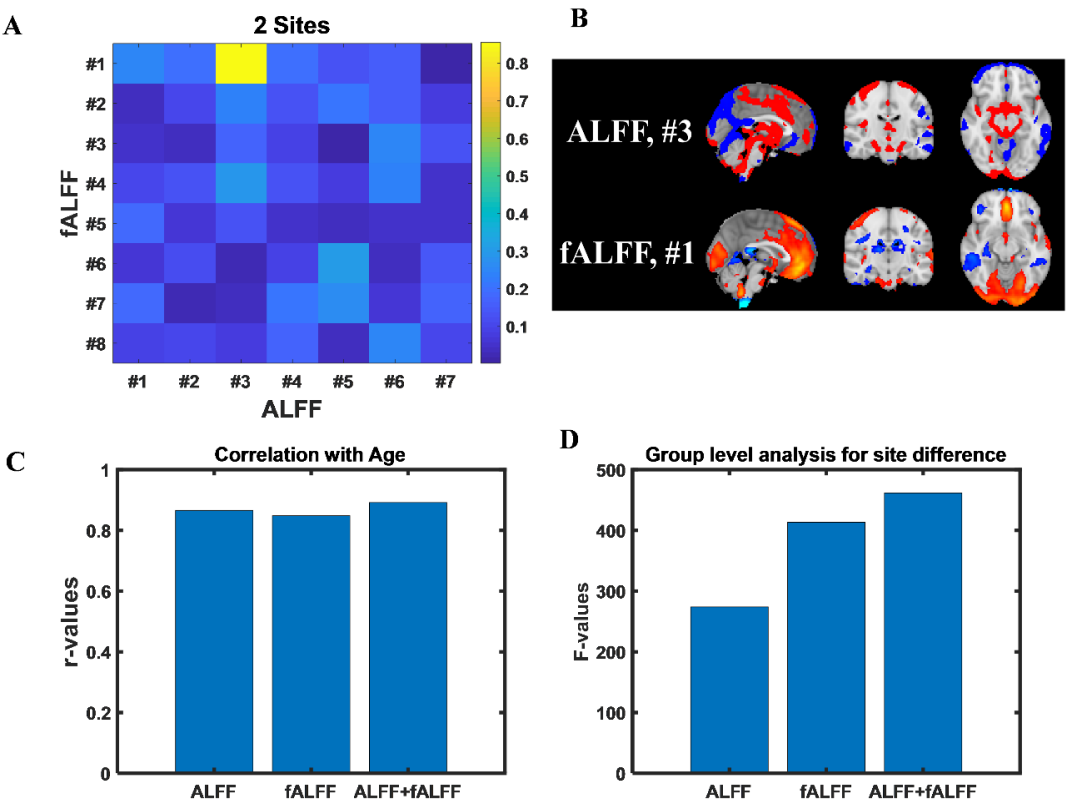


FIGURE 18 Components from LICA based on a single modality (ALFF, fALFF) and two modalities. (A) The correlation of subject courses among all the components generated by LICA on ALFF and fALFF; (B) The spatial maps of the significantly related components from ALFF and fALFF; (C) The correlation with age from a single modality and two modalities; (D) The correlation with site difference from a single modality and two modalities.

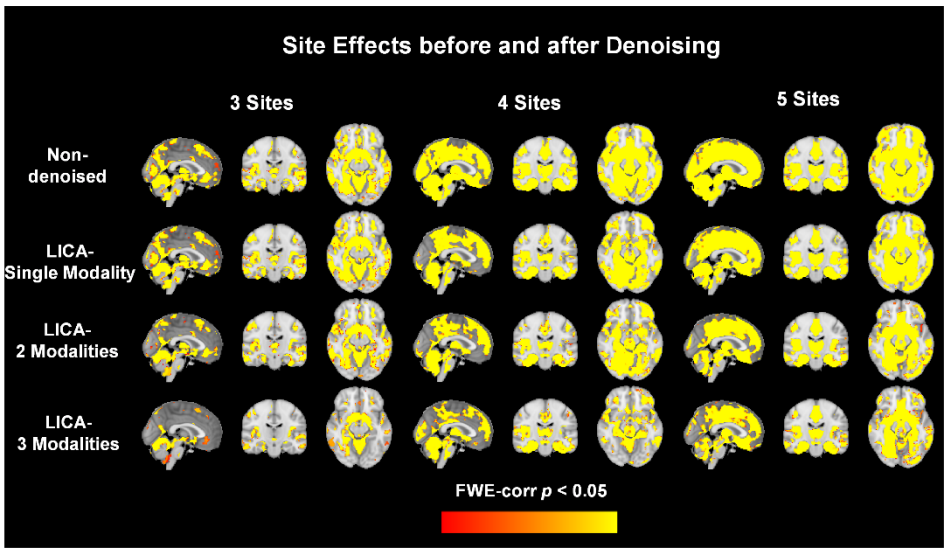


FIGURE 19 Group-level analysis for site effects before and after denoising. As the number of sites increases, more significantly different regions are related to sites. LICA based on multiple modalities can remove more site effects than LICA based on a single modality.

**Research contributions**

This study analyzed the impact of the number of modalities and the number of subjects on the results of the LICA method, and explored how to achieve optimal performance with LICA.

**Authors' contributions**

Huashuai Xu proposed the ideas of the whole study, analyzed the data, and wrote and revised the manuscript. Tommi Kärkkäinen, Huanjie Li, and Fengyu Cong supervised the whole study and revised the manuscript.

## 3.4 Article IV: Harmonization of multi-site MRI data with dual-projection based Linked ICA model

Huashuai Xu, Yuxing Hao, Yunge Zhang, Dongyue Zhou, Tommi Kärkkäinen, Lisa D. Nickerson, Huanjie Li, and Fengyu Cong. Harmonization of multi-site MRI data with dual-projection based Linked ICA model. To be submitted, 2023

**Methods**

In the study, we introduce an innovative multimodal harmonization strategy, employing a Linked Independent Component Analysis (LICA)-based Dual Projection (DP) methodology, crafted meticulously to mitigate the site effects. This technique possesses the capability to segregate the signal effects from the discerned site effects discretely. For the empirical validation of the proposed LICA-DP denoising methodology, we utilized a dataset derived from the Autism Brain Imaging Data Exchange II.

In order to gauge the efficacy of the harmonization methodologies, we employ a multitude of techniques, both for the visualization and quantification of site effects before and after the harmonization process. Moreover, we appraise the harmonization methods concerning their proficiency in maintaining the integrity of signal effects.

**Results**

For the structural MRI data, we utilized data from the first 13 sites within the ABIDE database and 9 sites in the traveling-subject dataset. Figures 20-23 show the results from unimodal MRI data, including the site and signal (Age, sex, group differences(ASD and HC)) effects before and after harmonization. LICA-DP is either equivalent to or surpasses traditional ICA methods in mitigating site-specific effects while simultaneously retaining the integrity of the biological signal.

Figures 24-27 show the results from unimodal MRI data, including the site and signal (Age, sex, group differences(ASD and HC))effects before and after harmonization. When data from the ALFF and ReHo modalities are fused, both LICA-SP and LICA-DP can achieve superior (or at least equivalent) outcomes

regarding removing site effects. In the statistical testing of signals, the results do not exhibit significant discrepancies compared to those obtained from unimodal analyses.

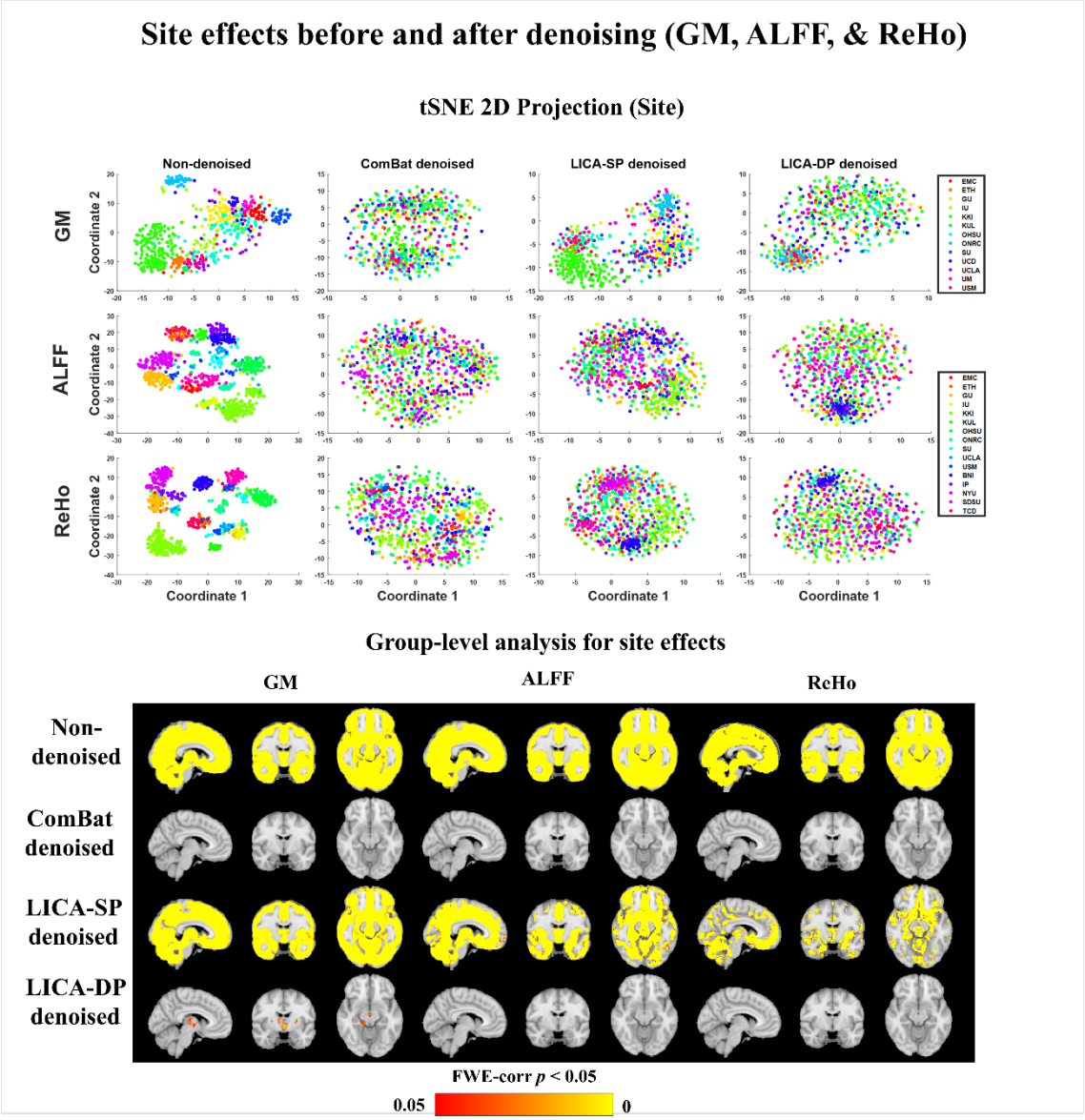


FIGURE 20 Site effects before and after harmonization. For all three modalities, data points tended to cluster by site, indicating the objective existence of site-specific effects. After harmonization, the distribution of data points became random and no longer tended to cluster by site. From the group-level analysis results, the impact of site effects was global across the brain. ComBat could eliminate site effects, while LICA-SP could partially remove them. Regarding ALFF and ReHo metrics, LICA-DP was proficient in eliminating site effects.

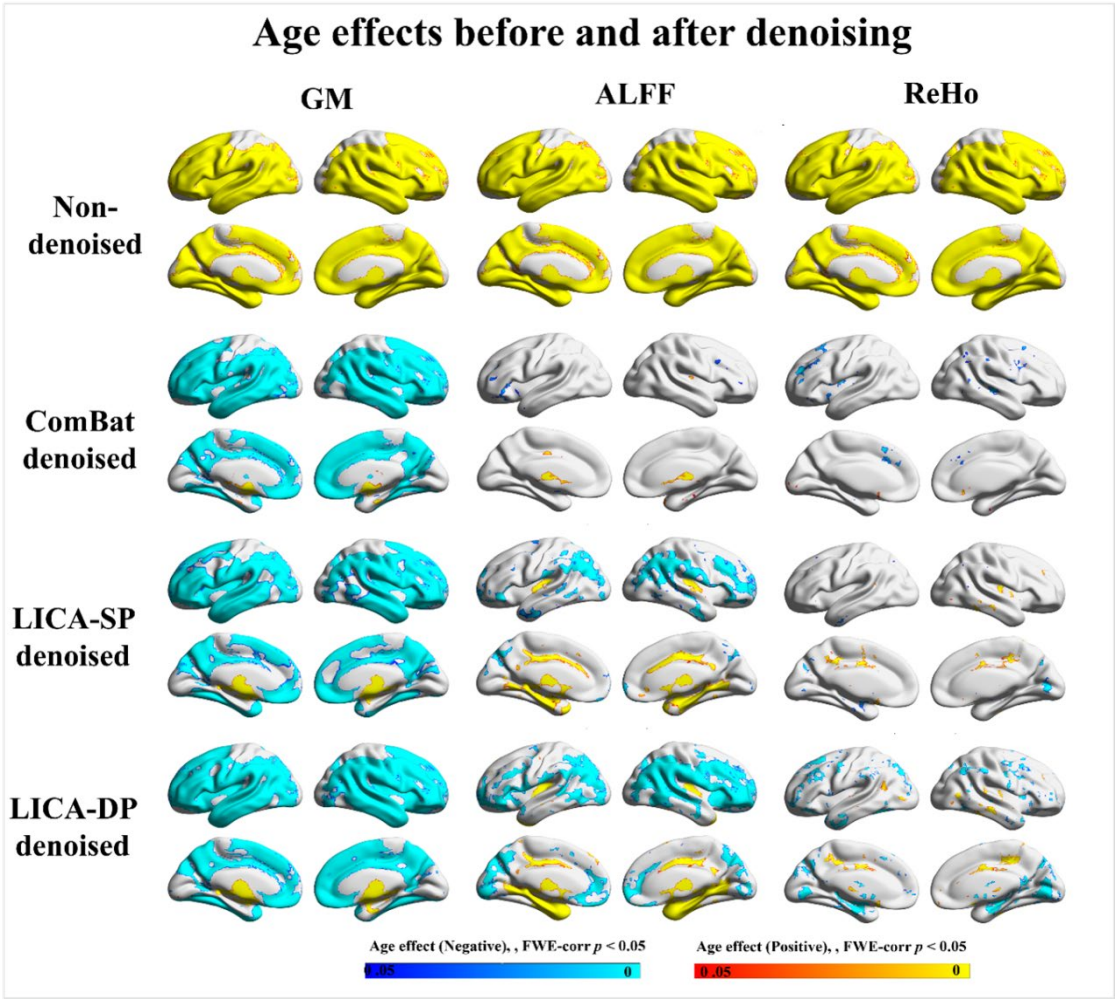


FIGURE 21 Age effects before and after harmonization. "Positive" association indicates increasing amplitude with increasing age, whereas "Negative" refers to decreasing amplitude with increasing age. Associations with age are enhanced by LICA-DP and weakened by ComBat.

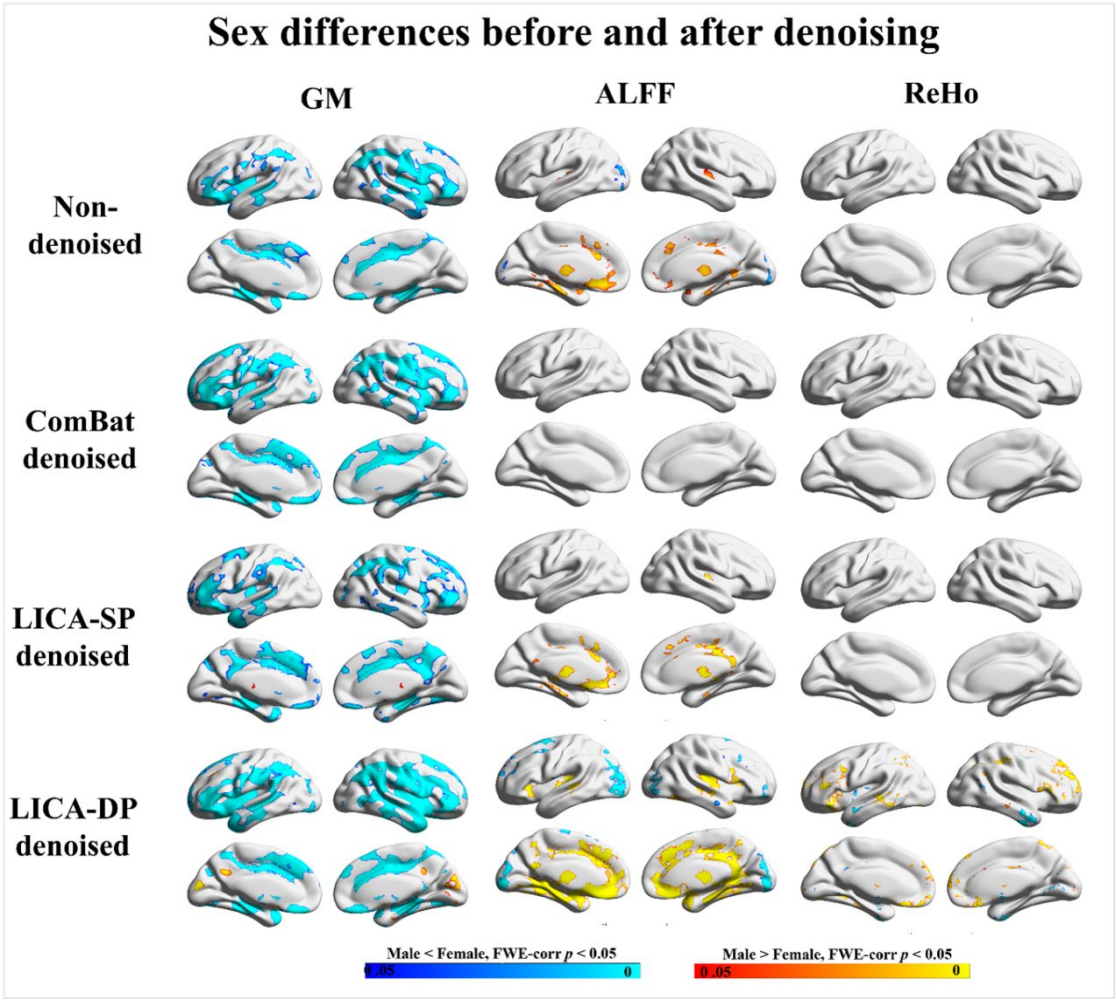


FIGURE 22 Sex effects before and after harmonization. "Male < Female" refers to significantly greater amplitude in females, whereas "Male > Female" refers to significantly greater amplitude in males. The sex effects are enhanced by LICA-DP while weakened by LICA-SP and ComBat.

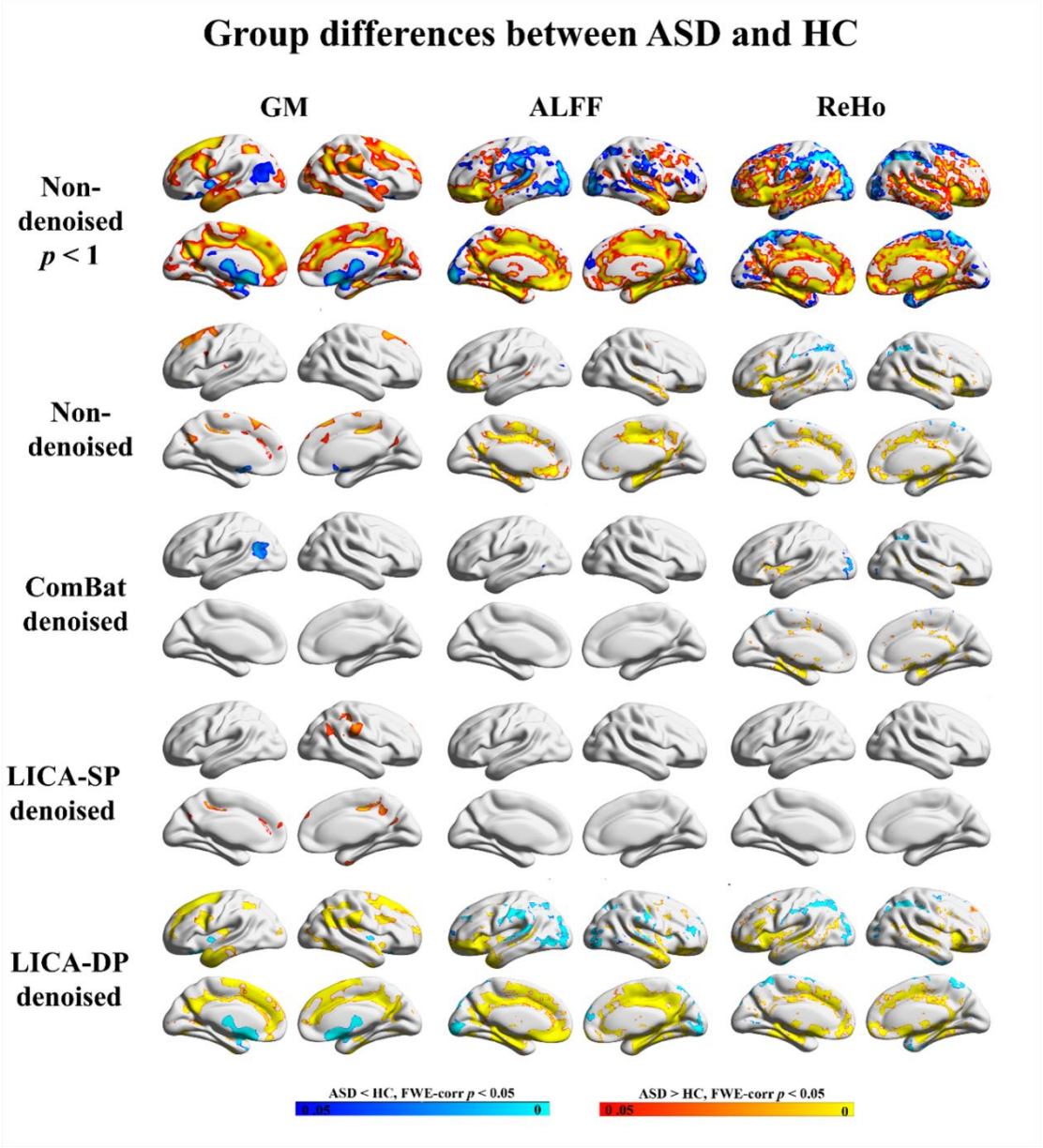


FIGURE 23 Group differences (ASD/HC) before and after harmonization. Fewer regions were found from the data denoised by ComBat and LICA-SP, while LICA-DP could increase the significance of the regions related to ASD/HC.

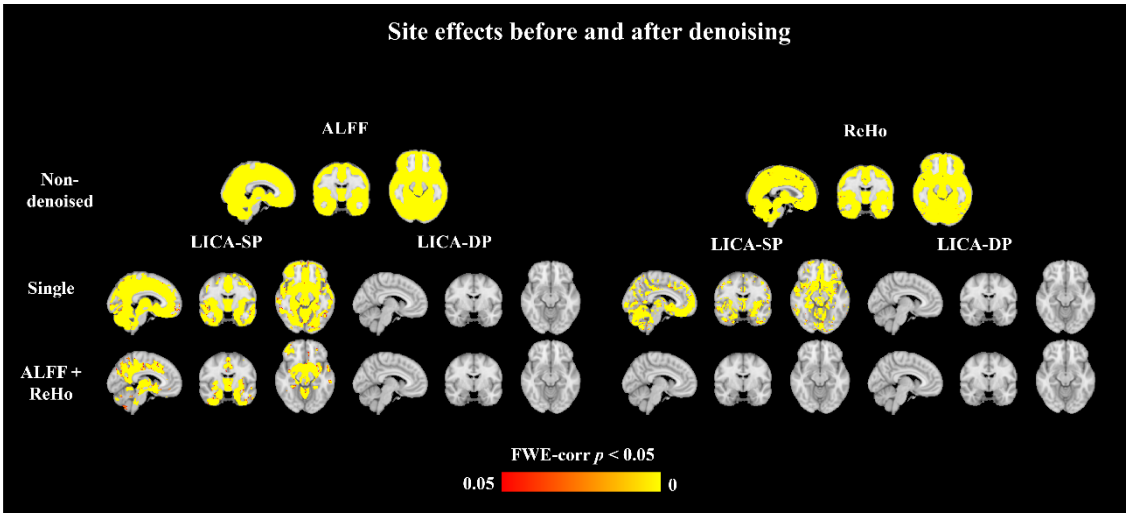


FIGURE 24 Site effects before and after harmonization (multimodal). Combining ALFF and ReHo modalities can eliminate site effects more effectively than unimodal. However, the mitigation of site effects is worse when ALFF, ReHo, and GM modalities are employed concurrently.

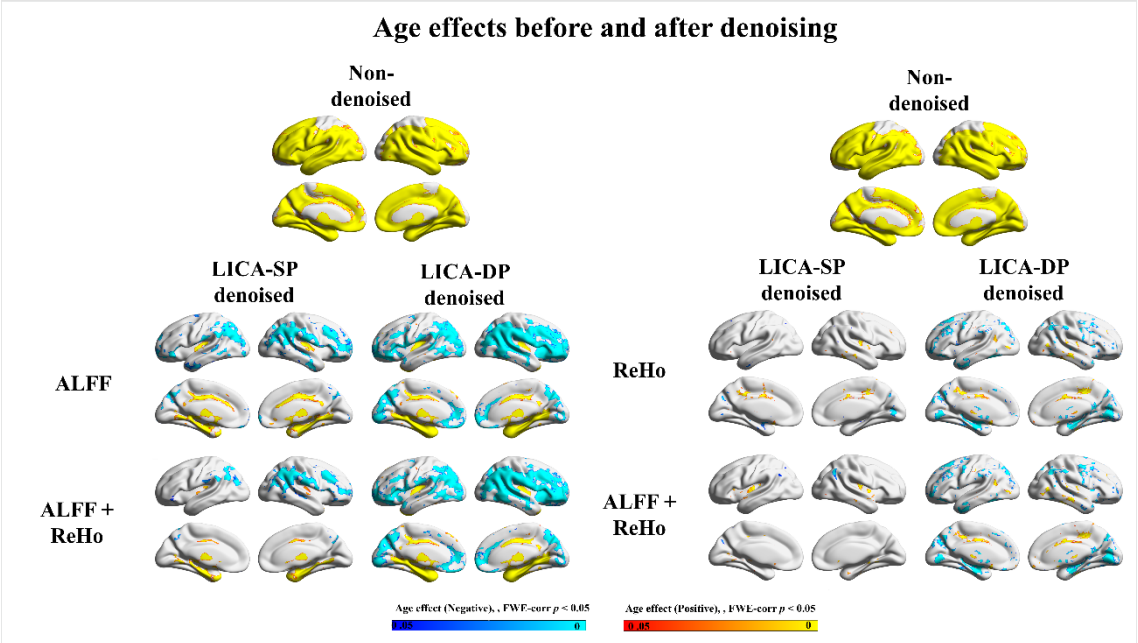


FIGURE 25 Age effects before and after harmonization (multimodal). "Positive" association indicates increasing amplitude with increasing age, whereas "Negative" refers to decreasing amplitude with increasing age.

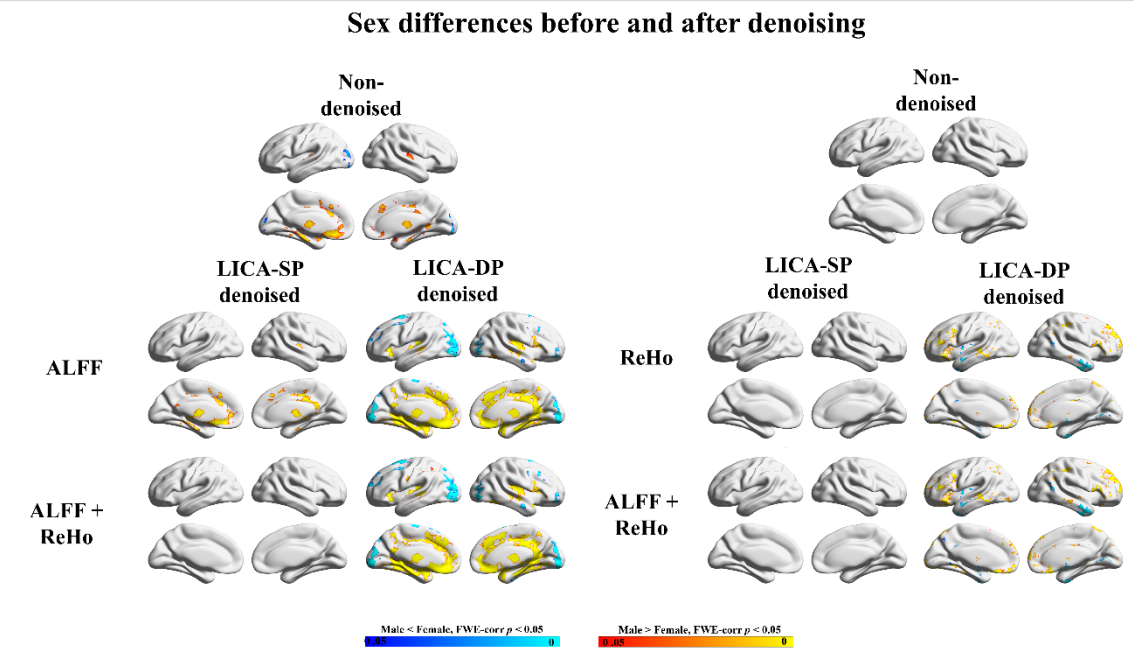


FIGURE 26 Sex effects before and after harmonization (multimodal). "Male < Female" refers to significantly greater amplitude in females, whereas "Male > Female" refers to significantly greater amplitude in males.

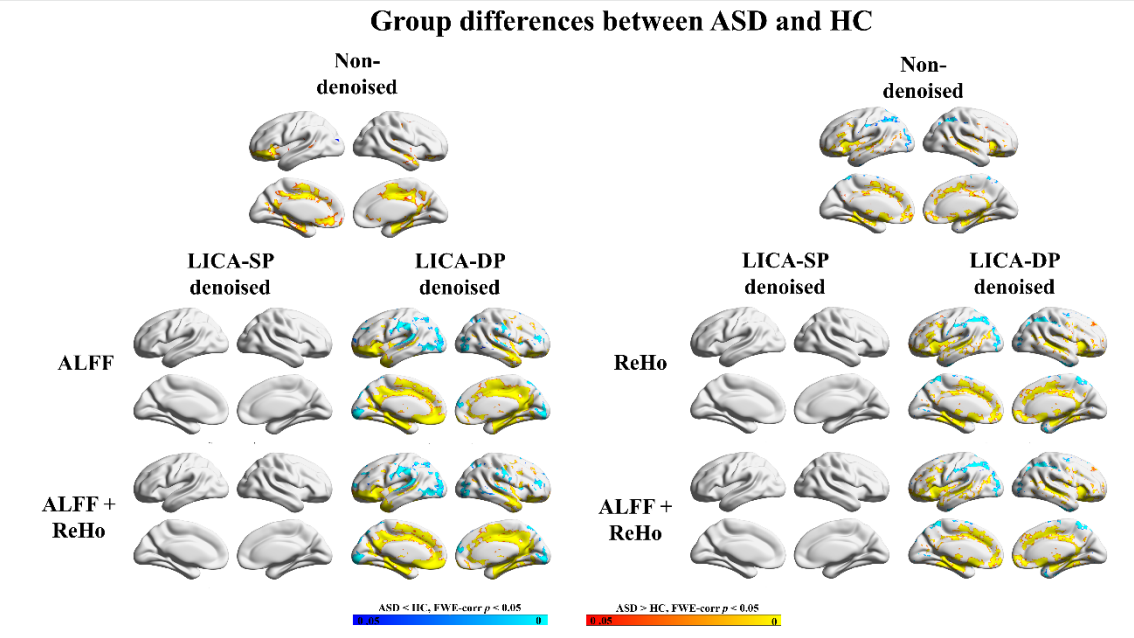


FIGURE 27 Group differences (ASD/HC) before and after harmonization (multimodal).

**Research contributions**

This study combined the DP and LICA methods to introduce the LICA-DP approach, which is designed to remove site effects from multi-modal magnetic resonance data.

**Authors' contributions**

Huashuai Xu proposed the ideas of the whole study, analyzed the data, and wrote and revised the manuscript. Yunge Zhang and Donagyue Zhou downloaded the data and preprocessed them. Yuxing Hao contributed to the guidance of methods. Tommi Kärkkäinen, Lisa D. Nickerson, Huanjie Li, and Fengyu Cong supervised the whole study and revised the manuscript.

# 4 DISCUSSION

This thesis investigated the harmonization of multi-site MRI data. This chapter will discuss our findings on MRI data harmonization methods and some limitations in the view of methodology. We will also point out some future directions of current research.

## 4.1 Findings of multi-site MRI data harmonization methods

We begin by discussing the performance of the ICA-DP method in the multi-site harmonization of unimodal MRI data. Then, we discuss the LICA-DP method's performance in harmonizing both single and multimodal data across multiple sites.

### 4.1.1 ICA-DP

We proposed the ICA-DP denoising method, which utilizes dual-projection in independent component analysis. This method adeptly and thoroughly eradicates site-related biases, simultaneously amplifying signal-related details. The performance of ICA-DP is particularly notable when the site effects are correlated with signal variables. ICA-DP could effectively eliminate site effects while enhancing signal-related information, regardless of the correlation between site and signal variables. Indeed, our method's enhancement of biological signals stems from our approach to dealing with noise components. We first regress out the influence of biological signals for each identified noise component prior to its application for denoising. As a result, the fraction of physiological signals in the denoised data becomes relatively larger, streamlining the identification of brain areas genuinely linked to the desired signals via statistical tests.

Based on the results from the simulated data, ICA-SP faces challenges in fully negating site interferences when the site-related components are

significantly correlated with signal variables. While the GLM-based denoising technique is proficient and parallels ICA-DP when the signal and noise variables lack significant correlation, GLM displays excessive purification, often sidelining signal-relevant details when these variables are highly correlated. ComBat surpasses GLM and the traditional ICA method, adeptly eradicating site effects and retaining or amplifying some signal effects when there is no correlation between signal and site variables. However, ComBat falters in wholly negating site effects when these variables overlap, especially when mixed components correlate more with the signal.

Analyzing results from the ABIDE II and traveling-subject datasets, ICA-DP outperformed both ICA-SP and ComBat in denoising site interferences. Significant areas impacted by site effects, which ICA-SP failed to negate fully, became undetectable post-ICA-DP and ComBat. Variations between sites in the traveling data and internal subject fluctuations across the nine sites in the traveling-subject set markedly diminished after denoising with ICA-DP. Moreover, subject heterogeneity within the traveling-subject dataset was well preserved after denoising with ICA-DP.

In addition, ICA-DP excelled in enhancing biological variability, such as age, sex effects, and group differences between individuals with ASD and healthy controls (HC), compared to ICA-SP and ComBat. 1) for structural MRI data, site effects hinder the detection of true age and group effects. However, after denoising with ICA-SP, ICA-DP, and ComBat, the true age effects on GM are discovered. Of these, ICA-DP identifies the most age-linked regions, strongly enhancing the negative correlation between age and GM. The median GM and age relationship improves from -0.4746 (non-denoised) to -0.8493 after denoising with ICA-DP (Figure 7). Moreover, ICA-DP enhances group differences (ASD/HC) (Figure 8). 2) for ALFF and ReHo, ICA-DP enhances the assessment of biological signals, including the effects of age, sex, and group difference (ASD/HC) (Figures 11-14).

These enhancements in biological variability may be attributed to the larger proportion of site-related components chosen for denoising with ICA-DP, which increases the weights of the signal of interest, making detection more straightforward. Additionally, other variables of interest are effectively preserved by incorporating them into the first projection of the ICA-DP denoising method (Eq. (10)). As a result, ICA-DP stands out as the premier technique for negating site effects while retaining biological variations among the discussed methods.

Moreover, unlike ICA-SP, ICA-DP's efficacy in denoising site effects and enhancing signals remains unswayed by the number of ICA decomposition components chosen. It consistently eradicates site biases and amplifies signals, irrespective of the component count chosen.

### 4.1.2 LICA-DP

LICA can be applied to single-modality data, uniquely differentiating it from the traditional ICA through an intrinsic characteristic: a definitive upper limit on the

number of components it can obtain. Intriguingly, LICA may not always achieve the anticipated number of components, introducing a layer of complexity in its application.

LICA skillfully intertwines dimensionality reduction within the standard ICA process, harnessing the prowess of components' Automatic Relevance Determination (ARD) priors, as (Roberts, 2001) expounded. This integration ensures that components deemed irrelevant—or part-components—are judiciously excised throughout the computational journey, thereby circumventing further analysis of zero-weight spatial maps, optimizing computational resources, and elevating the model's efficiency (Groves et al., 2011). In a comparative study involving our ICA-DP research (Hao et al., 2023; Xu, Hao, et al., 2023), where the number of components for ICA decomposition was set to 100, 150, and 200, LICA decomposition intriguingly yielded a scant 99 components for ALFF data, 83 for ReHo data, and a mere 80 for GM data.

Regarding denoising performance, LICA-DP mirrored ICA-DP in eradicating site effects from ALFF and ReHo data, adeptly obliterating site-specific effects. Conversely, its performance wavered with GM data, failing to fully eradicate site-specific effects, which appear intricately linked to the suboptimal number of components derived through LICA. A subsequent adjustment of mixed components for signal effects (Eq.12) revealed that these components persisted and correlated with site-specific variables post-denoising procedure (Eq.13). Thus, the inability to fully cleanse GM data of site-specific effects roots itself in the limited components derived through LICA, thereby impacting the denoising process. Given that LICA-SP does not address mixed components, it inherently cannot fully expunge site-specific effects. Conversely, the ComBat method consistently and successfully eliminates site-specific effects.

Results underscore that LICA-DP is adept at purging site-related effects while concurrently amplifying the detection of biologically relevant signals, including effects related to age, sex, and group differences (ASD/HC).

Despite LICA-SP's inability to fully extinguish site-specific effects, it outperforms in retaining signal effects compared to our previous ICA-SP results (Hao et al., 2023; Xu, Hao, et al., 2023). This is markedly observable in testing age effects on GM and ALFF data, where LICA-SP mirrors results nearly identical to those garnered using LICA-DP.

Originally conceived for the fusion analysis of multimodal data, the LICA method has paved the way for researchers to discern common features across diverse modalities. We have delved deeply into exploring modality quantity and its ensuing impact on LICA outcomes (Xu, Li, et al., 2023).

When data from ALFF and ReHo modalities are conjointly integrated, both LICA-SP and LICA-DP exhibit a capability to achieve, at a pinnacle, superior or at least equivalent outcomes in proficiently mitigating site effects. This is especially evident in ReHo, where even LICA-SP manages to obliterate the site effects entirely. This remarkable capability is attributed to the fusion of two modalities (ALFF and ReHo), leading to a more abundant component spectrum that robustly correlates with site effects. However, caution arises during the

statistical testing of signals: the utilization of a more significant component number to eliminate central effects concurrently diminishes signal variables (Age, Sex, and Group differences between ASD and HC). Contrarily, our LICA-DP method first regresses the signal contribution used in the denoising components, subsequently deploying it to eradicate site effects, thereby averting potential signal damage.

## 4.2 Limitations

It is essential to note that a limitation of our proposed harmonization method arises when the noise variable exhibits a strong relationship with the signal variable. In such cases, ICA-DP may struggle to eliminate the intersection effects of both site and signal variables. This limitation is intrinsic to the inherent correlation between the noise and signal variables. However, it is worth mentioning that in our simulations, high correlation values between noise and signal variables were deliberately chosen, and such situations are unlikely to occur frequently in real data studies.

Also, our DP-based methods focus on preserving the desired biological signals while potentially neglecting the preservation of certain non-target variables.

## 4.3 Future directions

Differences among sites pose a significant challenge when consolidating data from various scanners. Consequently, accurately identifying and removing scanner-related noise from MRI data is crucial for enhancing both the accuracy and reliability of data-sharing studies. Our futural research aims to identify stable, site-specific effects, thereby contributing to overcoming this barrier.

In addition, we plan to develop more accurate statistical methods to explore the specific contribution of different scanner parameters to specific site effects. This work is crucial for merging multi-site MRI data. It is fundamental to reveal the mechanism of site effects on MRI data and to develop the next-generation version of MRI data harmonization.

Our ongoing research is primarily dedicated to advancing the development of our methodology. Looking ahead, we are excited to direct our efforts toward practically implementing these methodologies in various clinical applications.

# 5 CONCLUSION

Integrating multi-site MRI data can improve the statistical power and the reliability and reproducibility of neuroimaging research. However, the presence of site effects complicates data analysis and makes it difficult to interpret the results. Removing these effects is crucial for successful multi-site data fusion. Additionally, preserving signals of interest is critical in any denoising strategy. Traditional ICA methods can not remove site effects inadequately or preserve the signal effects. To address these limitations, we propose a dual-projection data-driven method based on ICA that effectively eliminates noise while preserving the signal of interest.

ICA-DP method has effectively removed site effects while maintaining the integrity of biological variations. By harmonizing multi-site MRI data, our method bolsters the robustness and precision of analyses. This approach greatly uplifts the reliability of neuroimaging investigations, marking ICA-DP as a promising asset for upcoming research endeavors. A limitation of ICA-DP is that it cannot eliminate the intersection effects related to both site and signal variables when the site variable is strongly related to the signal variable (as illustrated in Figure 4); despite the harmonization, the refined data remain associated with site effects due to the intrinsic correlation between the site influences and signal elements.

Compared to unimodal, the advantage of multimodal fusion is that it can capitalize on the strength of each modality in a joint analysis. The estimation of site effects utilizing LICA on multimodal MRI data yields a more accurate model than those produced by single-modality ICA. This precision improvement reinforces the LICA method's effectiveness in accurately modeling site effects. So, we implement DP in LICA and denoise site effects from multi-modality data. The LICA-DP method underscores a significant advancement in eliminating site effects while preserving biological variability.

We emphatically advocate for researchers to adopt the ICA-DP and LICA-DP techniques to harmonize MRI data.

# YHTEENVETO (SUMMARY IN FINNISH)

Integroimalla usean paikan MRI-tietoja voidaan parantaa tilastollista tehoa sekä neurokuvantamistutkimusten luotettavuutta ja toistettavuutta. Kuitenkin paikkakohtaisten vaikutusten esiintyminen vaikeuttaa datan analysointia ja tulosten tulkintaa. Näiden vaikutusten poistaminen on välttämätöntä onnistuneelle usean paikan datan yhdistämiselle. On myös kriittistä säilyttää kiinnostuksen kohteena olevat signaalit missä tahansa kohinanpoistomenetelmässä. Perinteiset riippumattomien komponenttien analyysia (ICA) käyttävät menetelmät eivät kykene poistamaan paikkakohtaisia vaikutuksia riittävästi tai säilyttämään signaalivaikutuksia. Ratkaistaksemme nämä rajoitukset ehdotamme ICA:n pohjalta DP-ICA-kaksiprojektiomenetelmää.

ICA-DP-menetelmä poistaa tehokkaasti paikkakohtaiset vaikutukset säilyttäen biologisen vaihtelun. Yhdistämällä usean paikan MRI-tiedot, menetelmämme mahdollistaa robustimman ja tarkemman analyysin. Tämä lähestymistapa parantaa merkittävästi neurokuvantamistutkimusten validiteettia ja on lupaava työkalu tulevia tutkimuksia varten.

Väitöskirjassa esitetään myös uudenlainen monimuotoinen kohinanpoistomenetelmä paikkavaikutusten poistamiseksi, jossa kaksiprojektiomenetelmä (DP) yhdistetään linkitetyn riippumattomien komponenttien analyysin (LICA) kanssa. Tämän etuna on, että näin voidaan hyödyntää aineiston eri modaliteettien (esiintymismuotojen) vahvuuksia yhteisessä analyysissä. Paikkavaikutusten arvioiminen LICA:lla monimuotoisten MRI-tietojen pohjalta tuottaa tarkemman mallin kuin mitä saataisiin yksimuotoisen ICA:n avulla. Tämä tarkkuuden parantuminen korostaa LICA-menetelmän tehokkuutta paikkavaikutusten tarkassa mallintamisessa. LICA-DP-menetelmä on merkittävä edistysaskel paikkavaikutusten eliminointimenetelmissä, jotka samalla säilyttävät mittausaineiston biologisen vaihtelun.

Suosittelemme voimakkaasti, että tutkijat käyttävät ICA-DP- ja LICA-DP-menetelmiä MRI-datan kohinanpoistoon.

## REFERENCES


Beer, J. C., Tustison, N. J., Cook, P. A., Davatzikos, C., Sheline, Y. I., Shinohara, R. T., & Linn, K. A. (2020). Longitudinal ComBat: A method for harmonizing longitudinal multi-scanner imaging data. *NeuroImage*, *220*. https://doi.org/10.1016/j.neuroimage.2020.117129

Bell, T. K., Godfrey, K. J., Ware, A. L., Yeates, K. O., & Harris, A. D. (2022). Harmonization of multi-site MRS data with ComBat. *NeuroImage*, *257*, 119330. https://doi.org/10.1016/j.neuroimage.2022.119330

Bishop, C. M. (1999). Variational principal components. *IEE Conference Publication*, *1*(470), 509–514. https://doi.org/10.1049/cp:19991160

Biswal, B. B., Mennes, M., Zuo, X. N., Gohel, S., Kelly, C., Smith, S. M., Beckmann, C. F., Adelstein, J. S., Buckner, R. L., Colcombe, S., Dogonowski, A. M., Ernst, M., Fair, D., Hampson, M., Hoptman, M. J., Hyde, J. S., Kiviniemi, V. J., Kötter, R., Li, S. J., … Milham, M. P. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(10), 4734–4739. https://doi.org/10.1073/pnas.0911855107

Casey, B. J., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., Soules, M. E., Teslovich, T., Dellarco, D. V., Garavan, H., Orr, C. A., Wager, T. D., Banich, M. T., Speer, N. K., Sutherland, M. T., Riedel, M. C., Dick, A. S., Bjork, J. M., Thomas, K. M., … Dale, A. M. (2018). The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. In *Developmental Cognitive Neuroscience* (Vol. 32, pp. 43–54). Elsevier Ltd. https://doi.org/10.1016/j.dcn.2018.03.001

Chen, A. A., Beer, J. C., Tustison, N. J., Cook, P. A., Shinohara, R. T., & Shou, H. (2022). Mitigating site effects in covariance for machine learning in neuroimaging data. *Human Brain Mapping*, *43*(4), 1179–1195. https://doi.org/10.1002/hbm.25688

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, X., Lu, B., Li, H.-X., Li, X.-Y., Wang, Y.-W., Castellanos, F. X., Cao, L.-P., Chen, N.-X., Chen, W., Cheng, Y.-Q., Cui, S.-X., Deng, Z.-Y., Fang, Y.-R., Gong, Q.-Y., Guo, W.-B., Hu, Z.-J.-Y., Kuang, L., Li, B.-J., Li, L., … Yan, C.-G. (2022). The DIRECT consortium and the REST-meta-MDD project: towards neuroimaging biomarkers of major depressive disorder. *Psychoradiology*, *2*(1), 32–42. https://doi.org/10.1093/psyrad/kkac005

Da-ano, R., Masson, I., Lucia, F., Doré, M., Robin, P., Alfieri, J., Rousseau, C., Mervoyer, A., Reinhold, C., Castelli, J., De Crevoisier, R., Rameé, J. F., Pradier, O., Schick, U., Visvikis, D., & Hatt, M. (2020). Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-66110-w

Descoteaux, M., Gerson Unschuld, P., & Bayer, J. M. M. (2022). Site effects how-to and when: An overview of retrospective techniques to accommodate site effects in multi-site neuroimaging analyses. *Frontiers in Neurology*, *13*(923988). http://surfer.nmr.mgh.harvard.edu

Di Martino, A., O'Connor, D., Chen, B., Alaerts, K., Anderson, J. S., Assaf, M., Balsters, J. H., Baxter, L., Beggiato, A., Bernaerts, S., Blanken, L. M. E., Bookheimer, S. Y., Braden, B. B., Byrge, L., Castellanos, F. X., Dapretto, M., Delorme, R., Fair, D. A., Fishman, I., … Milham, M. P. (2017). Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Scientific Data*, *4*, 1–15. https://doi.org/10.1038/sdata.2017.10

Di Martino, A., Yan, C. G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D. A., Gallagher, L., Kennedy, D. P., Keown, C. L., Keysers, C., … Milham, M. P. (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, *19*(6), 659–667. https://doi.org/10.1038/mp.2013.78

Doan, N. T., Kaufmann, T., Bettella, F., Jørgensen, K. N., Brandt, C. L., Moberget, T., Alnæs, D., Douaud, G., Duff, E., Djurovic, S., Melle, I., Ueland, T., Agartz, I., Andreassen, O. A., & Westlye, L. T. (2017). Distinct multivariate brain morphological patterns and their added predictive value with cognitive and polygenic risk scores in mental disorders. *NeuroImage: Clinical*, *15*, 719–731. https://doi.org/10.1016/j.nicl.2017.06.014

Dudley, J. A., Maloney, T. C., Simon, J. O., Atluri, G., Karalunas, S. L., Altaye, M., Epstein, J. N., & Tamm, L. (2023). ABCD_Harmonizer: An Open-source Tool for Mapping and Controlling for Scanner Induced Variance in the Adolescent Brain Cognitive Development Study. *Neuroinformatics*, *21*(2), 323–337. https://doi.org/10.1007/s12021-023-09624-8

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(28), 7900–7905. https://doi.org/10.1073/pnas.1602413113

Eshaghzadeh Torbati, M., Minhas, D. S., Ahmad, G., O'Connor, E. E., Muschelli, J., Laymon, C. M., Yang, Z., Cohen, A. D., Aizenstein, H. J., Klunk, W. E., Christian, B. T., Hwang, S. J., Crainiceanu, C. M., & Tudorascu, D. L. (2021). A multi-scanner neuroimaging data harmonization using RAVEL and ComBat. *NeuroImage*, *245*. https://doi.org/10.1016/j.neuroimage.2021.118703

Feis, R. A., Smith, S. M., Filippini, N., Douaud, G., Dopper, E. G. P., Heise, V., Trachtenberg, A. J., van Swieten, J. C., van Buchem, M. A., Rombouts, S. A. R. B., & Mackay, C. E. (2015). ICA-based artifact removal diminishes scan site differences in multi-center resting-state fMRI. *Frontiers in Neuroscience*, *9*. https://doi.org/10.3389/fnins.2015.00395

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/10.1016/j.neuroimage.2017.08.047

Groves, A. R., Beckmann, C. F., Smith, S. M., & Woolrich, M. W. (2011). Linked independent component analysis for multimodal data fusion. *NeuroImage*, *54*(3), 2198–2217. https://doi.org/10.1016/j.neuroimage.2010.09.073

Groves, A. R., Smith, S. M., Fjell, A. M., Tamnes, C. K., Walhovd, K. B., Douaud, G., Woolrich, M. W., & Westlye, L. T. (2012). Benefits of multi-modal fusion analysis on a large-scale dataset: Life-span patterns of inter-subject variability in cortical morphometry and white matter microstructure. *NeuroImage*, *63*(1), 365–380. https://doi.org/10.1016/j.neuroimage.2012.06.038

Han, Q., Xiao, X., Wang, S., Qin, W., Yu, C., & Liang, M. (2023). Characterization of the effects of outliers on ComBat harmonization for removing inter-site data heterogeneity in multisite neuroimaging studies. *Frontiers in Neuroscience*, *17*. https://doi.org/10.3389/fnins.2023.1146175

Hao, Y., Xu, H., Xia, M., Yan, C., Zhang, Y., Zhou, D., Kärkkäinen, T., Nickerson, L. D., Li, H., & Cong, F. (2023). Removal of site effects and enhancement of signal using dual projection independent component analysis for pooling multi-site MRI data. *European Journal of Neuroscience*, *58*(6), 3466–3487. https://doi.org/10.1111/ejn.16120

Horng, H., Singh, A., Yousefi, B., Cohen, E. A., Haghighi, B., Katz, S., Noël, P. B., Shinohara, R. T., & Kontos, D. (2022). Generalized ComBat harmonization methods for radiomic features with multi-modal distributions and multiple batch effects. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-08412-9

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. In *NeuroImage* (Vol. 274). Academic Press Inc. https://doi.org/10.1016/j.neuroimage.2023.120125

Hyvärinen, A., & Oja, E. (2000). Independent Component Analysis: Algorithms and Applications. In *Neural Networks* (Vol. 13, Issue 5).

Jia, X. Z., Wang, J., Sun, H. Y., Zhang, H., Liao, W., Wang, Z., Yan, C. G., Song, X. W., & Zang, Y. F. (2019). RESTplus: an improved toolkit for resting-state functional magnetic resonance imaging data processing. *Science Bulletin*, *64*(14), 953–954. https://doi.org/10.1016/j.scib.2019.05.008

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

Li, H., Smith, S. M., Gruber, S., Lukas, S. E., Silveri, M. M., Hill, K. P., Killgore, W. D. S., & Nickerson, L. D. (2020). Denoising scanner effects from multimodal MRI data using linked independent component analysis. *NeuroImage*, *208*(116388). https://doi.org/10.1016/j.neuroimage.2019.116388

Maikusa, N., Zhu, Y., Uematsu, A., Yamashita, A., Saotome, K., Okada, N., Kasai, K., Okanoya, K., Yamashita, O., Tanaka, S. C., & Koike, S. (2021). Comparison of traveling-subject and ComBat harmonization methods for assessing structural brain characteristics. *Human Brain Mapping*, *42*(16), 5278–5287. https://doi.org/10.1002/hbm.25615

McKeown, M. J., Hansen, L. K., & Sejnowsk, T. J. (2003). Independent component analysis of functional MRI: What is signal and what is noise? *Current Opinion in Neurobiology*, *13*(5), 620–629. https://doi.org/10.1016/j.conb.2003.09.012

McKeown, M. J., Makeig, S., Brown, G. G., Jung, T. P., Kindermann, S. S., Bell, A. J., & Sejnowski, T. J. (1998). Analysis of fMRI data by blind separation into independent spatial components. *Human Brain Mapping*, *6*(3), 160–188. https://doi.org/10.1002/(SICI)1097-0193(1998)6:3<160::AID-HBM5>3.0.CO;2-1

Meyers, B., Lee, V. K., Dennis, L., Wallace, J., Schmithorst, V., Votava-Smith, J. K., Rajagopalan, V., Herrup, E., Baust, T., Tran, N. N., Hunter, J. V., Licht, D. J., Gaynor, J. W., Andropoulos, D. B., Panigrahy, A., & Ceschin, R. (2022). Harmonization of multi-center diffusion tensor tractography in neonates with congenital heart disease: Optimizing post-processing and application of ComBat. *Neuroimage: Reports*, *2*(3), 100114. https://doi.org/10.1016/j.ynirp.2022.100114

Nichols, T. E., Das, S., Eickhoff, S. B., Evans, A. C., Glatard, T., Hanke, M., Kriegeskorte, N., Milham, M. P., Poldrack, R. A., Poline, J. B., Proal, E., Thirion, B., Van Essen, D. C., White, T., & Yeo, B. T. T. (2017). Best practices in data analysis and sharing in neuroimaging using MRI. In *Nature Neuroscience* (Vol. 20, Issue 3, pp. 299–303). https://doi.org/10.1038/nn.4500

Nygaard, V., Rødland, E. A., & Hovig, E. (2016). Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, *17*(1), 29–39. https://doi.org/10.1093/biostatistics/kxv027

Onicas, A. I., Ware, A. L., Harris, A. D., Beauchamp, M. H., Beaulieu, C., Craig, W., Doan, Q., Freedman, S. B., Goodyear, B. G., Zemek, R., Yeates, K. O., & Lebel, C. (2022). Multisite Harmonization of Structural DTI Networks in Children: An A-CAP Study. *Frontiers in Neurology*, *13*. https://doi.org/10.3389/fneur.2022.850642

Orlhac, F., Lecler, A., Savatovski, J., Goya-Outi, J., Nioche, C., Charbonneau, F., Ayache, N., Frouin, F., Duron, L., & Buvat, I. (2021). How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. *European Radiology*, *31*(4), 2272–2280. https://doi.org/10.1007/s00330-020-07284-9

Parekh, P., Bhalerao, G. V., John, J. P., Venkatasubramanian, G., Viswanath, B., Rao, N. P., Narayanaswamy, J. C., Sivakumar, P. T., Kandasamy, A., Kesavan, M., Mehta, U. M., Mukherjee, O., Purushottam, M., Kannan, R., Mehta, B., Kandavel, T., Binukumar, B., Saini, J., Jayarajan, D., … Jain, S. (2022). Sample size requirement for achieving multisite harmonization using structural brain MRI features. *NeuroImage*, *264*(119768). https://doi.org/10.1016/J.NEUROIMAGE.2022.119768

Radua, J., Vieta, E., Shinohara, R., Kochunov, P., Quidé, Y., Green, M. J., Weickert, C. S., Weickert, T., Bruggemann, J., Kircher, T., Nenadić, I., Cairns, M. J., Seal, M., Schall, U., Henskens, F., Fullerton, J. M., Mowry, B., Pantelis, C., Lenroot, R., … Pineda-Zapata, J. (2020). Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA. *NeuroImage*, *218*. https://doi.org/10.1016/j.neuroimage.2020.116956

Roberts, R. A. C. and S. J. (2001). Flexible Bayesian independent component analysis for blind source separation. *Proc. Int. Conf. on Independent Component Analysis*, *I*(4), 90–95.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., & Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, *23*(SUPPL. 1), 208–219. https://doi.org/10.1016/j.neuroimage.2004.07.051

Sui, J., Adali, T., Yu, Q., Chen, J., & Calhoun, V. D. (2012). A review of multivariate methods for multimodal fusion of brain imaging data. *Journal of Neuroscience Methods*, *204*(1), 68–81. https://doi.org/10.1016/j.jneumeth.2011.10.031

Sui, J., Yu, Q., He, H., Pearlson, G. D., & Calhoun, V. D. (2012). A Selective Review of Multimodal Fusion Methods in Schizophrenia. *Frontiers in Human Neuroscience*, *6*, 1–11. https://doi.org/10.3389/fnhum.2012.00027

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, *15*(3), 1–18. https://doi.org/10.1371/journal.pbio.2000797

Takao, H., Hayashi, N., & Ohtomo, K. (2014). Effects of study design in multi-scanner voxel-based morphometry studies. *NeuroImage*, *84*, 133–140. https://doi.org/10.1016/j.neuroimage.2013.08.046

Tanaka, S. C., Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., Kunimatsu, A., Okada, N., Hashimoto, R., Okada, G., Sakai, Y., Morimoto, J., Narumoto, J., Shimada,

Y., Mano, H., Yoshida, W., … Imamizu, H. (2021). A multi-site, multi-disorder resting-state magnetic resonance image database. *Scientific Data*, *8*(1). https://doi.org/10.1038/s41597-021-01004-8

Thompson, P. M., Stein, J. L., Medland, S. E., Hibar, D. P., Vasquez, A. A., Renteria, M. E., Toro, R., Jahanshad, N., Schumann, G., Franke, B., Wright, M. J., Martin, N. G., Agartz, I., Alda, M., Alhusaini, S., Almasy, L., Almeida, J., Alpert, K., Andreasen, N. C., … Drevets, W. (2014). The ENIGMA Consortium: Large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging and Behavior*, *8*(2), 153–182. https://doi.org/10.1007/s11682-013-9269-5

Van Der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Xia, M., Wang, J., & He, Y. (2013). BrainNet Viewer: A Network Visualization Tool for Human Brain Connectomics. *PLoS ONE*, *8*(7). https://doi.org/10.1371/journal.pone.0068910

Xu, H., Hao, Y., Zhang, Y., Zhou, D., Kärkkäinen, T., Nickerson, L. D., Li, H., & Cong, F. (2023). Harmonization of multi-site functional MRI data with dual-projection based ICA model. *Frontiers in Neuroscience*, *17*. https://doi.org/10.3389/fnins.2023.1225606

Xu, H., Li, H., Kärkkäinen, T., & Cong, F. (2023). Enhancing Performance of Linked Independent Component Analysis: Investigating the Influence of Subjects and Modalities. *Proceedings - 2023 International Conference on Computers, Information Processing and Advanced Education (CIPAE)*, 726–732. https://doi.org/10.1109/CIPAE60493.2023.00141

Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., Kunimatsu, A., Okada, N., Yamagata, H., Matsuo, K., Hashimoto, R., Okada, G., Sakai, Y., Morimoto, J., Narumoto, J., Shimada, Y., Kasai, K., … Imamizu, H. (2019). Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. *PLoS Biology*, *17*(4), 1–34. https://doi.org/10.1101/440875

Yan, C. G., Chen, X., Li, L., Castellanos, F. X., Bai, T. J., Bo, Q. J., Cao, J., Chen, G. M., Chen, N. X., Chen, W., Cheng, C., Cheng, Y. Q., Cui, X. L., Duan, J., Fang, Y. R., Gong, Q. Y., Guo, W. Bin, Hou, Z. H., Hu, L., … Zang, Y. F. (2019). Reduced default mode network functional connectivity in patients with recurrent major depressive disorder. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(18), 9078–9083. https://doi.org/10.1073/pnas.1900390116

Yan, C. G., Wang, X. Di, Zuo, X. N., & Zang, Y. F. (2016). DPABI: Data Processing & Analysis for (Resting-State) Brain Imaging. *Neuroinformatics*, *14*(3), 339–351. https://doi.org/10.1007/s12021-016-9299-4

Yeung, A. W. K. (2018). An Updated Survey on Statistical Thresholding and Sample Size of fMRI Studies. *Frontiers in Human Neuroscience*, *12*, 1–7. https://doi.org/10.3389/fnhum.2018.00016

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/10.1002/hbm.24241

Zang, Y. F., Yong, H., Chao-Zhe, Z., Qing-Jiu, C., Man-Qiu, S., Meng, L., Li-Xia, T., Tian-Zi, J., & Yu-Feng, W. (2007). Altered baseline brain activity in children with ADHD revealed by resting-state functional MRI. *Brain and Development*, *29*(2). https://doi.org/10.1016/j.braindev.2006.07.002

Zang, Y., Jiang, T., Lu, Y., He, Y., & Tian, L. (2004). Regional homogeneity approach to fMRI data analysis. *NeuroImage*, *22*(1), 394–400. https://doi.org/10.1016/j.neuroimage.2003.12.030

Zou, Q. H., Zhu, C. Z., Yang, Y., Zuo, X. N., Long, X. Y., Cao, Q. J., Wang, Y. F., & Zang, Y. F. (2008). An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: Fractional ALFF. *Journal of Neuroscience Methods*, *172*(1). https://doi.org/10.1016/j.jneumeth.2008.04.012

# ORIGINAL PAPERS

# I

## REMOVAL OF SITE EFFECTS AND ENHANCEMENT OF SIGNAL USING DUAL PROJECTION INDEPENDENT COMPONENT ANALYSIS FOR POOLING MULTI-SITE MRI DATA

by

Hao, Yuxing, Huashuai Xu, Mingrui Xia, Chenwei Yan, Yunge Zhang, Dongyue Zhou, Tommi Kärkkäinen, Lisa D. Nickerson, Huanjie Li, and Fengyu Cong, 2023

European Journal of Neuroscience, 58(6): 3466-3487

https://doi.org/10.1111/ejn.16120

# Removal of site effects and enhancement of signal using dual projection independent component analysis for pooling multi-site MRI data

Yuxing Hao[a,1], Huashuai Xu[b,1], Mingrui Xia[c,d,e], Chenwei Yan[a], Yunge Zhang[a], Dongyue Zhou[a], Tommi Kärkkäinen[b], Lisa D. Nickerson [f,g,*], Huanjie Li[a,*], Fengyu Cong[a,b,h,i]

*a School of Biomedical Engineering, Dalian University of Technology, Dalian, China*

*b Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland*

*c State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, China*

*d Beijing Key Laboratory of Brain Imaging and Connectomics, Beijing Normal University, Beijing, China*

*e IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, China*

*f McLean Imaging Center, McLean Hospital, Belmont, MA, United States*

*g Department of Psychiatry, Harvard Medical School, Boston, MA, United States*

*h School of Artificial Intelligence, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, China*

*i Key Laboratory of Integrated Circuit and Biomedical Electronic System, Liaoning Province. Dalian University of Technology, Dalian, China*

***1 Yuxing Hao and Huashuai Xu contributed equally to this study.***

## Conflict of interest

The authors declare no conflict of interest.

## Acknowledgment

This work was supported by STI 2030 - Major Projects 2022ZD0211500, Science and Technology Planning Project of Liaoning Province (no. 2022JH2/10700002 and 2021JH1/10400049), National Natural Science Foundation of China [grant numbers 81601484], National Foundation in China [grant number JCKY 2019110B009], National Institutes of Health [NIA RF1 AG078304].

## Data availability statement

The simulated MRI data are from https://github.com/ThomasYeoLab/CBIG/tree/master /stable_projects/brain_ parcellation/Schaefer2018_LocalGlobal/Parcellations/.

All real MRI data used in this study are publicly available, including Autism Brain Imaging Data Exchange II dataset (ABIDE II) (http://fcon_1000.projects.nitrc.org/indi/abide/ abide_II.html), travelling subject dataset from DecNef Project Brain Data Repository website (https://bicr-resource.atr.jp/srpbsts/).

* Corresponding authors:
E-mail addresses: hj_li@dlut.edu.cn (Huanjie Li), lisa_nickerson@hms.harvard.edu (Lisa D. Nickerson)

# Removal of site effects and enhancement of signal using dual projection independent component analysis for pooling multi-site MRI data

Yuxing Hao[a,1], Huashuai Xu[b,1], Mingrui Xia[c,d,e], Chenwei Yan[a], Yunge Zhang[a], Dongyue Zhou[a], Tommi Kärkkäinen[b], Lisa D. Nickerson [f,g,*], Huanjie Li[a,*], Fengyu Cong[a,b,h,i]

*a School of Biomedical Engineering, Dalian University of Technology, Dalian, China*

*b Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland*

*c State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, China*

*d Beijing Key Laboratory of Brain Imaging and Connectomics, Beijing Normal University, Beijing, China*

*e IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, China*

*f McLean Imaging Center, McLean Hospital, Belmont, MA, United States*

*g Department of Psychiatry, Harvard Medical School, Boston, MA, United States*

*h School of Artificial Intelligence, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, China*

*i Key Laboratory of Integrated Circuit and Biomedical Electronic System, Liaoning Province. Dalian University of Technology, Dalian, China*

***1 Yuxing Hao and Huashuai Xu contributed equally to this study.***

## Abstract

Combining magnetic resonance imaging (MRI) data from multi-site studies is a popular approach for constructing larger datasets to greatly enhance the reliability and reproducibility of neuroscience research. However, the scanner/site variability is a significant confound that complicates the interpretation of the results, so effective and complete removal of the scanner/site variability is necessary to realize the full advantages of pooling multi-site datasets. Independent component analysis (ICA) and general linear model (GLM) based harmonization methods are the two primary methods used to eliminate scanner/site effects. Unfortunately, there are challenges with both ICA-based and GLM-based harmonization methods to remove site effects completely when the signals of interest and scanner/site effects-related variables are correlated, which may occur in neuroscience studies. In this study, we propose an effective and powerful harmonization strategy that implements dual projection (DP) theory based on ICA to remove the scanner/site effects more completely. This method can separate the signal effects correlated with site variables from the identified site effects for removal without losing signals of interest. Both simulations and vivo structural MRI datasets, including a dataset from Autism Brain Imaging Data Exchange II and a traveling subject dataset from the Strategic Research Program for Brain Sciences, were used to test the performance of a DP-based ICA

harmonization method. Results show that DP-based ICA harmonization has superior performance for removing side effects and enhancing the sensitivity to detect signals of interest as compared with GLM-based and conventional ICA harmonization methods.

***Keywords***: dual projection, harmonization, independent component analysis, magnetic resonance imaging, multi-site, site effects

# 1 Introduction

It is now common practice to pool multi-site magnetic resonance imaging (MRI) datasets to study brain biomarkers of neuroscience, neuropsychiatry, and neurology to promote rigor and reproducibility of results (Button et al., 2013; Eickhoff et al., 2016; Van Horn & Toga, 2009). However, combining multiple datasets does introduce site-related effects, which confounds effects of interest, and complicates the interpretation of the final results (Casey et al., 1998; Focke et al., 2011; Friedman et al., 2008; Pohl et al., 2016; Takao et al., 2011; Venkatraman et al., 2015; Vollmar et al., 2010; Wegner et al., 2008; Zivadinov & Cox, 2008). Site-related effects arises from differences in scanners manufacturers, field strengths, hardware, software, pulse sequences, quality control, and data quality across sites (Jovicich et al., 2009). It has been shown that differences in acquisition parameters and software and hardware upgrades during data collection using the same scanner have non-negligible effects on almost all image-derived phenotypes from structural images (such as cortical surface and gray matter volume), diffusion-weighted images (such as diffusion tensor image (DTI) measures), and functional MRI (fMRI) data (Groves et al., 2011; Li et al., 2020). Hence, effective removal or deconfounding of site-related variability from the MRI data is a critical step to ensure the accuracy and reproducibility of findings generated from combined datasets.

Several approaches have been proposed for the harmonization of multi-site MRI data, including methods based on the general linear model (GLM) (Fennema-Notestine et al., 2007; Glover et al., 2012; Venkatraman et al., 2015), and data-driven unsupervised learning methods such as independent component analysis (ICA) (Chen et al., 2014; Li et al., 2020) and recently proposed deep learning methods (C Monte-Rubio et al., 2022; Dinsdale et al., 2021; Tian et al., 2022). Two of the most popular methods to eliminate or minimize the site-related effects are based on GLM and ICA (Chen et al., 2014). GLM-based harmonization method is easy to implement and often used in multi-site MRI data studies to minimize the site-related effects, in this case, it utilizes site/study variables as covariates of no interest in group-level GLM analysis to control for site-related effects. Fortin et al. (Fortin et al., 2017) have adapted a GLM-based technique called ComBat (Johnson et al., 2007), an empirical Bayesian method for data harmonization that is popular in the field of genetics, to remove unwanted variation induced by sites while preserving the signal-related variation in neuroimaging studies. ComBat has been applied to harmonize DTI measures (Fortin et al., 2017), cortical thicknesses and functional connectivity measures (Yu et al., 2018), magnetic resonance spectroscopy measures (T. K. Bell

et al., 2022), and positron emission tomography (PET) outcomes (Orlhac et al., 2018) showing good performance for removing site effects. ICA is an unsupervised data-driven statistical method that factorizes or decomposes the image data into a set of statistically independent non-Gaussian components reflecting different sources that generate the measured imaging data. And the site-related components can be removed by regressing them from the original data to generate a harmonized clean dataset for further analysis (Chen et al. 2014). In this case, the ICA has been used to do a data-driven estimation of the site/scanner-related covariates of no interest that are regressed out of the data rather than creating covariates to model the site-scanner effects based on strong assumptions (e.g., regressors are used that assume a constant effect for each site/scanner, which ignores within-site/data to day variations in these effects). ICA is typically applied to harmonize individual MRI modalities, however, our previous work (Li et al., 2020) proposed a harmonization method for multi-modal imaging measures that implemented linked ICA (LICA) (Groves et al., 2011) as a novel approach to eliminate scanner effects from multi-study data. LICA simultaneously decomposes the multi-modal imaging data (for example, structural plus diffusion MRI-derived measures) into a set of multi-modal components and a set of subject loadings quantifying the strength of each multi-modal component in each individual, with components reflecting true signals of interest as multi-modal covariance patterns, as well as artifacts and variability related to uninteresting effects like scanner and site differences. We found that several of the resulting LICA components from an analysis of multi-study data with scanner effects were associated with scanner variations and that these patterns could be effectively regressed from the data to obtain harmonized data relatively free from scanner effects. We showed that multi-modal ICA-based harmonization was more effective at removing scanner-related effects compared with the conventional GLM and single-modality ICA harmonization methods. The reason for its superior performance is that even though all three approaches involve regression to remove scanner effects from the data, the data-driven estimates of the scanner effects from LICA of multi-modal MRI data provided more accurate model of site effects than assuming a constant effect or estimating effects based on single modality ICA to use as nuisance covariates for harmonization.

In the present study, we aim to address another limitation of current methods for harmonization scanner/site effects, namely existing methods for harmonization site/scanner effects ignore the possibility of correlations between these effects and the effects of interest. For the conventional GLM approach, site-related variables are included as covariates of no interest or may be regressed out of data prior to higher-level statistical modeling, which may lead to the removal of interesting signals that are correlated with scanner/site variables and to weaker specificity of harmonization. While ComBat tries to preserve the signal-related variation when harmonization scanner/site effects, similar to the conventional GLM approach, it also assumes a constant effect for all datasets collected from the same site or the same scanner state, thus also ignoring the day-to-day variations in scanner performance. While ICA and LICA can identify scanner/site effects that capture day to day variations in scanner performance (Li

et al., 2020), both approaches are vulnerable to identifying components that reflect a mixture of signal and scanner/site effects, rather than separating the effects into two separate components. In our previous work, to retain signals of interest, only components that were associated with site effects and not signals of interest (e.g., had subject loadings that correlated only with site variables and not variables of interest) were removed from the data while mixed components were retained as a conservative approach to harmonization (Chen et al., 2014; Li et al., 2020). One possibility to address this limitation for ICA-based techniques is to run the ICA with several different model orders to identify a decomposition with stable pure site effects related components not mixed with signals of interest. In practice, it is challenging to do this as different mixtures may arise at different model orders such that it may not be possible to have full separation at any model order.

To solve this problem for ICA-based methods, we propose a new ICA with dual-projection (ICA-DP) technique for harmonization scanner/site effects. In this study, we focus on single modality ICA, with extension to multi-modal ICA to be done in future work. For ICA-DP, mixed components from single-modality ICA are separated into a part related to signal only and a part related only to site effects by applying a projection procedure. The site effects extracted from the mixed components via the projection step are combined with the other ICA components that reflected only site/scanner variance and are then removed from the data using a second projection procedure. Our new method is tested using simulated MRI data and in vivo multi-site datasets to assess the performance of ICA-DP compared to conventional ICA and ComBat harmonization methods.

# 2 Methods

## 2.1 Dual-projection harmonization improvement

### 2.1.1 Traditional ICA-based harmonization

ICA decomposition model for group structural MRI data can be expressed as:

$$\mathbf{Y} = \mathbf{LS}, \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^{M \times N}$ denotes the original data to be decomposed, $\mathbf{S} \in \mathbb{R}^{R \times N}$ contains independent spatial maps and $\mathbf{L} \in \mathbb{R}^{M \times R}$ contains their corresponding loadings. And $M$, $N$ and $R$ denote the number of subjects, voxels and components, respectively.

When removing site effects of multi-site structural MRI data, the original ICA-based harmonization methods (Chen et al., 2014; Li et al., 2020) only eliminate the pure site-related components (only related to site effects) to avoid discarding useful information such as diagnoses or symptom measures. For comparison, we designate this ICA-based harmonization method as ICA-SP (single projection) as it only uses one step of projecting:

$$\mathbf{Y}_{clean}^{ICA-SP} = \mathbf{Y}_{orig} - \mathbf{L}_N \cdot pinv(\mathbf{L}_N) \cdot \mathbf{Y}_{orig}, \tag{2}$$

where $\mathbf{L}_N$ is the loadings of pure site effects components (components related to site effects)

and $\mathbf{Y}_{clean}^{ICA-SP}$ is the harmonized data derived from ICA-SP harmonization methods.

Though it may preserve signal-related information well, it is too soft to remove the site effects as it does nothing with the mixed components and is more than likely to find site effects in its harmonized data.

### 2.1.2 Proposed ICA-DP harmonization

The ICA-DP harmonization procedure is summarized in Fig. 1. ICA-DP is inspired by the dual-regression approach for projecting a participant's fMRI data onto a set of spatial maps derived from ICA of multi-subject fMRI data to identify subject-level spatial maps corresponding to each group level component (Beckmann et al., 2009; Filippini et al., 2009; Nickerson et al., 2017).

First, the subject series is decomposed by ICA, and the resulting subject loadings **L**, of each component that reflects the strength of the corresponding variables represented in the IC map (could be interesting signal, scanner/site effects, or a mixture) are labeled as loadings for pure site effects components $\mathbf{L}_N$, pure signal components $\mathbf{L}_S$, or mixed components $\mathbf{L}_M$ by calculating the correlations between the loadings and all the signal and site effects variables (i.e., components exhibiting significant associations $p$<0.05 are identified as signal- and/or site effects-related ones).

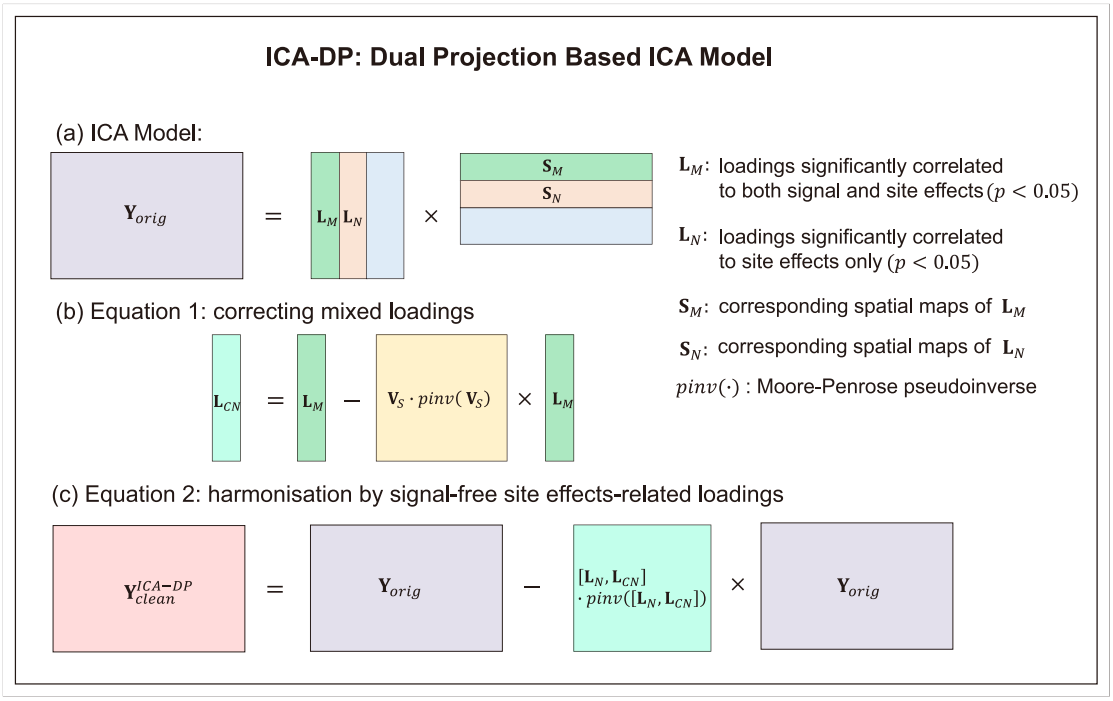


Fig. 1. The procedures of ICA-DP harmonization method. (a) Identifying the loadings extracted by ICA that related to site effects variables (including mixed ones that significantly correlated to both site effects and signal, and the ones only significantly correlated to site effects). (b) Correcting the mixed loadings to only site effects-related ones ($\mathbf{L}_{CN}$) by projecting out signal-related information. (c) Obtaining cleaned data by removing the integral site effects-related components ($\mathbf{L}_N, \mathbf{L}_{CN}$).

The first projection procedure is used to separate the signal effects out from $\mathbf{L}_M$ as below:

$$\mathbf{L}_{CN} = \mathbf{L}_M - \mathbf{V}_S \cdot pinv(\mathbf{V}_S) \cdot \mathbf{L}_M, \quad (3)$$

[1]https://github.com/ThomasYeoLab/CBIG/tree/master/stable_projects/brain_parcellation/Schaefer2018_LocalGlobal/Parcellations/

where $\mathbf{V}_S$ denotes the variables of interest (signals to be preserved, e.g., age, gender or health condition), $\mathbf{L}_{CN}$ denotes the corrected site effects contributions to the mixed components, and $pinv(\cdot)$ denotes the Moore-Penrose inverse (pseudoinverse) of a non-square matrix. Thus, the signal information is projected out from $\mathbf{L}_M$ and we can identify the site effects for the mixed components.

Then $[\mathbf{L}_N, \mathbf{L}_{CN}]$ represent the total site-related effects present in the data, which are then cleaned from the subject series via a second projection procedure:

$$\mathbf{Y}_{clean}^{ICA-DP} = \mathbf{Y}_{orig} - [\mathbf{L}_N, \mathbf{L}_{CN}] \cdot pinv([\mathbf{L}_N, \mathbf{L}_{CN}]) \cdot \mathbf{Y}_{orig}, \quad (4)$$

where $\mathbf{Y}_{orig}$ denotes the subject series of spatial maps and $\mathbf{Y}_{clean}^{ICA-DP}$ denotes the harmonized MRI data free from site/scanner effects that can be used for further analysis.

### 2.2 Study Data

#### 2.2.1 Simulated Data

The simulated structural MRI data, including 100 subjects, were generated in this study. For each subject, the data was generated by computing 10 spatial maps and one set of ground truth subject loadings (Eq. 1). Each component map was multiplied by the corresponding subject loading, and then they were added together to obtain the simulated MRI data for each subject. The spatial maps were gotten by combining different areas of the standard brain template as shown in Fig. 2[1]. To make our simulated data much closer to real MRI data, two kinds of spatial maps were simulated, one is all the spatial maps of 10 components were spatially independent and the other is two components' spatial maps were overlapped (Fig. 2). For each condition, 100 subjects were generated, and the subject-specific data shared the same spatial maps, and the difference was the weights in its loadings corresponding to the spatial maps. Three different types of relationships between subject loadings and signal/site effects variables were simulated in this study: (1) signal variable was not significantly correlated to site effects variables; subject loadings were linearly correlated to signal and (or) site effects variables (Table 1). Among the 10 components, the first four components were significantly related to signal and (or) site effects variables, and the other components were not related to the variables we are interested in. Components 1 and 2 are mixed components, which are related with both signal and site effects variables. The difference is that component 1 is much more related to signal, and component 2 is more correlated to site effects. Component 3 is pure site effects components, which only significantly correlated with site effects variable. Component 4 is a pure signal component that only significantly correlated with signal variable; (2) Signal variable is significantly correlated to site effects variable, subject loadings are linearly correlated to signal and (or) site effects variables (Table 2). Since the signal variable was significantly correlated to the site effects variable, there were no pure signal or site effects components under this condition, so we selected the first 2 components as mixed components, component 1 is much more related to the signal, and component 2 is much more correlated to site effects. Three

[1] https://github.com/ThomasYeoLab/CBIG/tree/master/stable_projects/brain_parcellation/Schaefer2018_LocalGlobal/Parcellations/

different correlation levels (from low to high) between signal and site effects variables were simulated in this study to show the harmonization power of ICA-DP.

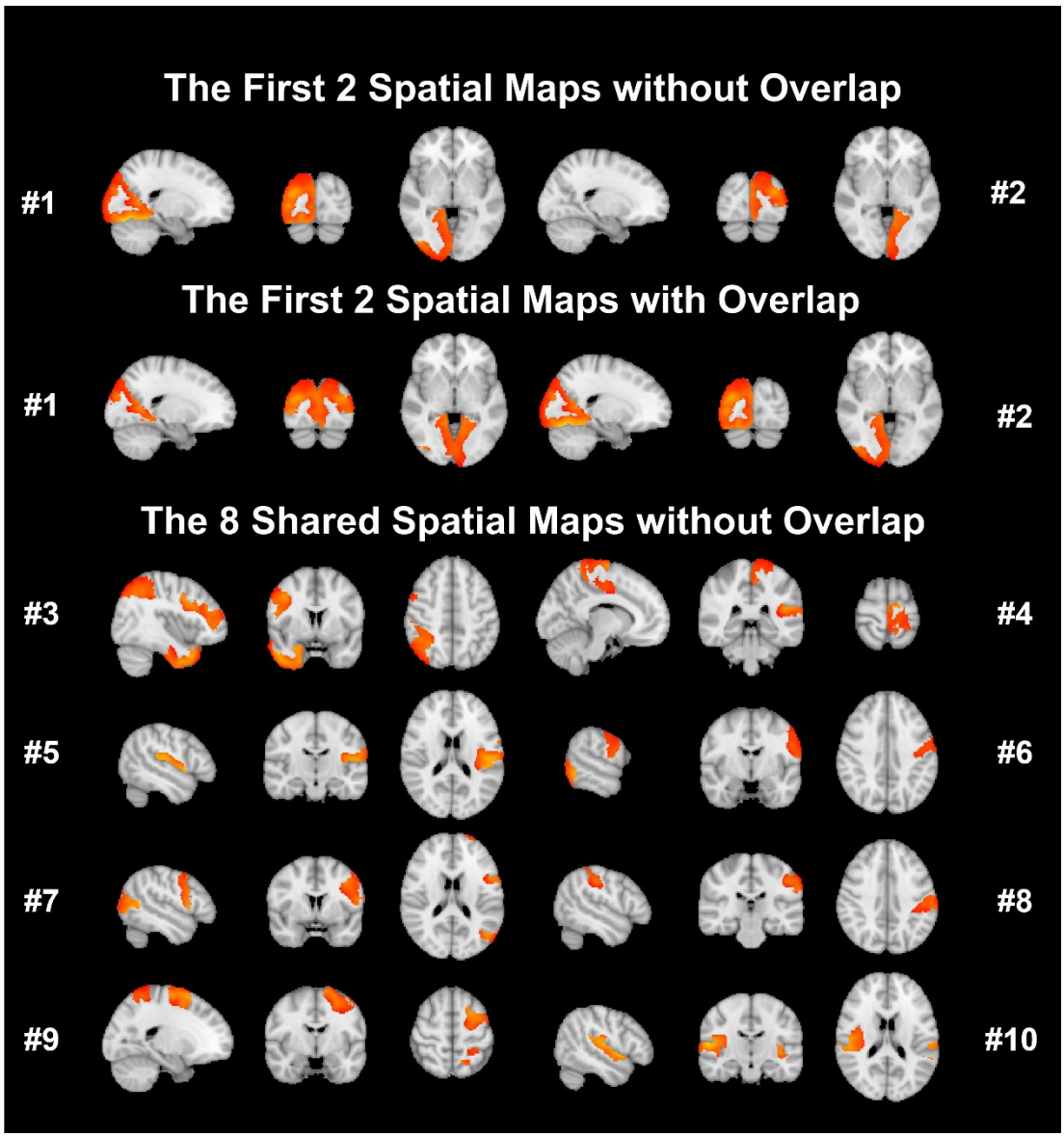


Fig. 2. Ten independent brain spatial maps used to simulate MRI data. Situation1: there is no overlap among all the spatial maps; Situation 2: The first two components were spatially overlapped, and the other 8 components share the same maps with situation 1.

**Table 1** Pearson correlation coefficients and corresponding *p* values by correlating variables and loadings.

| #Component | Signal Variable (*r*/*p*) | Site Effects Variable (*r*/*p*) |
|---|---|---|
| 1 | 0.9425(<0.0001) | 0.3999(<0.0001) |
| 2 | 0.2999(0.0024) | 0.9728(<0.0001) |
| 3 | -- | 0.5999(<0.0001) |
| 4 | 0.5999(<0.0001) | -- |

**Note:** Component loadings are linearly correlated with signal and site effects variables, while the signal variable is not significantly correlated to the site effects variable. Components 1 (more related to signal) and 2 (more related to site effects) are mixed components. Component 3 is only related to the site effects variable, and component 4 is only related to the signal variable. The relationship of loadings and variables is expressed by *r*-value and *p*-value. -- denotes not significantly correlated.

**Table 2** Pearson correlation coefficients and corresponding *p* values by correlating variables and loadings.

| Correlation between Signal and Site Effects | #Component | Signal | Site effects |
|---|---|---|---|
| 0.2999 (2.4e-3) | 1 | 0.7946(<0.0001) | 0.2412(<0.0001) |
| | 2 | 0.2590(0.0093) | 0.7959(<0.0001) |
| 0.4999 (<0.0001) | 1 | 0.7962(<0.0001) | 0.4279(<0.0001) |
| | 2 | 0.3447(0.0005) | 0.7859(<0.0001) |
| 0.6999 (<0.0001) | 1 | 0.7957(<0.0001) | 0.5932(<0.0001) |
| | 2 | 0.4993(<0.0001) | 0.7761(<0.0001) |

**Note:** Component loadings are linearly correlated with signal and site effects variables, while signal variable is significantly correlated to site effects variable. Components 1 (more related to signal) and 2 (more related to site effects) are mixed components. The relationship of loadings and variables is expressed by *r*-value and *p*-value. Three different correlation levels between signal and site effects variables are simulated in this study with *r*-values of 0.2999 ($p$=0.0024), 0.4999 ($p$=1.2e-7), and 0.6999 ($p$=5.6e-16), corresponding to low, medium and high correlation levels.

### 2.2.2 Multi-site MRI data from ABIDE II

High spatial resolution structural MRI data of 606 subjects (including Autism Spectrum Disorder (ASD) patients: 225, Healthy Controls (HC): 381) were obtained from Autism Brain Imaging Data Exchange II dataset (ABIDE II) (http://fcon_1000.projects.nitrc.org/indi/abide/abide_II.html). The data were collected from 13 different sites, all the data were acquired from 3T scanner with different manufacturers (Siemens, Philips, and GE) (Di Martino et al., 2017). The acquisition parameters: scanner/site and imaging-related details, including repetition time (TR), echo time (TE), flip angle (FA), and voxel size in Table 3. The demographic information: ASD/HC, gender, and age are summarized in Table 4. Subjects with ASD could be divided into two categories: ASD only and ASD with comorbidity (Attention-Deficit/Hyperactivity Disorder, anxiety or others) (Di Martino et al., 2017).

**Table 3** Scanning parameters and demographic information of the multi-site ABIDE II data.

| Sites | Scanners | TR/TE (ms) | FA (degree) | Voxel Size |
|---|---|---|---|---|
| EMC | GE MR750 | 1664/4.24 | 16 | 0.9×0.9×0.9 |
| ETH | PhilipsAchieva | 3000/3.9 | 8 | 0.9×0.9×0.9 |
| GU | Siemens TriTim | 2530/3.5 | 7 | 1×1×1 |
| IU | Siemens TriTim | 2400/2.3 | 8 | 0.7×0.7×0.7 |
| KKI | PhilipsAchieva | 3500/3.7 | 8 | 1×1×1 |
| KUL | PhilipsAchieva | 2000/4.6 | 8 | 1×1×1.2 |
| OHSU | Siemens TriTim | 2300/3.58 | 10 | 1×1×1.1 |
| ONRC | Siemens Skyra | 2200/2.88 | 13 | 0.8×0.8×0.8 |
| SU | GE SIGNA | 5.9/1.8 | 11 | 1×1×1 |
| UCD | Siemens TriTim | 2000/3.16 | 8 | 1×1×1 |
| UCLA | Siemens TriTim | 2300/2.86 | 9 | 1×1×1.2 |
| UM | GE Healthcare | -/- | 12 | 1×1×1 |
| USM | Siemens TriTim | 2300/2.91 | 9 | 1×1×1.2 |

**Note:** The data were collected from 13 different sites: Erasmus University Medical Center (EMC), ETH Zürich (ETH,), Georgetown University (GU), Indiana University (IU), Kennedy Krieger Institute (KKI), Katholieke Universiteit Leuven (KUL), Oregon Health and Science University (OHSU), Olin Neuropsychiatry Research Center (ONRC), Stanford University (SU), University of California Davis (UCD), University of California Los Angeles (UCLA), University of Miami (UM), University of Utah School of Medicine (USM).

In this study, the site differences are defined as nuisance variables to eliminate, while group differences (ASD/HC), age and gender are regarded as signal variables. The correlation coefficients among these variables are summarized in Table 5. Since site differences are categorical variables and calculating the correlation coefficients between categorical variables and numeric variables directly is not achievable, we used ANOVA to calculate the significant levels of signal variables and site effects variables to identify the independent components related to site effects significantly. For ICA analysis, the components that only significantly correlated to site variables were regarded as pure site effects components, and the components that significantly correlated to both site and signal variables were regarded as mixed components.

**Table 4** Demographic information of the multi-site ABIDE II data collected from 13 sites.

| Sites | ASD/HC | ASD with comorbidity | Male/Female | Age Range (Mean) | Full IQ Standard Score (Mean) |
|---|---|---|---|---|---|
| EMC | 18/20 | 11 | 31/7 | 6.40~10.66 (8.28) | -- |
| ETH | 8/17 | -- | 25/0 | 13.83~29.42 (22.43) | 82~133 (112.84) |
| GU | 33/43 | -- | 51/25 | 8.06~13.88 (10.74) | 92~149 (119.17) |
| IU | 18/18 | -- | 28/8 | 17~54 (24.61) | 80~135 (116.36) |
| KKI | 32/133 | 29 | 103/62 | 8.02~12.99 (10.36) | 83~143 (113.26) |
| KUL | 7/0 | 1 | 7/0 | 18~25 (21.71) | 73~146 (103.86) |
| OHSU | 33/51 | 23 | 52/32 | 7~15 (10.94) | 72~140 (113.11) |
| ONRC | 16/29 | 5 | 32/13 | 18~31 (23.24) | 86~146 (111.76) |
| SU | 15/17 | 2 | 29/3 | 8.42~12.99 (10.99) | 93~151 (115.31) |
| UCD | 13/13 | 4 | 19/7 | 12~17.83 (15.03) | 83~128 (107.96) |
| UCLA | 12/12 | -- | 19/5 | 7.75~15.03 (11.04) | 78~141 (107.96) |
| UM | 7/12 | -- | 14/5 | 7.3~14.3 (10.32) | 98~144 (115.72) |
| USM | 13/16 | -- | 24/5 | 9.12~38.86 (21.21) | 73~144 (108.5) |

**Table 5** The relationship of signal and site effects variables for real MRI data (*p* values).

| Correlation | Site | ASD vs HC | Age | Gender |
|---|---|---|---|---|
| Site | <0.0001 | <0.0001 | <0.0001 | 0.0004 |
| ASD vs HC | <0.0001 | <0.0001 | - | <0.0001 |
| Age | <0.0001 | - | <0.0001 | - |
| Gender | 0.0004 | <0.0001 | - | <0.0001 |

### 2.2.3 Traveling subjects

To further validate the site effects removing efficiency of ICA-DP, the high spatial resolution structural MRI data from a traveling-subject dataset including 9 healthy participants (all males, age: 27±2.6) scanned at 12 different sites from the DecNef Project Brain Data Repository (https://bicr-resource.atr.jp/srpbsts/) were used in this study. For T1-weighted MRI data of the 12 different sites, there were two phase-encoding directions (PA and AP), three MRI manufacturers (Siemens, GE, and Philips) with seven scanner types (TimTrio, Verio, Skyra, Spectra, MR750W, SignaHDxt, and Achieva) and four channels per coil (8, 12, 24, and 32) (Maikusa et al., 2021; Tanaka et al., 2021). Scanning parameters, including repetition time (TR), echo time (TE), flip angle (FA), and voxel size, are summarized in Table 6. Three sites (i.e., ATT, UTO, and YC2) were excluded for harmonization analysis because they contain duplicate data (there are 7 duplicate images in ATT and ATV, 2 same images within both YC2 and UTO).

As the images from this dataset are the same groups under different sites, site effects variables and subject variables (including subject labels, age, etc.) are uncorrelated.

**Table 6** Scanning parameters of the traveling-subject dataset.

| Sites | Scanners | TR/TE (ms) | FA (degree) | Voxel Size |
|---|---|---|---|---|
| ATT | SiemensTimTrio | 2300/2.98 | 9 | 1×1×1 |
| ATV | Siemens Verio | 2300/2.98 | 9 | 1×1×1 |
| COI | Siemens Verio | 2300/2.98 | 9 | 1×1×1 |
| HKH | Siemens Spectra | 1900/2.38 | 10 | 0.8×0.75×0.75 |
| HUH | GE Signa HDxt | 6788/1.928 | 20 | 1×1×1 |
| KPM | Philips Achieva | 7.1/3.31 | 10 | 1×1×1 |
| KUS | Siemens Skyra | 2300/2.98 | 9 | 1×1×1 |
| KUT | SiemensTimTrio | 2000/3.4 | 8 | 0.9375×0.9375×1 |
| SWA | Siemens Verio | 2300/2.98 | 9 | 1×1×1 |
| UTO | GE MR750W | 7.7/3.1 | 11 | 1×1.0156×1.0156 |
| YC1 | Philips Achieva | 6.99/3.176 | 9 | 1×1×1 |

**Note:** The datasets include 9 healthy subjects undergoing T1-weighted MRI scans at 12 different sites, and all of them used 3T scanners and the same acquisition parameters but with different manufacturers and hardware versions (Siemens, GE, and Philips).

### 2.3 Data analysis

For both the ABIDE II dataset and the traveling dataset, the modulated gray matter (GM) images were analyzed with FSL-VBM (Douaud et al., 2007) (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/ FSLVBM), an optimized VBM protocol (Good et al., 2001) carried out with FSL tools (Smith et al., 2004). First, structural images were brain-extracted and grey matter-segmented before being registered to the MNI 152 standard space using non-linear registration (Andersson et al., 2007). The resulting images were averaged and flipped along the x-axis to create a left-right symmetric, study-specific grey matter template. Second, all native grey matter images were non-linearly registered to this study-specific template and "modulated" to correct for local expansion (or contraction) due to the non-linear component of the spatial transformation. The modulated GM images were then smoothed with an isotropic Gaussian kernel with a sigma of 3 mm.

### 2.4 Comparison of ICA-DP with other harmonization methods

General linear model (GLM) harmonization(Maikusa et al., 2021; Venkatraman et al., 2015; Yamashita et al., 2019) and ComBat harmonization(Fortin et al., 2017) are two main model-baesd methods for removing site effect differences (see the Supplementary Information for detailed descriptions of these methods). To examine the advantages of our proposed methods, we compared ICA-DP with ICA-SP and these model-based harmonization methods in terms of site effect removal and biological variability preservation.

#### 2.4.1 Site effects removing with ICA-SP/DP methods

Firstly, ICA was applied to the non-harmonized data to identify pure site effects components, pure signal components and mixed components. For simulated data, the Pearson correlation

coefficient between subject loadings and signal (or site effects) variable was used to measure the properties of components. Those components only related with the signal variable ($p < 0.05$) were identified as pure signal components; those components only related with the site effects variable ($p < 0.05$) were identified as pure site effects components; those components related with the signal and site effects variables ($p < 0.05$) were identified as mixed components.

For the real MRI data from ABIDE II and the traveling-subject data, we used the Pearson correlation coefficient and ANOVA to identify signal, site effects and mixed components (age, gender, and group difference(ASD/HC) were signal variables of interest). The subject loadings from one site were divided into one variable, then 13 levels- for ABIDE II and 9 levels- for the traveling-subject data. ANOVA was used to calculate the significant levels of subject loadings and site differences and identify the independent components related to site effects significantly. Finally, those components whose *p*-values of ANOVA were significant (using Bonferroni correction to adjust for multiple comparisons, adjusted $p < 0.05$) were identified as site effects components. The intersection of signal components and site-related components were mixed components.

For the ICA-SP method, only pure site effects components were used to regress the site effects. For the ICADP method, all the site effects components, including the mixed ones, were used for site effects removal. The site effects extracted from the mixed and pure site effects components will be regarded as the integral site-related effects for removal by ICA-DP. All the ICA analyses were based on MATLAB and FSL MELODIC.

**2.4.2 Site effects removing with GLM and ComBat methods**

For the GLM-based harmonization method, the site difference is regarded as the covariates to be regressed out. For ComBat harmonization method, firstly, ComBat normalizes the data by removing the effect of the overall mean and signal variables. Then, using an empirical Bayes framework, ComBat estimates additive and multiplicative site effects. The final harmonized data could be obtained by removing these site effects and adding the signal-related information back. In our study, the performance of GLM was only evaluated with simulation data. The main reason is that ComBat is a GLM-derived model and is more powerful than the original GLM model. Thus, for real MRI data, only the performance of ICA model and ComBat model were compared.

**2.4.3 Evaluating the harmonization results**

For simulated MRI data, ICA was utilized to the non-harmonized and harmonized data to extract and identify the signal- and site effects-related components to compare the harmonization effects of all the methods.

For the real MRI data, a set of analyses were applied to show the performance of site effects elimination and biological variability preservation (including age effects and group difference (ASD/HC)) for all the harmonization methods. For site effects removal evaluation, T-distributed stochastic neighbor embedding (t-SNE) was used to visualize the heterogeneity

related to sites of non-harmonized data and harmonized data by projecting their dominant features into a 2D space. We could assess the efficiency of the harmonization methods by checking whether there are site-clustered distributions or noticeable inter-site heterogeneities after site effects removal. Group F-test was also applied to both the non-harmonized data and the harmonized data derived from different methods to test the significant difference regions caused by the site difference. For age effects evaluation, the Pearson correlation coefficient between median GM and age of all the subjects from ABIDE II data was calculated to show their relationship. The median GM per subject was obtained by calculating the median of 100 regions of interest (using a standard brain template, FSLMNI152_1mm, in FSL parcelled by Schaefer et al., 2018). One-way ANOVA analysis was also applied to both the non-harmonized data and the harmonized data derived from different methods to test the significant difference regions caused by age. For group difference (ASD/HC) evaluation, group t-test was applied to both the non-harmonized data and the harmonized data derived from different methods to test the significant difference regions caused by the group difference. FWE-corrected $p < 0.05$ using non-parametric permutation testing with threshold-free cluster enhancement (TFCE) (Smith & Nichols, 2009) in FSL's Randomise (Winkler et al., 2014), with 5,000 permutations was used to find the significant regions.

# 3 Results

## 3.1 Simulation Harmonization Results

For simulated data, the data were decomposed into 10 components based on the simulation. Fig. 3 shows the signal- and site-effects-related components of simulated data extracted by ICA, before and after harmonization, when the signal variable is not significantly correlated to the site-effects variable. The results are shown in Fig. 3(a) when the spatial maps of all 10 components are spatially independent. Fig. 3(b) shows the results when the first two components are spatially overlapped. When the signal variable is not correlated to the site effects variable, the results for spatially independent and spatial dependent data are similar. All the harmonization methods could remove pure site effects component #3 and preserve pure signal component #4. However, the site effects cannot be removed from the mixed components #1 (more related to signal for the original data) and #2 (more related to site effects for the original data) by ICA-SP method. The performance of ICA-DP, GLM and ComBat were comparable under this situation, the site effects in the mixed components #1 and #2 were effectively removed and the signal effect was enhanced by increasing its correlation levels with the signal variable after ICA-DP, GLM and ComBat harmonization. The two mixed components were combined into one component that is only significantly related with signal variables. The site effects-related regions were also removed after harmonization with ICA-DP, GLM, and ComBat methods.

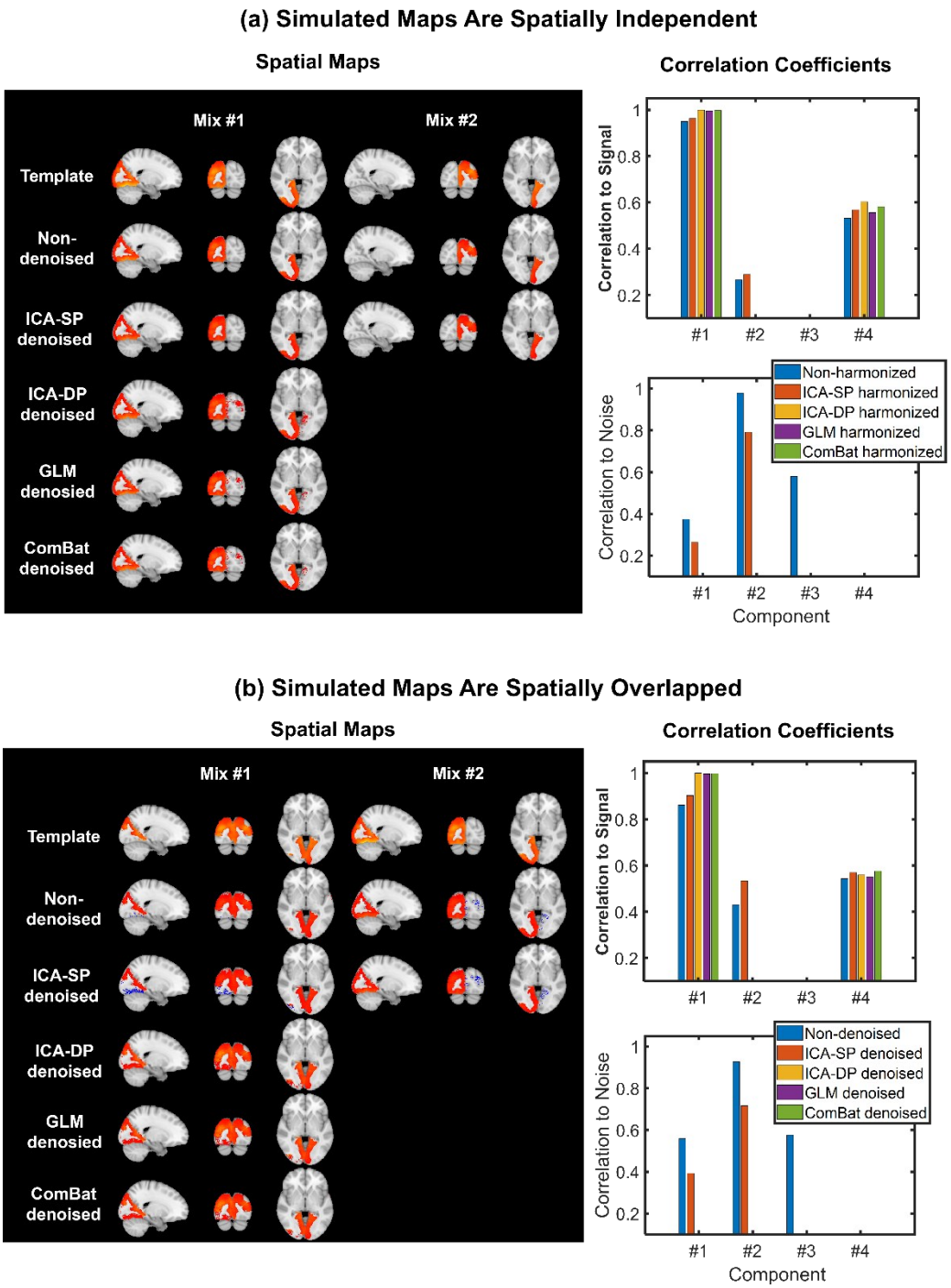


Fig. 3. Harmonization effects on the signal- and site-effects-related components when the signal variable is not significantly correlated to the site-effects variable. (a) Results when the spatial maps of all 10 components were spatially independent. (b) Results when the spatial maps of the first two components were spatially overlapped. For the non-harmonized data, components #1 and #2 are two mixed components, named Mix #1 (more related to the signal variable) and Mix #2 (more related to the site effects variable), component #3 is pure site effects component and component #4 is a pure signal component. The pure site effects component was removed by all the harmonization methods. The two mixed components were combined into one component that is only significantly related with signal variable. The site effects-related regions were also removed after harmonization with ICA-DP, GLM, and ComBat methods. ICA-SP cannot harmonize the site effects from the two mixed components.

Fig. 4 shows the harmonization effects on the two mixed components when the signal variable significantly correlates to the site effects variable. Three different correlation levels between the signal and site effects variables were simulated. Fig. 4(a) shows the results when the spatial maps of all 10 components were spatially independent. Fig. 4(b) shows the results when the spatial maps of the first two components were spatially overlapped. Among all the harmonization methods, only ICA-DP could effectively weaken the site effects while strengthening the signal effects when signal and site effects variables are correlated for two types of simulated data.

ICA-SP harmonization could not remove the site effects from the mixed components, as both the extracted spatial maps and correlation coefficients between subject loadings and site effects variable of the mixed components were not changed after being harmonized by ICA-SP, compared to that of non-harmonized data.

GLM harmonization showed the most aggressive harmonization performance while eliminating the site effects-related information at the expense of destroying the signal-related information. After being harmonized by GLM, both components #1 and #2 were not correlated to site effects variable. Besides, component #2 was not correlated to the signal variable any longer and the correlation between component #1 and signal variable also became lower, which became more severe with the increasing correlation levels between signal and site effects variables.

ComBat harmonization could not remove the site effects when the mixed component is more related with signal effects (component #1) and showed aggressive harmonization performance that also removes signal effects when the mixed component was more correlated with site effects variable (component #2). When all the 10 components are spatially independent, after ComBat harmonization, both spatial maps and subject loadings of the mixed component #1 were not changed, in contrast, the mixed component #2 was not related with both signal and site effects variables any longer. When the two mixed components are spatially overlapped, the site effects could not be effectively removed by ComBat when the mixed component was more related to signal variable (component #1) and showed aggressive harmonization performance that removed some signal effects when the mixed component was more related with site effects (component #2). Both spatial maps and subject loadings of the mixed component #1 (more related to signal) did not change significantly, in contrast the mixed component #2 (more related to site effects) was less correlated to both signal and site effects variables.

After being harmonized by ICA-DP, the original mixed components #1 and #2 were merged into a single component that was more related to signal variable and the site-related effects were effectively decreased. Some spatial areas related to site effects were also removed (highlighted with white circles), especially for lower correlation between signal and site effects variables. Fig. 4(b) shows that the removed spatial parts only cover the unique parts of component #2 and do not involve the overlapping parts. Though the merged component after being harmonized by ICA-DP was still mixed component, the correlation coefficient between its loading and

signal variable was strengthened for all the signal to site effects correlation levels, and the correlation levels to site effects variable were contributed by the inherent relationship between signal and site effects variables. Thus, there is no site-specific effect in the mixed component after being harmonized by ICA-DP. ICA-DP showed the most powerful harmonization performance, which could remove all the site-specific effects and enhance the signal effects.


Fig. 4. Harmonization effects on the two mixed components when the signal variable is significantly correlated to the site effects variable. Mix #1 is more related to the signal variable, and Mix #2 is more related to the site effects variable for the non-harmonized data. Three different correlation levels between signal and site effects variables

were simulated in this study to test the harmonization performance of all the harmonization methods. (a) When the spatial maps of all 10 components were spatially independent. (b) Results when the spatial maps of the first two components were spatially overlapped. After being harmonized by ICA-DP, Mix #1 and #2 were merged into a single component more related to signal variable, and the site-related effects were effectively decreased. Some spatial areas related to site effects were also removed (highlighted with a white circle), especially for lower correlation between signal and site effects variables. Among all the harmonization methods, only ICA-DP could effectively weaken the site effects while strengthening the signal effects when signal and site effects variables are correlated.

### 3.2 Real Datasets Harmonization Results

After harmonization, we performed a set of analyses to show the elimination of site effects and the preservation of biological variability, i.e., HC/ASD and age for the ABIDE II dataset, and subject heterogeneity for the traveling subject dataset. For the data from ABIDE II, they were decomposed into 50, 100 and 150 independent components, by calculating the correlation levels of subjects' loadings and variables with the analysis of variance (ANOVA) for each component, we identified the numbers of pure site effects components were 27, 56 and 96, respectively, and the numbers of mixed components were 23, 42 and 49, respectively. The traveling-subject data were decomposed into 50 independent components, and 10 pure site effects components and 16 mixed components were identified.

Fig. 5 shows the tSNE-2D projection of ADIDE II and traveling subject datasets before and after harmonization. The t-SNE was utilized to project the data into two dimensions by using the two dominant features of the non-harmonized data and harmonized data, to visualize the distribution of site effects and indicate whether it could be eliminated after harmonization. For the ABIDE II dataset, the data points of the non-harmonized data showed site-clustered distribution as most of the centers had their own specific cluster area, except for some intersections among centers UCLA, OHSU, and ETH. And the site-clustered distribution disappeared after being harmonized by any of the harmonization methods. For the traveling subject dataset, the projected data points of the non-harmonized data from the same subject tend to be clustered into one cluster, i.e., the first two projected features were dominated by the subject heterogeneity rather than site effects. Though significant difference was not found before and after harmonization, the subject heterogeneity was well preserved after harmonization.

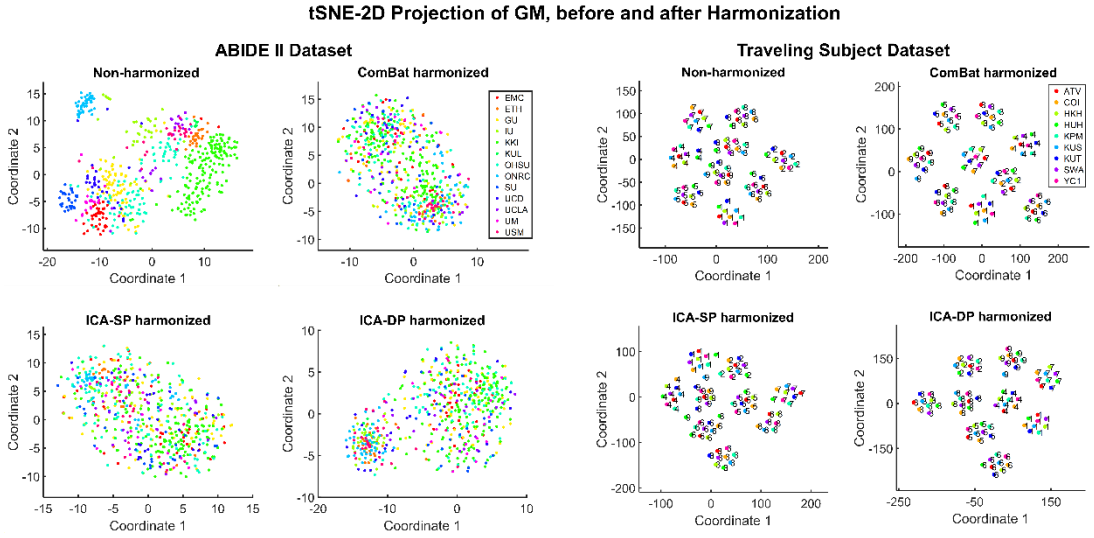


Fig. 5. Dimension reduction visualization by t-SNE before and after harmonization for ABIDE II and traveling subject datasets. The site-cluster distribution of the ABIDE II dataset before harmonization indicated the site effects, and it decreased when the data points were randomly distributed after harmonization. For the traveling subject dataset, the subject-cluster distribution indicated the dominance of subject heterogeneity, as the subjects from this dataset are the same ones scanned at different centers (the data points were labeled by subject numbers). There was no significant difference before and after harmonization, and subject heterogeneity was well preserved after harmonization.

In Fig. 6(a), diagnostic plots were presented for all the subjects from the two datasets, and the different colors represent different sites. For each subject, the GM measurements were summarized into a boxplot. The different range of GM values among sites was reduced after harmonization, and the ICA-based harmonization showed efficient reduction. Fig. 6(b) shows the median GM values distribution of the subjects from different sites. The standard deviation values for the medians of Median GM were calculated across subjects. After harmonization, the site effects decreased noticeably for all the harmonization methods.

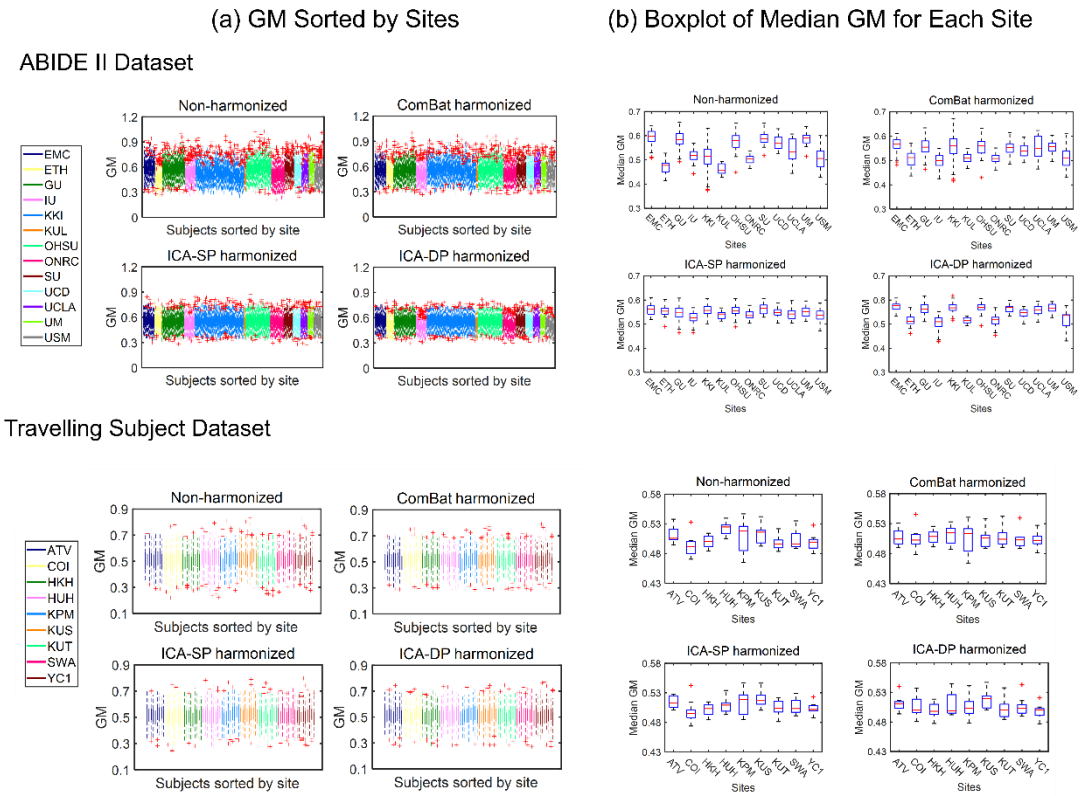


Fig. 6. (a) Site-sorted boxplots of GM. Each boxplot represents the GM values distribution of 100 regions of interest (ROI) for every subject. The ranges indicated differences among sites and among subjects. (b) Site-sorted boxplots of median GM. Each boxplot represents the distribution of median GM values for all subjects from the same site. The fluctuates indicated the inter-site difference. The standard deviation values for the medians of Median GM across subjects, before and after harmonization, are 1)ABIDE II Dataset: 0.0472 (Non-harmonized), 0.0250 (ComBat harmonized), 0.0209 (ICA-SP harmonized), 0.0251 (ICA-DP harmonized); 2)Traveling Subject Dataset: 0.0115 (Non-harmonized), 0.0046 (ComBat harmonized), 0.0078 (ICA-SP harmonized), 0.0071 (ICA-DP harmonized).

The boxplots presented in Fig. 7, for non-harmonization data and harmonization data, summarized the distribution of the median GM for each subject, revealing heterogeneity among different subjects. The subject heterogeneity was not destroyed by all the harmonization methods and preserved well after being harmonized by ICA-DP, and the trends of median GM for each subject, before and after harmonization, are shown in Fig. 7(b). Besides, the intra-subject difference (represented by the height of each box in Fig. 7 (a) and the standard variation of median GM for each subject in Fig. 7 (c)), as a representation of site effects, had been most significantly reduced after ComBat and ICA-DP harmonization (Fig. 7(c)).

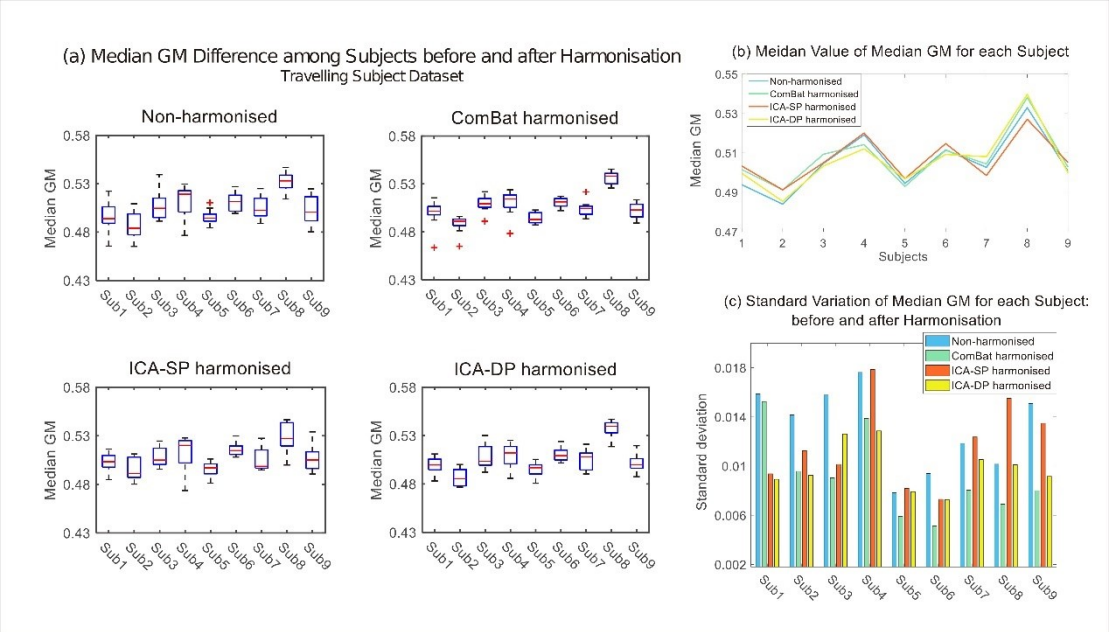


Fig. 7. (a) Subject-sorted boxplots of median GM. The fluctuates indicated the inter-subject difference and the height of each box indicated the intra-subject (inter-site) difference. The standard deviation values for the medians of Median GM across subjects, before and after harmonization, are 0.0147 (Non-harmonized), 0.0139 (ComBat harmonized), 0.0116 (ICA-SP harmonized), and 0.0149 (ICA-DP harmonized). (b) The median values of Meidan GM for each subject, before and after harmonization. The trends show the difference among subjects. (c) The standard deviation value of median GM for each subject, before and after harmonized.

Fig. 8 shows the group-level analysis for site effects from the two datasets. The non-harmonized GM data was globally affected by the site effects for both datasets. Although the site effects had been alleviated by the ICA-SP method, it could not remove them sufficiently. After being harmonized by ICA-DP and ComBat, no significant regions were associated with site variable for both datasets (FWE-corrected $p < 0.05$).

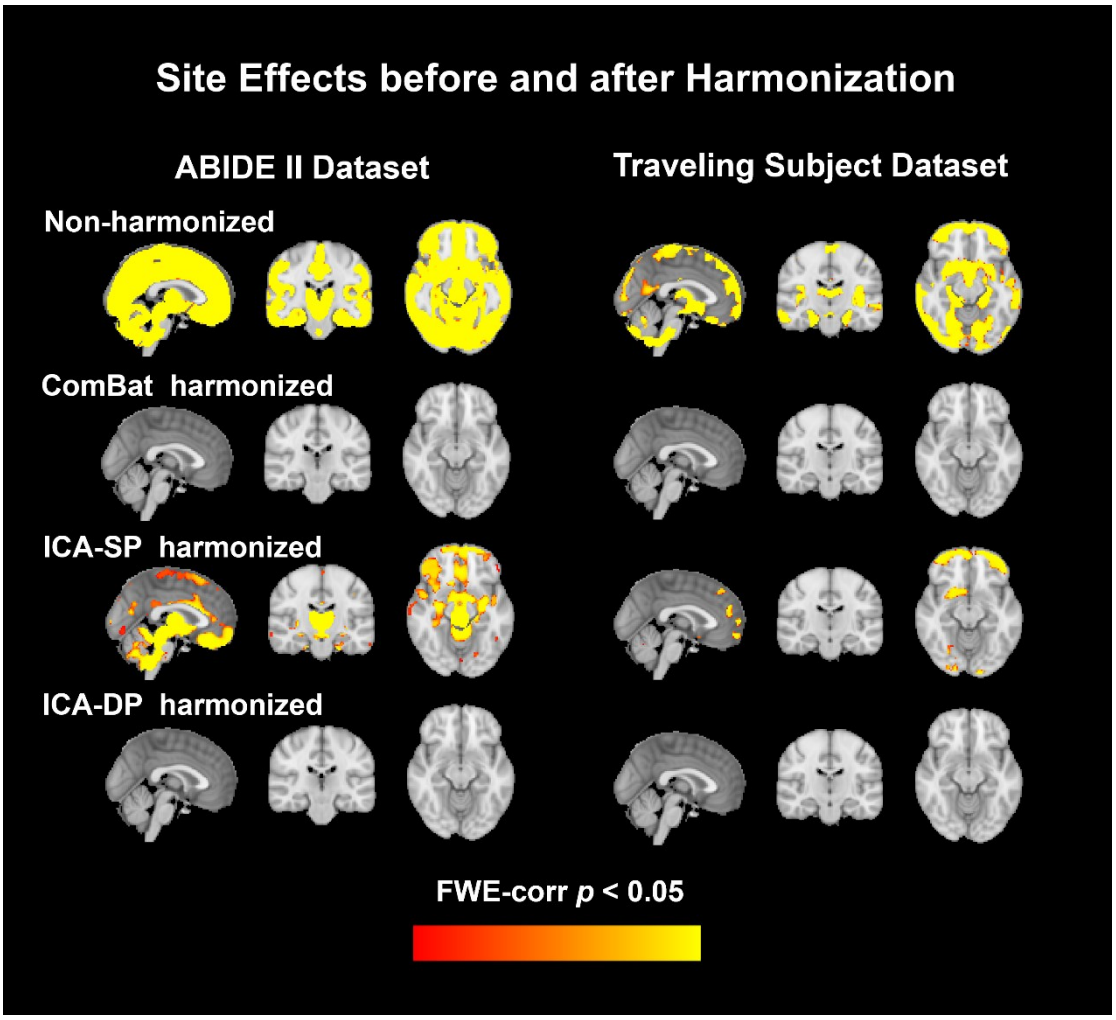


Fig. 8. Group-level analysis for site effects before and after harmonization. Site effects were removed completely by Combat and ICA-DP. Though ICA-SP reduced the site effects, some significant regions still could be found.

As a biological variable of interest, age effects of the ABIDE II dataset, before and after harmonization, were shown in Fig. 9. The correlation between age and median GM before and after harmonization with the ABIDE dataset was shown in Fig. 9(a). The median GM were sorted by age and the data from different scanning centers were in different colors. Pearson correlation coefficients between age and median GM from non-harmonized data and harmonized data were calculated, which were -0.4746 (Non- harmonized), -0.5689 (ComBat harmonized), -0.3617 (ICA-SP harmonized), -0.8493 (ICA-DP harmonized), respectively. The correlation coefficients indicated that the negative correlation between GM and age was strengthened by ICA-DP and ComBat, especially for ICA-DP. Fig. 9(b) shows the group-level analyses for age. Site effects confound us to find the true age effects. The negative age effects were not found in the non-harmonized data because of the existence of site effects, removal of the effects by all the harmonization methods, especially for ICA-DP, could reveal the negative age effects that are not detected from the non-harmonized data.

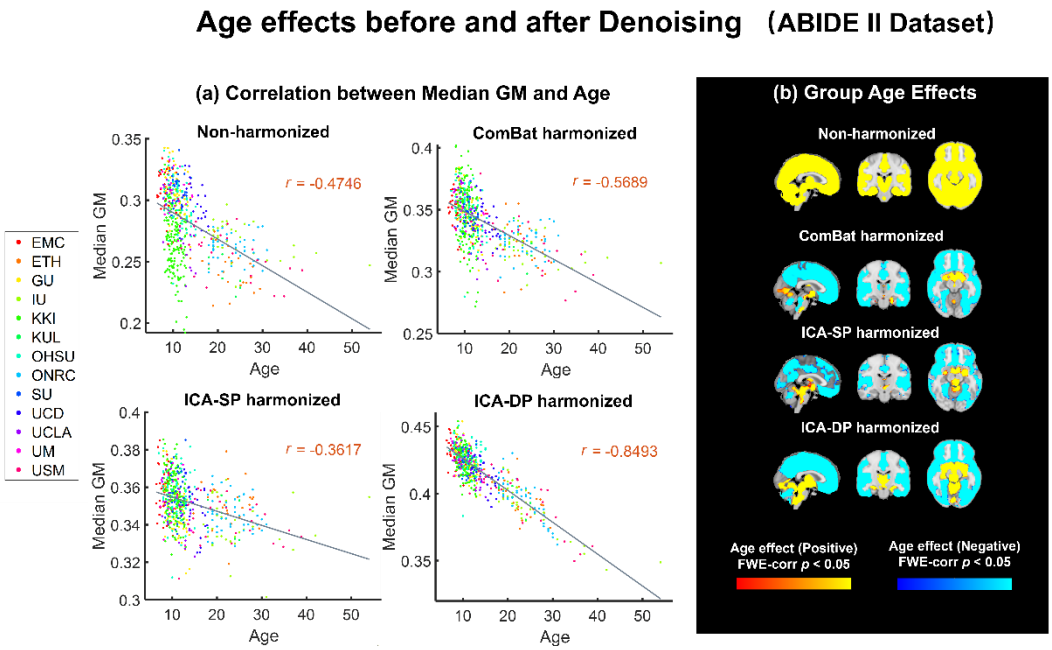


Fig. 9. (a) Relationship between age and median GM before and after site effects harmonization. The Pearson correlation coefficient was -0.4746 (Non-harmonized), -0.5689 (ComBat harmonized), -0.3617 (ICA-SP harmonized), and -0.8493 (ICA-DP harmonized). (b) Group-level analysis of GM maps for age effects before and after data harmonization. The negative age effects (significance level) were enhanced after harmonization, and they could not be detected in the non-harmonized data when testing age group differences.

Fig. 10 shows the group difference (ASD/HC) before and after harmonization. ICA-DP increased the group effects by detecting more significantly different regions related to ASD and HC, while ComBat and ICA-SP decreased the group effects as no significant regions were tested from the data harmonized by them. Compared to the non-thresholded group difference maps from non-harmonized data (first row), the regions associated with group difference (ASD/HC) from ICA-DP-based harmonized data could also be found in the non-harmonized data. In other words, ICA-DP only strengthened the signal that should be there rather than reintroducing artifacts.

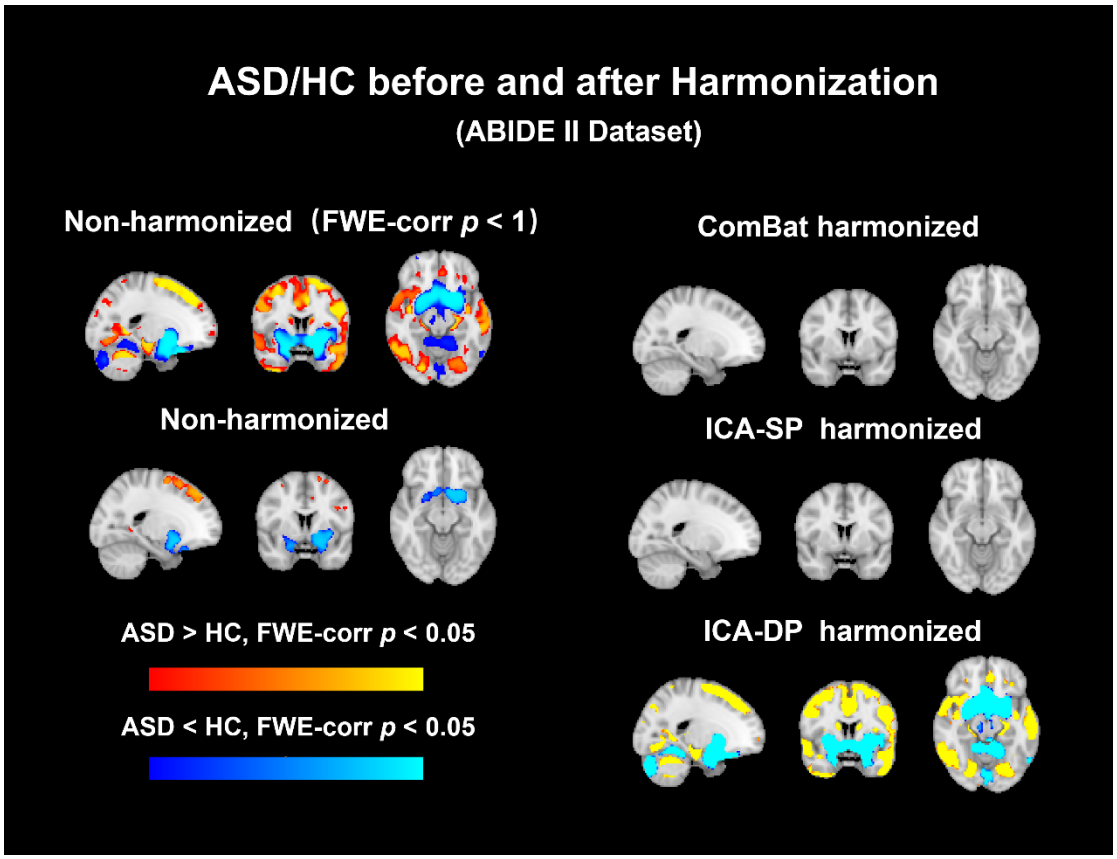


Fig. 10. Group-level analysis of GM maps for group difference (ASD/HC) before and after data harmonization. No significant regions were detected from the data harmonized by ComBat and ICA-SP, while ICA-DP could increase the significance of the regions related to ASD/HC. FWE-corr $p < 1$ was shown for non-harmonized data to indicate that the regions tested from ICA-DP harmonized data were not reintroduced artifacts.

Fig. 11 shows the harmonization performance of the two ICA-based harmonization methods under different choices of component numbers when running ICA algorithms. The ICA-SP could not remove the site effect completely, though it decreased more site effects as the number of components increased, and the information related to signal variable (ASD/HC) could not be detected from the data harmonized by it under any component number choosing, indicating that this kind of soft harmonization based on ICA could neither remove the site effects completely nor reveal the information related to covariates of interest. In contrast, after being harmonized by ICA-DP, the information that related to site effects could not be tested and there were some regions that significantly correlated to ASD/HC could be revealed from the harmonized data, indicating good performance and importance for eliminating site effects and the ability to unveil the signal related information concurrently and showing no affection of which number of components were chosen.

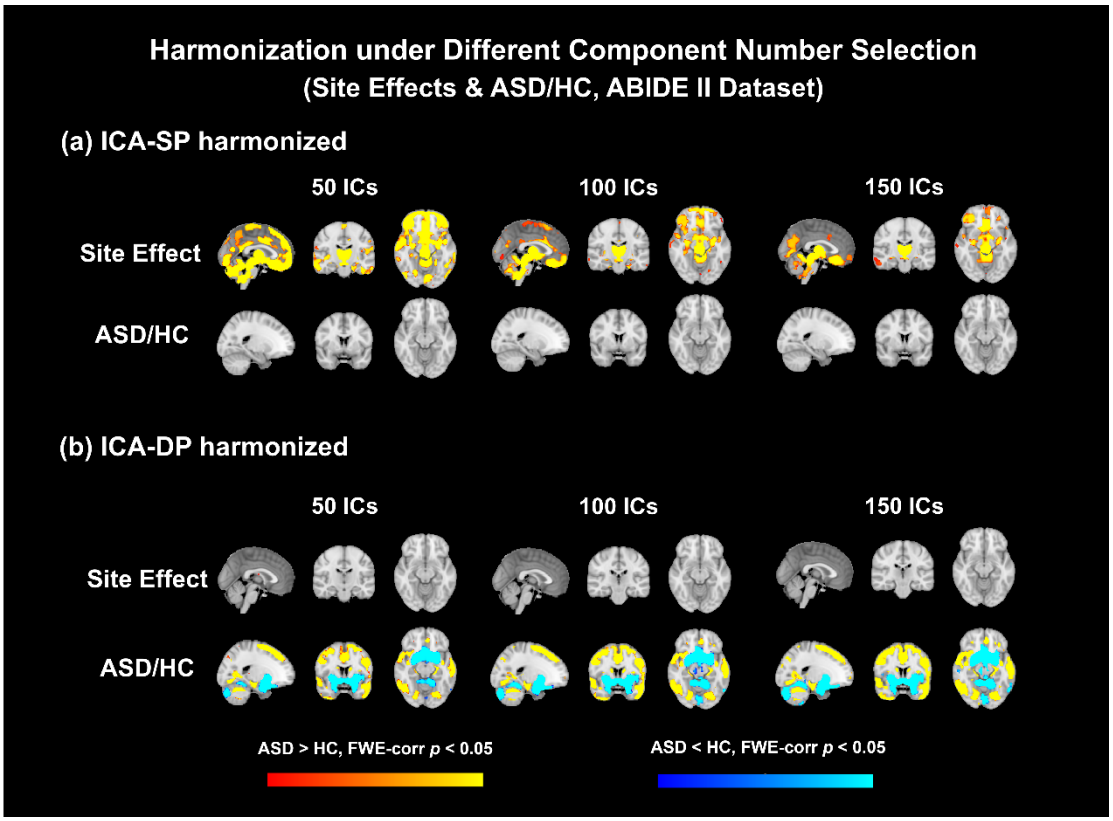


Fig. 11. Harmonization performance of the two ICA-based harmonization methods under different component numbers (50, 100, and 150). After being harmonized by ICA-DP, site effects were removed completely and ASD/HC group differences were significantly enhanced (the regions related to ASD/HC could not be detected before harmonization) without the affection of component number.

## 4 Discussion

In this paper, we proposed a dual-projection ICA-based harmonization method that can remove the site effects more effectively and completely while enhancing the signal effects. This method shows superior performance when the site effects are also related with signal variables. The Matlab codes of ICA-DP harmonization are available at https://github.com/Yuxing-Hao/ICA-DP_Harmonization.git.

ICA-DP was developed to clean the site effects more completely and effectively when the site and signal effects are correlated. Based on the simulation results (Figs. 3, 4), it was found that ICA-DP can eliminate the site effects effectively while enhancing the signal-related information, whether the site and signal variables are correlated. The benefits of ICA-DP are contributed by two reasons: first, ICA-DP can extract all the site-related effects with the first projection even though some site effects are mixed with the signal effects that could not be extracted from original ICA method; second, ICA-DP is more stable than original ICA method, as it can extract and remove all the site effects without the limitation of model order selection; third, compared to standard GLM and ComBat model whose regressors have the same value for all participants in the same sites, site-related effects are better captured by ICA-DP, which can capture day-to-day variations in scanner performance. Compared with ICA-DP, ICA-SP,

GLM and ComBat showed different defects when harmonization the site effects. For ICA-SP, it encountered a problem that it failed to eliminate the site effects completely when the site-related components also significantly correlated to signal variables (Figs. 3, 4). The GLM-based harmonization method performed well and was comparable to ICA-DP when the signal and site effects variables are not significantly correlated (Fig. 3(a)). However, GLM showed aggressive harmonization performance that will remove or decrease the signal-related information when the signal and site effects variables are significantly correlated. The signal distortion became worse when the correlation coefficients between the signal and site effects variables increased (Fig. 4). Though ComBat-based harmonization method showed better performance than GLM and original ICA-based harmonization performance which can effectively remove site effects while keeping or strengthening some signal effects when signal and site variables are not correlated. However, ComBat could not completely remove the site effects when signal and site variables are correlated when the mixed components are more correlated with the signal (Fig. 4). Meanwhile, we found that ComBat showed aggressive harmonization that also removed signal effects when the mixed components are more related with site effects.

Based on the results of the ABIDE II dataset and the traveling-subject dataset, ICA-DP also shows superior performance on harmonization site effects than ICA-SP and ComBat. The significant regions of GM that related to site effects, which cannot be completely removed by ICA-SP, were not detectable after harmonizing with ICA-DP and ComBat (Fig. 8). The inter-site variation of the traveling data and the intra-subject variation among nine sites for the traveling subject dataset were most significantly reduced after harmonization with ICA-DP (Fig. 6). Moreover, subject heterogeneity of the traveling subject dataset was also well preserved after being harmonized by ICA-DP (Figs. 5, 7).

In addition, ICA-DP also shows superior performance in enhancing biological variability (i.e., age effects and group difference between ASD and HC) compared with ICA-SP and ComBat based on the results of the ABIDE II dataset. Site effects hinder us from finding the true age and group effects. The age effects detected with GM of the non-harmonized ABIDE II are opposite with the recognized results (Gennatas et al., 2017; Groves et al., 2012). After being harmonized by ICA-SP, ICA-DP, and ComBat, the true age effects on GM were discovered. Among all the three methods, ICA-DP finds the most significant regions related to age and the negative correlation between age and GM was most strongly enhanced by ICA-DP. The relationship of median GM and age was enhanced from -0.4746 (non-harmonized) to -0.8493 after being harmonized by ICA-DP (Fig. 9). In addition, compared to ICA-SP and ComBat, only ICA-DP enhanced the group effects (ASD vs. HC) by detecting more significantly group different regions which cannot be effectively detected by the non-harmonized data (Fig. 10), indicating the importance of removing site effects in multi-site data. The notable enhancements of the biological variabilities (i.e., age effects and ASD vs. HC) may be attributed to the larger proportion of site-related components that we selected for ICA-DP harmonization, which could increase the weights of signal we are interested in and make the signal-related information

easier to be detected. On the other hand, this may lead to the other variables that we are not interested in not being well preserved. To protect other variables that we may be interested in, we just need to add these variables to $\mathbf{V}_S$ in the first projection of the ICA-DP harmonization method (Eq. (1)). Thus, the ICA-DP is the most effective method for harmonizing site effects and preserving biological variability among the methods discussed above. Moreover, unlike ICA-SP, the performance of ICA-DP in site effects harmonization and signal enhancement was not affected by the number of components chosen for ICA decomposition. It could clean the site effects and strengthen the signal under any selected component number (Fig. 11).

Finally, as a limitation of the proposed harmonization method, when the site effects variable is strongly related to the signal variable (Fig. 4), ICA-DP could not eliminate the intersection effects that are related with both site and signal variables (neither do other methods except GLM, the most aggressive one destroying signal-related information severely), thus the harmonized data are still correlated to site effects because of the inherent correlation between site effects and signal variables (Nevertheless, the correlation values in our simulation is really high and hardly appear in real data study). Another limitation to consider is the selection of the number of components to be extracted by ICA. Though we validated that the performance of ICA-DP in site effects harmonization and signal enhancement were not affected by the number of components chosen for ICA decomposition under three different selections for the number of components (see Figure. 11), it is not sufficient and verification methods should be developed further. However, to some extent, ICA-DP allows users to choose the number of components according to their own standard.

Overall, the dual-projection harmonization method is more effective and powerful in removing site effects while preserving signal-related information than other methods mentioned above, and can enhance the sensitivity to detect signals of interest and remove all the effects that are only contributed by site difference. Compared to Combat, it is a data-driven method rather than utilizing the manually designed covariates for regressing. ICA-DP harmonization method has great potential for large-scale multi-site studies to produce combined data free from study-site confounds.

## 5 Conclusion

While combing the multi-site MRI data has great convenience that enhances the statistical results and obviates some of the shortcomings of the single-site study, the site effects come naturally, confounding the MRI data analysis and making the results hard to interpret. The traditional methods designed to eliminate the site effects encounter incomplete or aggressive harmonization, i.e., cannot eliminate the site effects well or may destroy the signal-related information. To tackle these shortcomings, we proposed a dual-projection data-driven method based on ICA, which can better eliminate the site effects and preserve the signals of interest. And we strongly recommend that researchers use the ICA-DP method to harmonize the MRI

data as it can extract subject-specific loadings that correspond to the signal or site effects variable.

**Abbreviations:** DP, dual-projection; DTI, diffusion tensor image; GLM, general linear model; GM, gray matter; ICA, independent component analysis; ICA-DP, ICA-dual projection; ICA-SP, ICA-single projection; LICA, linked ICA; MRI, magnetic resonance imaging; PET, positron emission tomography; t-SNE, t-distributed stochastic neighbor embedding.

## References


Andersson, J. L., Jenkinson, M., & Smith, S. (2007). Non-linear registration, aka Spatial normalisation FMRIB technical report TR07JA2. *FMRIB Analysis Group of the University of Oxford*, *2*(1), e21.

Beckmann, C. F., Mackay, C. E., Filippini, N., & Smith, S. M. (2009). Group comparison of resting-state FMRI data using multi-subject ICA and dual regression. *NeuroImage*, *47*(Supp1 1), S148.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*(6), 1129–1159. https://doi.org/10.1162/neco.1995.7.6.1129

Bell, T. K., Godfrey, K. J., Ware, A. L., Yeates, K. O., & Harris, A. D. (2022). Harmonization of multi-site MRS data with ComBat. *NeuroImage*, *257*, 119330. https://doi.org/10.1016/j.neuroimage.2022.119330

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.

C Monte-Rubio, G., Segura, B., P Strafella, A., van Eimeren, T., Ibarretxe-Bilbao, N., Diez-Cirarda, M., Eggers, C., Lucas-Jimenez, O., Ojeda, N., & Pena, J. (2022). Parameters from site classification to harmonize MRI clinical studies: Application to a multi-site Parkinson's disease dataset. *Human Brain Mapping*, *43*(10), 3130–3142.

Casey, B. J., Cohen, J. D., O'Craven, K., Davidson, R. J., Irwin, W., Nelson, C. A., Noll, D. C., Hu, X., Lowe, M. J., Rosen, B. R., Truwitt, C. L., & Turski, P. A. (1998). Reproducibility of fMRI results across four institutions using a spatial working memory task. *NeuroImage*, *8*(3), 249–261. https://doi.org/10.1006/nimg.1998.0360

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

Di Martino, A., O'connor, D., Chen, B., Alaerts, K., Anderson, J. S., Assaf, M., Balsters, J. H., Baxter, L., Beggiato, A., & Bernaerts, S. (2017). Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Scientific Data*, *4*(1), 1–15.

Dinsdale, N. K., Jenkinson, M., & Namburete, A. I. L. (2021). Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. *NeuroImage*, *228*(December 2020), 117689. https://doi.org/10.1016/j.neuroimage.2020.117689

Douaud, G., Smith, S., Jenkinson, M., Behrens, T., Johansen-Berg, H., Vickers, J., James, S., Voets, N., Watkins, K., & Matthews, P. M. (2007). Anatomically related grey and white matter abnormalities in adolescent-onset schizophrenia. *Brain*, *130*(9), 2375–2386.

Eickhoff, S., Nichols, T. E., Van Horn, J. D., & Turner, J. A. (2016). Sharing the wealth: Neuroimaging data repositories. *NeuroImage*, *124*, 1065–1068. https://doi.org/10.1016/j.neuroimage.2015.10.079

Fennema-Notestine, C., Gamst, A. C., Quinn, B. T., Pacheco, J., Jernigan, T. L., Thal, L., Buckner, R., Killiany, R., Blacker, D., Dale, A. M., Fischl, B., Dickerson, B., & Gollub, R. L. (2007). Feasibility of multi-site clinical structural neuroimaging studies of aging using legacy data. *Neuroinformatics*, *5*(4), 235–245. https://doi.org/10.1007/s12021-007-9003-9

Filippini, N., MacIntosh, B. J., Hough, M. G., Goodwin, G. M., Frisoni, G. B., Smith, S. M., Matthews, P. M., Beckmann, C. F., & Mackay, C. E. (2009). Distinct patterns of brain activity in young carriers of the APOE-$\varepsilon$4 allele. *Proceedings of the National Academy of Sciences*, *106*(17), 7209–7214.

Focke, N. K., Helms, G., Kaspar, S., Diederich, C., Tóth, V., Dechent, P., Mohr, A., & Paulus, W. (2011). Multi-site voxel-based morphometry - Not quite there yet. *NeuroImage*, *56*(3), 1164–1170. https://doi.org/10.1016/j.neuroimage.2011.02.029

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*(July), 149–170. https://doi.org/10.1016/j.neuroimage.2017.08.047

Friedman, L., Stern, H., Brown, G. G., Mathalon, D. H., Turner, J., Glover, G. H., Gollub, R. L., Lauriello, J., Lim, K. O., Cannon, T., Greve, D. N., Bockholt, H. J., Belger, A., Mueller, B., Doty, M. J., He, J., Wells, W., Smyth, P., Pieper, S., … Potkin, S. G. (2008). Test-retest and between-site reliability in a multicenter fMRI study. *Human Brain Mapping*, *29*(8), 958–972. https://doi.org/10.1002/hbm.20440

Gennatas, E. D., Avants, B. B., Wolf, D. H., Satterthwaite, T. D., Ruparel, K., Ciric, R., Hakonarson, H., Gur, R. E., & Gur, R. C. (2017). Age-related effects and sex differences in gray matter density, volume, mass, and cortical thickness from childhood to young adulthood. *Journal of Neuroscience*, *37*(20), 5065–5073.

Glover, G. H., Mueller, B. A., Turner, J. A., Van Erp, T. G. M., Liu, T. T., Greve, D. N., Voyvodic, J. T., Rasmussen, J., Brown, G. G., Keator, D. B., Calhoun, V. D., Lee, H. J., Ford, J. M., Mathalon, D. H., Diaz, M., O'Leary, D. S., Gadde, S., Preda, A., Lim, K. O., … Potkin, S. G. (2012). Function biomedical informatics research network recommendations for prospective multicenter functional MRI studies. *Journal of Magnetic Resonance Imaging*, *36*(1), 39–54. https://doi.org/10.1002/jmri.23572

Good, C. D., Johnsrude, I. S., Ashburner, J., Henson, R. N., Friston, K. J., & Frackowiak, R. S. (2001). A voxel-based morphometric study of ageing in 465 normal adult human brains. *Neuroimage*, *14*(1), 21–36.

Groves, A. R., Beckmann, C. F., Smith, S. M., & Woolrich, M. W. (2011). Linked independent component analysis for multi-modal data fusion. *Neuroimage*, *54*(3), 2198–2217.

Groves, A. R., Smith, S. M., Fjell, A. M., Tamnes, C. K., Walhovd, K. B., Douaud, G., Woolrich, M. W., & Westlye, L. T. (2012). Benefits of multi-modal fusion analysis on a large-scale dataset: life-span patterns of inter-subject variability in cortical morphometry and white

matter microstructure. *Neuroimage*, *63*(1), 365–380.

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

Jovicich, J., Czanner, S., Han, X., Salat, D., Kouwe, A. Van Der, Quinn, B., Pacheco, J., Albert, M., Killiany, R., Blacker, D., Rosas, D., Makris, N., Gollub, R., & Dale, A. (2009). MRI-derived measurements of human subcortical, ventricular andintracranial brain volumes: Reliability effects of scan sessions,acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. *NeuroImage*, *46*(1), 177–192. https://doi.org/10.1016/j.neuroimage.2009.02.010.MRI-derived

Li, H., Smith, S. M., Gruber, S., Lukas, S. E., Silveri, M. M., Hill, K. P., Killgore, W. D., & Nickerson, L. D. (2020). Denoising scanner effects from multi-modal MRI data using linked independent component analysis. *Neuroimage*, *208*, 116388.

Maikusa, N., Zhu, Y., Uematsu, A., Yamashita, A., Saotome, K., Okada, N., Kasai, K., Okanoya, K., Yamashita, O., & Tanaka, S. C. (2021). Comparison of traveling-subject and ComBat harmonization methods for assessing structural brain characteristics. *Human Brain Mapping*, *42*(16), 5278–5287.

Nickerson, L. D., Smith, S. M., Ongur, D., & Beckmann, C. F. (2017). Using dual regression to investigate network shape and amplitude in functional connectivity analyses. *Frontiers in Neuroscience*, *11*, 115.

Orlhac, F., Boughdad, S., Philippe, C., Stalla-Bourdillon, H., Nioche, C., Champion, L., Soussan, M., Frouin, F., Frouin, V., & Buvat, I. (2018). A postreconstruction harmonization method for multicenter radiomic studies in PET. *Journal of Nuclear Medicine*, *59*(8), 1321–1328. https://doi.org/10.2967/jnumed.117.199935

Pohl, K. M., Sullivan, E. V, Rohlfing, T., Chu, W., Kwon, D., Nichols, B. N., Zhang, Y., Brown, S. A., Tapert, S. F., Cummins, K., Thompson, W. K., Brumback, T., Colrain, I. M., Baker, F. C., Prouty, D., De Bellis, M. D., Voyvodic, J. T., Clark, D. B., Schirda, C., … Pfefferbaum, A. (2016). Harmonizing DTI measurements across scanners to examine the development of white matter microstructure in 803 adolescents of the NCANDA study. *NeuroImage*, *130*, 194–213. https://doi.org/10.1016/j.neuroimage.2016.01.061

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, *28*(9), 3095–3114.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., & Flitney, D. E. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, *23*, S208--S219.

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*, *44*(1), 83–98.

Takao, H., Hayashi, N., & Ohtomo, K. (2011). Effect of scanner in longitudinal studies of brain volume changes. *Journal of Magnetic Resonance Imaging*, *34*(2), 438–444. https://doi.org/10.1002/jmri.22636

Tanaka, S. C., Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N.,

Takamura, M., Yoshihara, Y., & Kunimatsu, A. (2021). A multi-site, multi-disorder resting-state magnetic resonance image database. *Scientific Data*, *8*(1), 1–15.

Tian, D., Zeng, Z., Sun, X., Tong, Q., Li, H., He, H., Gao, J. H., He, Y., & Xia, M. (2022). A deep learning-based multisite neuroimage harmonization framework established with a traveling-subject dataset. *NeuroImage*, *257*(March), 119297. https://doi.org/10.1016/j.neuroimage.2022.119297

Van Horn, J. D., & Toga, A. W. (2009). Multisite neuroimaging trials. *Current Opinion in Neurology*, *22*(4), 370–378. https://doi.org/10.1097/WCO.0b013e32832d92de

Venkatraman, V. K., Gonzalez, C. E., Landman, B., Goh, J., Reiter, D. A., An, Y., & Resnick, S. M. (2015). Region of interest correction factors improve reliability of diffusion imaging measures within and across scanners and field strengths. *Neuroimage*, *119*, 406–416.

Vollmar, C., O'Muircheartaigh, J., Barker, G. J., Symms, M. R., Thompson, P., Kumari, V., Duncan, J. S., Richardson, M. P., & Koepp, M. J. (2010). Identical, but not the same: Intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0T scanners. *NeuroImage*, *51*(4), 1384–1394. https://doi.org/10.1016/j.neuroimage.2010.03.046

Wegner, C., Filippi, M., Korteweg, T., Beckmann, C., Ciccarelli, O., De Stefano, N., Enzinger, C., Fazekas, F., Agosta, F., Gass, A., Hirsch, J., Johansen-Berg, H., Kappos, L., Barkhof, F., Polman, C., Mancini, L., Manfredonia, F., Marino, S., Miller, D. H., … Matthews, P. M. (2008). Relating functional changes during hand movement to clinical parameters in patients with multiple sclerosis in a multi-centre fMRI study. *European Journal of Neurology*, *15*(2), 113–122. https://doi.org/10.1111/j.1468-1331.2007.02027.x

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. *Neuroimage*, *92*, 381–397.

Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., Kunimatsu, A., Okada, N., Yamagata, H., Matsuo, K., Hashimoto, R., Okada, G., Sakai, Y., Morimoto, J., Narumoto, J., Shimada, Y., Kasai, K., … Imamizu, H. (2019). Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. In *PLoS Biology*. https://doi.org/10.1101/440875

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/10.1002/hbm.24241

Zivadinov, R., & Cox, J. L. (2008). Is functional MRI feasible for multi-center studies on multiple sclerosis? *European Journal of Neurology*, *15*(2), 109–110. https://doi.org/10.1111/j.1468-1331.2007.02030.x

# II

# HARMONIZATION OF MULTI-SITE FUNCTIONAL MRI DATA WITH DUAL-PROJECTION BASED ICA MODEL

by

Xu, Huashuai, Yuxing Hao, Yunge Zhang, Dongyue Zhou, Tommi Kärkkäinen, Lisa D. Nickerson, Huanjie Li, and Fengyu Cong, 2023

Frontiers in Neuroscience, 17, 1225606

https://doi.org/10.3389/fnins.2023.1225606

frontiers | Frontiers in Neuroscience

TYPE Methods
PUBLISHED 20 July 2023
DOI 10.3389/fnins.2023.1225606



OPEN ACCESS

EDITED BY
Kamen Atanasov Tsvetanov,
University of Cambridge, United Kingdom

REVIEWED BY
Guoyuan Yang,
Beijing Institute of Technology, China
Qihong Zou,
Peking University, China

*CORRESPONDENCE
Lisa D. Nickerson
✉ lisa_nickerson@hms.harvard.edu
Huanjie Li
✉ hj_li@dlut.edu.cn

RECEIVED 19 May 2023
ACCEPTED 06 July 2023
PUBLISHED 20 July 2023

CITATION
Xu H, Hao Y, Zhang Y, Zhou D, Kärkkäinen T, Nickerson LD, Li H and Cong F (2023) Harmonization of multi-site functional MRI data with dual-projection based ICA model. *Front. Neurosci.* 17:1225606. doi: 10.3389/fnins.2023.1225606



# Harmonization of multi-site functional MRI data with dual-projection based ICA model

Huashuai Xu[1,2], Yuxing Hao[1], Yunge Zhang[1], Dongyue Zhou[1], Tommi Kärkkäinen[2], Lisa D. Nickerson[3,4]*, Huanjie Li[1]* and Fengyu Cong[1,2,5,6]

[1]School of Biomedical Engineering, Dalian University of Technology, Dalian, China, [2]Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland, [3]McLean Imaging Center, McLean Hospital, Belmont, MA, United States, [4]Department of Psychiatry, Harvard Medical School, Boston, MA, United States, [5]School of Artificial Intelligence, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, China, [6]Key Laboratory of Integrated Circuit and Biomedical Electronic System, Dalian University of Technology, Dalian, China

Modern neuroimaging studies frequently merge magnetic resonance imaging (MRI) data from multiple sites. A larger and more diverse group of participants can increase the statistical power, enhance the reliability and reproducibility of neuroimaging research, and obtain findings more representative of the general population. However, measurement biases caused by site differences in scanners represent a barrier when pooling data collected from different sites. The existence of site effects can mask biological effects and lead to spurious findings. We recently proposed a powerful denoising strategy that implements dual-projection (DP) theory based on ICA to remove site-related effects from pooled data, demonstrating the method for simulated and *in vivo* structural MRI data. This study investigates the use of our DP-based ICA denoising method for harmonizing functional MRI (fMRI) data collected from the Autism Brain Imaging Data Exchange II. After frequency-domain and regional homogeneity analyses, two modalities, including amplitude of low frequency fluctuation (ALFF) and regional homogeneity (ReHo), were used to validate our method. The results indicate that DP-based ICA denoising method removes unwanted site effects for both two fMRI modalities, with increases in the significance of the associations between non-imaging variables (age, sex, etc.) and fMRI measures. In conclusion, our DP method can be applied to fMRI data in multi-site studies, enabling more accurate and reliable neuroimaging research findings.

KEYWORDS

**multi-site, site effects, functional magnetic resonance imaging, independent component analysis, dual-projection**

## 1. Introduction

Functional magnetic resonance imaging (fMRI) has become a popular tool for understanding the human brain and detecting brain diseases since its inception in the 1990s (Eklund et al., 2016). Over the past three decades, countless methods and paradigms have been adopted to utilize and interpret fMRI data. One popular approach is frequency-domain analysis. Zang et al. proposed the amplitude of low frequency fluctuations (ALFF) for a voxel's time series, and it measures the total signal power in the low-frequency range is computed as the total power in the low frequency range (0.01–0.1 Hz; Zang et al., 2007). Another popular approach is based on

regional homogeneity analysis, or ReHo (Zang et al., 2004), which computes a voxel-based measure of brain activity that evaluates the synchronization between the time series of a given voxel and its nearest neighbors using Kendall's coefficient of concordance that is used to investigate the local coherence of fMRI signals in the brain (Yang et al., 2020).

Most neuroimaging studies are conducted within a single research site, with limited capabilities for collecting large sample-size datasets. Smaller sample sizes and lack of harmonization across independent studies present challenges in achieving acceptable reliability and reproducibility of neuroimaging research (Nichols et al., 2017). As a result, multi-site fMRI studies are becoming increasingly common to increase the power of statistical analyses to detect group differences, longitudinal changes, and, in turn the reliability and reproducibility of neuroimaging research. While combining multiple datasets across different studies is beneficial for the development of neuroscience, the existence of site effects makes pooling multi-site datasets challenging. Site effects can confound actual effects of interest and make the final results hard to interpret for fMRI data (Biswal et al., 2010; Groves et al., 2011; Li et al., 2020). Hence, effectively eliminating or minimizing the site effect is necessary for the fusion of multi-site fMRI data.

Recently, a new technique, independent component analysis (ICA) with dual-projection (ICA-DP), was proposed for removal of site effects in multi-site structural MRI data (Hao et al., 2023). For ICA-DP, mixed components are separated into a part related to signal only and a part related only to noise by applying a projection procedure. The noise effects extracted from the mixed components *via* the projection step and other pure noise components are then removed from the data using a second projection procedure. Compared with traditional ICA and ComBat (COMbining BATches), which is a general linear modal (GLM)-derived method based on the empirical Bayes approach (Johnson et al., 2007; Stein et al., 2015; Fortin et al., 2017, 2018; Beer et al., 2020; Cetin-Karayumak et al., 2020; Da-ano et al., 2020; Pinto et al., 2020; Cackowski et al., 2021; Eshaghzadeh Torbati et al., 2021; Maikusa et al., 2021; Bell et al., 2022; Orlhac et al., 2022), ICA-DP method demonstrates superior denoising while preserving the signals of interest.

In this paper, we introduce the use of ICA-DP to harmonize fMRI data collected from multiple sites. We apply ICA-DP to the data from Autism Brain Imaging Data Exchange II (ABIDE II) and compare its performance with two other common harmonization methods: ICA and ComBat. To assess the effectiveness of the harmonization methods, we utilize various techniques for visualizing and quantifying site effects before and after denoising. Additionally, we evaluate the denoising methods in terms of their ability to preserve signal effects.

# 2. Methods

## 2.1. Multi-site fMRI data

We utilized data from Autism Brain Imaging Data Exchange II (ABIDE II) to investigate the impact of site effects on ALFF and ReHo, and the performance of ICA-DP for denoising site effects and preserving signal effects.

Neuroimaging data from 1,114 subjects collected by 18 different sites with various scanner manufacturers (Simens, Philips, and GE) (Di Martino et al., 2017) were obtained from the ABIDE II dataset.[1] We excluded images with obvious artifacts, large head movement (larger than one voxel size), and incomplete scanning of the whole brain. After strict quality control, functional MRI data of 795 subjects [Autism Spectrum Disorder (ASD) patients: 341, Healthy Controls (HC): 454] in 16 sites (data from two centers were fully excluded) were included in our study. The acquisition parameters: scanner and imaging-related details, including repetition time (TR), echo time (TE), flip angle (FA), voxel size, and demographic information (ASD/HC, sex, and age), are summarized in Table 1.

In this study, the site differences are defined as noise variables, and group differences (ASD/HC), age, and sex are regarded as signal variables. The correlation coefficients among these variables are summarized in Table 2. Since site differences are categorical variables, it is not achievable to directly calculate the correlation coefficients between categorical and numeric variables. We used ANOVA to calculate the significant levels of signal variables and site variables.

## 2.2. Data preprocessing

The raw fMRI data were preprocessed with FSL FEAT, including removing the first six volumes, motion correction, and spatial normalization to standard MNI space. Two functional modalities, ALFF and ReHo, were generated from the preprocessed fMRI data with DPABI (Yan et al., 2016). For ReHo, spatial smoothing (with Full Width at Half Maximum (FWHM) of 6 mm) was performed after ReHo calculation, but for ALFF, spatial smoothing was completed before the calculation (Jia et al., 2019).

## 2.3. Harmonization methods

Two most widely used harmonization methods: ComBat and traditional ICA, were applied in this study to show the performance of ICA-DP on site-effects removal. We now describe the three different strategies below.

### 2.3.1. ComBat

ComBat is a GLM-derived method based on empirical Bayes approach. The method assumes that the data can be modeled as a linear combination of signal variables and site effects, which includes additive and multiplicative factors:

$$Y_{Non-denoised} = \alpha + X_{signal}\beta_{signal} + \gamma + \delta\varepsilon \quad (1)$$

where $\alpha$ is the average value, $X_{signal}$ is the design matrix for the signal variables and $\beta_{signal}$ is the corresponding regression coefficient, $\gamma$ and $\delta$ are the additive and multiplicative factors, respectively. Then ComBat normalizes the data by removing the effects of average and signal variables:

$$Y_{Normalized} = Y_{Non-denoised} - \alpha - X_{signal}\beta_{signal} \quad (2)$$

1 http://fcon_1000.projects.nitrc.org/indi/abide/ abide_II.html

TABLE 1 Scanning parameters and demographic information of the multi-site ABIDE II data.

| Sites | Scanners | TR/TE (ms) | FA (degree) | Voxel Size | ASD/HC | Male/Female | Age |
|---|---|---|---|---|---|---|---|
| EMC | GE MR750 | 2000/30 | 85 | 3.6×3.6×4.0 | 14/13 | 22/5 | 8.39 ± 1.03 |
| ETH | Philips Achieva | 2000/25 | 90 | 3×3×3 | 7/22 | 29/0 | 23.36 ± 4.59 |
| GU | Siemens TriTim | 2000/30 | 90 | 3×3×3 | 27/41 | 46/22 | 10.89 ± 1.62 |
| IU | Siemens TriTim | 813/28 | 60 | 3.4×3.4×3.4 | 18/19 | 28/9 | 24.62 ± 7.59 |
| KKI | Philips Achieva | 2500/30 | 75 | 3×3×3 | 25/123 | 89/59 | 10.37 ± 1.27 |
| KUL | PhilipsAchieva | 2500/30 | 90 | 1.6×1.6×3.1 | 25/0 | 25/0 | 23.76 ± 5.10 |
| OHSU | Siemens TriTim | 475/30 | 60 | 3×3×3 | 33/51 | 52/32 | 11.00 ± 2.04 |
| ONRC | Siemens Skyra | 2500/30 | 90 | 3.8×3.8×3.8 | 15/26 | 30/11 | 23.24 ± 4.09 |
| SU | GE SIGNA | 2000/30 | 80 | 3.4×3.4×3.5 | 14/17 | 28/3 | 10.94 ± 1.14 |
| UCLA | Siemens TriTim | 3000/28 | 90 | 3×3×4 | 12/12 | 19/5 | 11.04 ± 2.46 |
| USM | Siemens TriTim | 2000/28 | 90 | 3.1×3.1×4 | 9/12 | 17/4 | 24.34 ± 7.49 |
| BNI | Philips Ingenia | 3000/25 | 80 | 3.8×3.8×4 | 29/28 | 57/0 | 38.86 ± 15.41 |
| IP | Philips Achieva | 2700/45 | 90 | 3.6×3.7×4 | 13/21 | 16/18 | 22.37 ± 10.97 |
| NYU | Siemens Allegra | 2000/15 | 90 | 3×3×4 | 61/28 | 81/8 | 9.24 ± 4.78 |
| SDSU | GE MR750 | 2000/30 | 90 | 3.4×3.4×3.4 | 30/24 | 46/8 | 13.19 ± 3.06 |
| TCD | Philips Achieva | 2000/27 | 90 | 3×3×3.2 | 9/17 | 26/0 | 15.98 ± 3.23 |

The data were collected from 13 different sites: Erasmus University Medical Center (EMC), ETH Zürich (ETH), Georgetown University (GU), Indiana University (IU), Kennedy Krieger Institute (KKI), Katholieke Universiteit Leuven (KUL), Oregon Health and Science University (OHSU), Olin Neuropsychiatry Research Center (ONRC), Stanford University (SU), University of California Los Angeles (UCLA), University of Utah School of Medicine (USM), Barrow Neurological Institute (BNI), Institut Pasteur and Robert Debré Hospital (IP), NYU Langone Medical Center (NYU), San Diego State University (SDSU), Trinity Centre for Health Sciences (TCD).

TABLE 2 The relationship of signal and noise variables.

| Correlation | Site | ASD vs. HC | Age | Sex |
|---|---|---|---|---|
| Site | 1.000(0.0000) | 3.26e-18 | 2.29e-183 | 3.38e-18 |
| ASD vs. HC | 3.26e-18 | 1.000(0.0000) | – | 0.1984(1.69e-8) |
| Age | 2.29e-183 | – | 1.000 (0.0000) | −0.1223(5.50e-4) |
| Sex | 3.38e-18 | 0.1984(1.69e-8) | −0.1223(5.50e-4) | 1.000(0.0000) |

Finally, ComBat uses an empirical Bayes (EB) framework to get an improved estimate of site effects $\gamma^*$ and $\delta^*$, after removing these site effects and adding the effects of average and signal variables back, we finally get the denoised data by ComBat:

$$Y_{Denoised}^{ComBat} = \frac{Y_{Non-denoised} - \alpha - X_{signal}\beta_{signal} - \gamma^*}{\delta^*} + \alpha + X_{signal}\beta_{signal} \quad (3)$$

## 2.3.2. ICA and ICA-DP

ICA is a data-driven strategy that decomposes the data matrix into a set of statistically independent non-Gaussian maps together with associated courses (e.g., time, subject).

$$Y_{Non-denoised} = A * S \quad (4)$$

where S is the spatial map and A is the corresponding courses. Compared with our ICA-DP, we rename the traditional ICA as ICA-SP (single-projection). To preserve the signal effects, ICA-SP method only removes those pure site-related components (related to site effects only), and leaves those mixed components without any process.

$$Y_{Denoised}^{ICA-SP} = Y_{Non-denoised} - A_{Sites}\, pinv(A_{Sites})\, Y_{Non-denoised} \quad (5)$$

where $A_{Sites}$ is the course of pure site-related components.

In order to eradicate the site effects, we proposed the ICA-DP method in our previous study (Hao et al., 2023). Firstly, ICA-DP separates the signal effects from the mixed components:

$$A_{Sites}^{'} = A_{Mixed} - Var_{Signal}\, pinv\left(Var_{Signal}\right) A_{Mixed} \quad (6)$$

where $A_{Mixed}$ is the course of mixed components and $Var_{Signal}$ is the signal variable. Then $\left[\ A_{Sites}\ A_{Sites}^{'}\ \right]$ is utilized as the whole site effects to be regressed out.

$$Y_{Denoised}^{ICA-DP} = Y_{Non-denoised} - \left[\ A_{Sites}\ A_{Sites}^{'}\ \right] pinv\left(\left[\ A_{Sites}\ A_{Sites}^{'}\ \right]\right) Y_{Non-denoised} \quad (7)$$

## 2.4. Denoising process

For the ComBat-based method, the input $X_{signal}$ was set as group difference (ASD/HC), age, and sex.

For ICA-based methods, the data were decomposed into 100, 150, and 200 independent components. Pearson correlation and Analysis of Variance (ANOVA) were applied to identify signal, noise, and mixed components based on Hao et al. (2023). Components that only significantly correlated with the signal variables ($p < 0.05$, with Bonferroni correction) were classified as pure signal components. Conversely, those solely correlated with the noise variable were identified as pure noise components. Components related to both signal and noise variables were categorized as mixed components. For ALFF, we identified 79, 120, and 166 pure site-related components, and 21, 29, and 34 mixed components; For ReHo, we identified 83, 136, and 179 pure site-related components, and 15, 10, and 11 mixed components. The ICA-SP method exclusively utilized pure noise components to eliminate site effects, whereas the ICA-DP method employed all noise-related components, including mixed ones, for site effects removal. In the ICA-DP approach, both mixed and pure noise components were used to extract noise effects that are considered as integral site-related noise effects and used for removal. Both ICA methods were implemented using widely used software packages for neuroimaging data analysis, namely MATLAB and FSL MELODIC. FSLeyes[2] and BrainNet (Xia et al., 2013) were used to present the results.

## 2.5. Evaluation of data denoising

We used several strategies to assess the performance of the three different denoising methods in terms of eliminating the site effects and preserving the signal effects. To visualize the site effects, we used t-distributed stochastic neighbor embedding (t-SNE) to observe the distribution of the data points, with a tendency to be clustered by site or not. Group F-test was also used to find the significant differences in ALFF and ReHo for brain regions associated with site differences. It is also important to show whether the methods can preserve the signal effects well. In this study, we used age, sex, and group difference (ASD/HC) as variables of interest. In addition to t-SNE and group-level tests of ASD vs. HC, the Pearson correlation coefficient between median ALFF, ReHo and age was also assessed. For each modality, the median value for each subject was obtained by calculating the median of 100 regions of interest.[3] Then, the obtained values were sorted by age distribution, where different colors represent data from different sites.

# 3. Results

## 3.1. Visualization and quantification of site effects

Figure 1 shows the tSNE-2D projection of ALFF and ReHo before and after site effects denoising. The tSNE can project the data into two vectors, which can be regarded as the two dominant features of the data. The data points of the non-denoised data showed site-clustered distribution as most of the centers had their own specific cluster areas, while this site-clustered distribution disappeared after being denoised by any of the denoising methods.

Figure 2 shows the group-level analysis for site effects. The analysis was based on a generalized linear model implementation of one-way ANOVA (factor: sites; covariates: age, sex, and group difference (ASD/HC)). Both two non-denoised modalities were globally contaminated by site effects. Though ICA-SP method decreased site effects, it was not very effective at denoising them. However, compared with our previous results applying ICA-SP to denoise site effects from structural MRI data (Hao et al., 2023), ICA-SP method did denoise site effects better for fMRI data, since purer site-related components were identified. After denoising with ICA-DP and ComBat, there were no brain regions with ALFF or ReHo that were associated with site difference (FWE-corrected $p < 0.05$).

## 3.2. Visualization and quantification of signal effects

### 3.2.1. Age effects

Figure 3 shows the tSNE-2D projection of ALFF and ReHo before and after site effects denoising. The data points of the non-denoised data did not show age-clustered distribution, while this age-clustered distribution appeared after being denoised by ICA-DP.

Figure 4 displays the correlation between the median values of the connectivity measures across the whole brain and age for each of the two modalities. For ALFF, the Pearson correlation coefficients were −0.3552 (Non-denoised), −0.1287 (ComBat denoised), −0.2603 (ICA-SP denoised), −0.8525 (ICA-DP denoised), respectively. For ReHo, the Pearson correlation coefficients were −0.0684 (Non-denoised), −0.0090 (ComBat denoised), 0.0513 (ICA-SP denoised), −0.4640 (ICA-DP denoised), respectively. From a whole-brain perspective, only the ICA-DP method enhanced the correlation between the two modalities and age.

Figures 5, 6 show the group-level analyses for age on ALFF and ReHo, and correlations between age-related regions and age. In order to better rule out the influence of ASD, we only analyzed the age effect of healthy people. The group-level analyses were based on a generalized linear model implementation of one-way ANOVA (factor: age; covariates: sex). Figure 5 shows the results from ALFF. The negative age effects were not found in the non-denoised data because of the existence of site effects, removal of the effects by all the denoising methods, especially for ICA-DP, could reveal the negative age effects not detected from the non-denoised data. From the results of ICA-DP, regions positively associated with age included Cerebellum, Thalamus, Temporal Lobe, and Frontal Lobe; regions negatively associated with age included Parietal Lobe, Temporal Lobe, and Frontal Lobe for the non-denoised data. Figure 6 shows the results from ReHo. The results had the same tendency as those from ALFF. Site effects masked the negative age effects. Removal of the effects by ComBat and ICA-DP could reveal the negative age effects not detected from the non-denoised data. There were no age effects detected after denoising by ICA-SP. From the results of ICA-DP, regions positively associated with age included the Frontal Lobe, Parietal Lobe, and Temporal Lobe; regions negatively associated with age included Occipital Lobe, and Parietal Lobe.

2 https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSLeyes

3 https://github.com/ThomasYeoLab/CBIG/tree/master /stable_projects/brain_ parcellation/Schaefer2018_LocalGlobal/Parcellations/

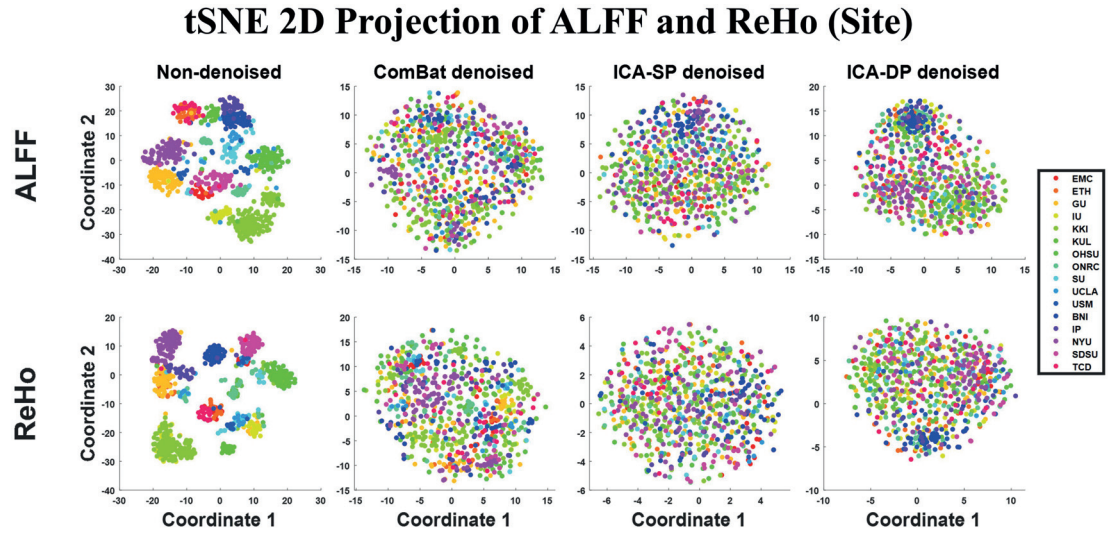


FIGURE 1
Dimension reduction visualization by t-SNE before and after denoising (Sites). The site-clustered distribution before denoising indicated the site effects, and it decreased when the data points were randomly distributed after denoising.

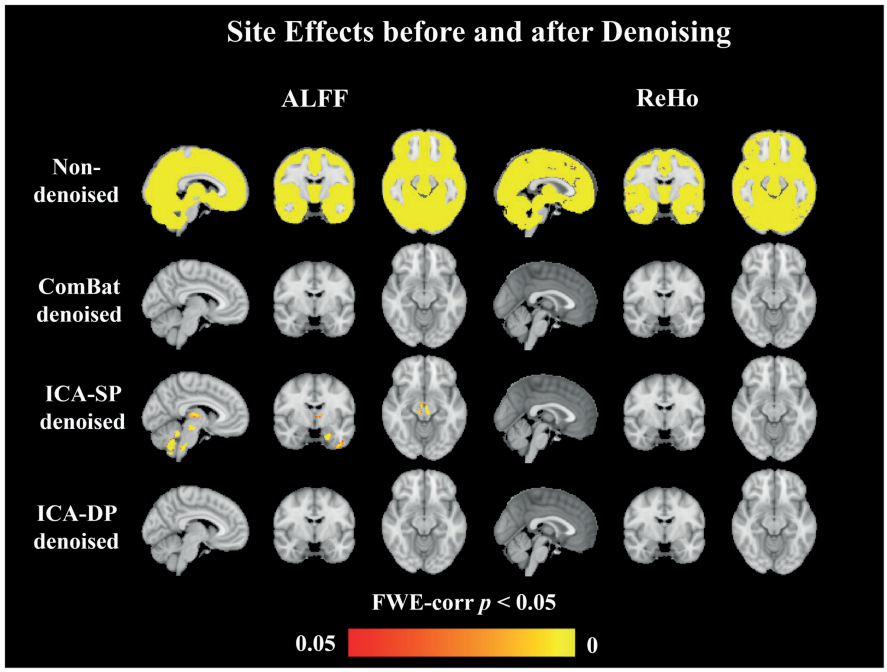


FIGURE 2
Group-level analysis for site effects before and after denoising. Site effects were removed entirely by ComBat and ICA-DP. Though ICA-SP reduced the site effects, some significant regions still could be found.

In summary, ICA-DP increased the age effects by detecting more significantly different regions related to age, while ComBat and ICA-SP decreased the age effects with fewer or no significant regions.

### 3.2.2. Sex effects

Figure 7 shows the tSNE-2D projection of ALFF and ReHo before and after site effects denoising. The data points of the non-denoised

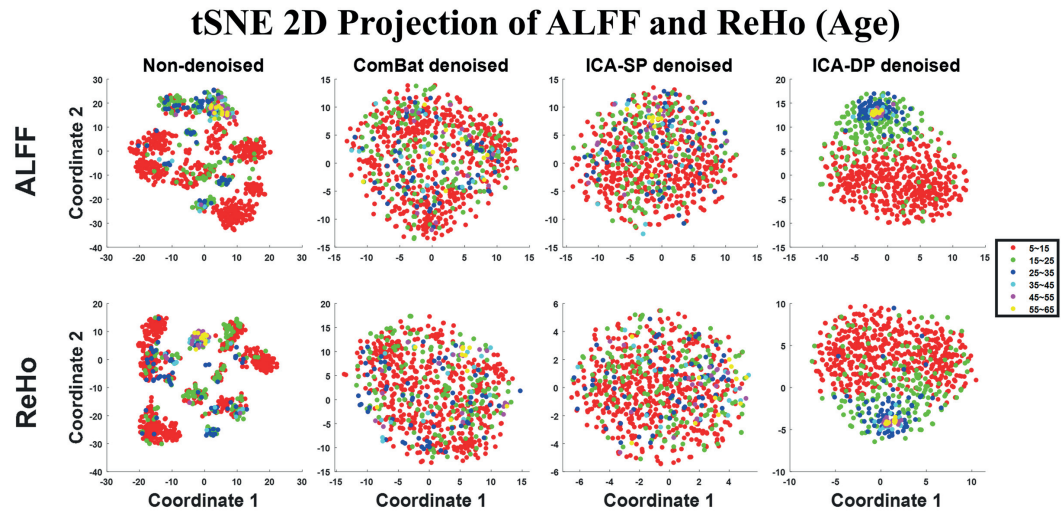


FIGURE 3
Dimension reduction visualization by t-SNE before and after denoising (Age). No age-clustered distribution before denoising, and the age-clustered distribution appeared after denoising by ICA-DP.

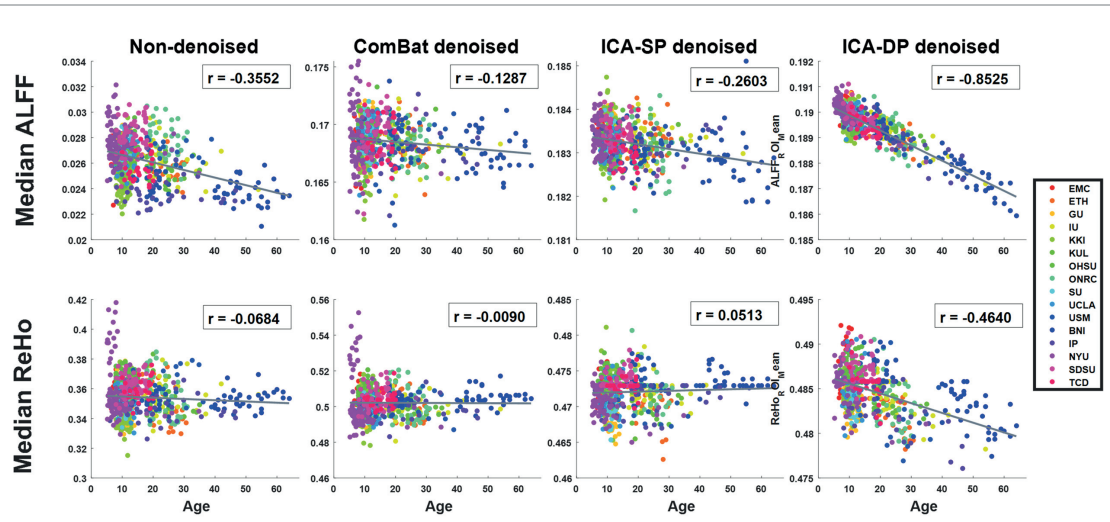


FIGURE 4
Relationship between age and whole-brain median ALFF, and ReHo. For ALFF, the Pearson correlation coefficients were – 0.3552 (Non-denoised), –0.1287 (ComBat denoised), –0.2603 (ICA-SP denoised), –0.8525 (ICA-DP denoised), respectively; for ReHo, the Pearson correlation coefficients were – 0.0684 (Non-denoised), –0.0090 (ComBat denoised), 0.0513 (ICA-SP denoised), –0.4640 (ICA-DP denoised), respectively.

data did not show sex-clustered distribution, which appeared after being denoised by ICA-DP.

Figure 8 displays the group-level analyses for sex on the two fMRI modalities. The group-level analyses were based on a generalized linear model implementation of one-way ANOVA (factor: sex; covariates: age, and group difference (ASD/HC)). For ALFF, we observed several regions that were significantly greater in males, including the Frontal Lobe, Thalamus, and Temporal Lobe.; regions significantly greater in females included Occipital Lobe for the non-denoised data. After denoising, our ICA-DP widened the boundaries of these regions, while the other two methods resulted in the disappearance of these regions. For ReHo, no regions were associated with sex. After denoising with ICA-DP, we identified regions associated with sex. Specifically, regions significantly greater in males included Frontal Lobe, Parietal Lobe, and Occipital Lobe; regions significantly greater in females included Cerebellum, and Temporal Lobe.

Similar to the results for age effects, ICA-DP increased the sex effects by detecting more significantly different regions related to sex, while ComBat and ICA-SP decreased the sex effects with fewer or less significant regions.

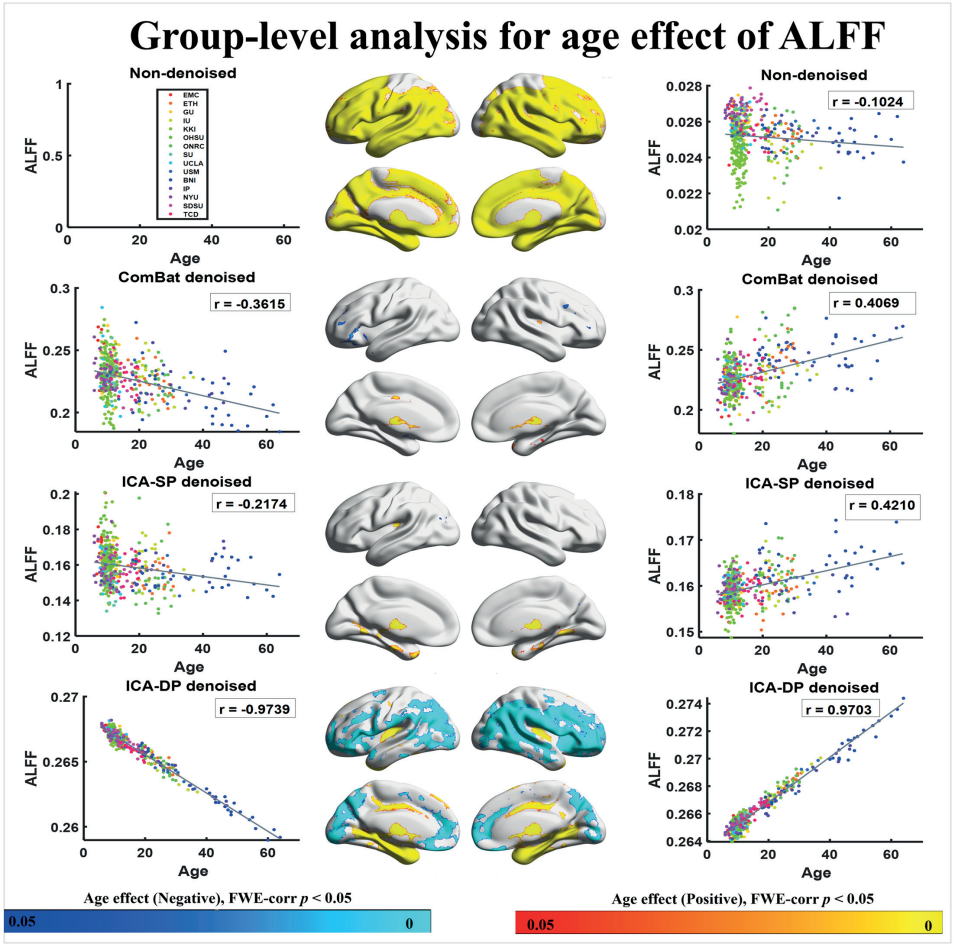


FIGURE 5
Associations between age and ALFF with different denoising strategies. "Positive" association indicates increasing amplitude with increasing age, whereas "Negative" refers to decreasing amplitude with increasing age. Associations with age are enhanced by ICA-DP and weakened by ICA-SP and ComBat.

### 3.2.3. Group difference (ASD/HC)

Figure 9 shows the tSNE-2D projection of ALFF and ReHo before and after site effects denoising. The data points of the non-denoised data could not be divided into two groups according to the group difference (ASD/HC), and only ICA-DP method could enhance the group effects by distinguishing ASD and HC.

Figure 10 demonstrates the impact of denoising on group differences between individuals with autism spectrum disorder (ASD) and healthy controls (HC). The group-level analyses were based on a generalized linear model implementation of one-way ANOVA (factor: group difference (ASD/HC); covariates: age and sex). The results revealed that ICA-DP enhanced the group effects by identifying more regions that were significantly different between the two groups, whereas ComBat and ICA-SP decreased the group effects by detecting fewer or less significant regions. When compared to the non-thresholded group difference maps from the original data (first row), it could be seen that the regions associated with group differences (ASD/HC) from ICA-DP-based denoised data were also present in the original data. This suggested that ICA-DP only amplified the existing signal and did not introduce new information.

## 4. Discussion

In this study, we applied the ICA-DP method to the multi-site harmonization of ALFF and ReHo, and compared it to traditional ICA and ComBat methods for removing site effects and preserving biological variability. The results showed that our ICA-DP method can better remove site effects and preserve physiological signals compared with two other approaches for denoising, ICA-SP, and ComBat.

In the non-denoised data, site effects objectively exist in both modalities: 1) original ALFF and ReHo both show a trend of clustering by site (Figure 1), even if the data from the same site have different distributions of age, sex, and group difference (ASD/HC). To some extent, the statistical differences caused by site differences are greater than those caused by other biological variables (Figures 3, 7, 9).

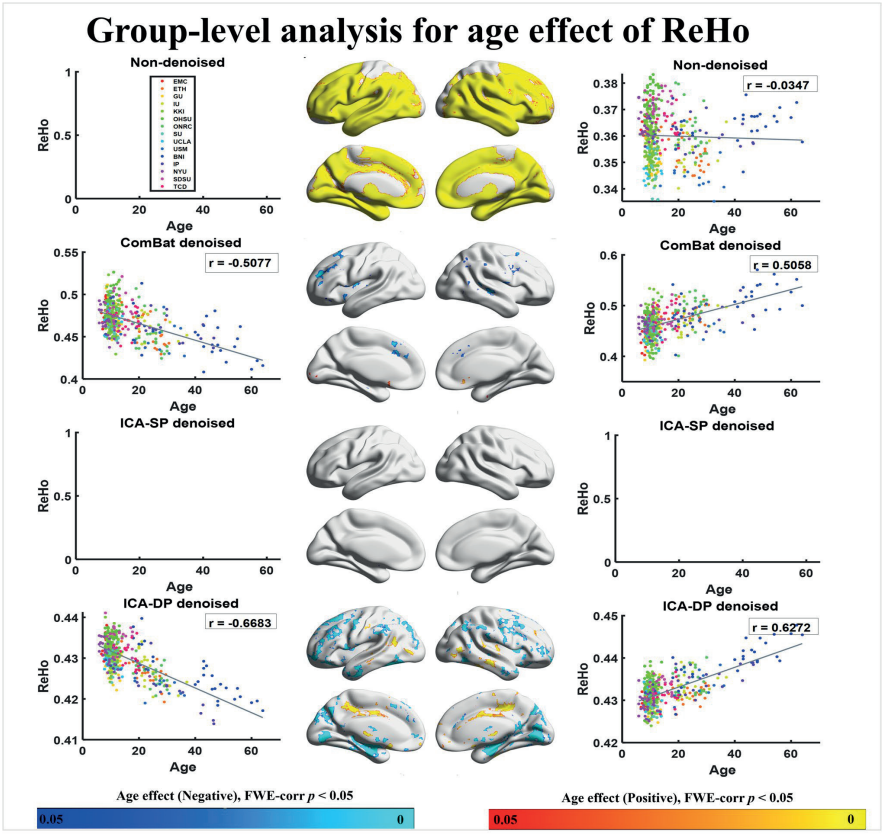


FIGURE 6
Associations between age and ReHo with different denoising strategies. "Positive" refers to significantly increasing amplitude with increasing age, whereas "Negative" refers to significantly increasing amplitude with decreasing age. The age effects are enhanced by ICA-DP, while weakened by ICA-SP and ComBat.

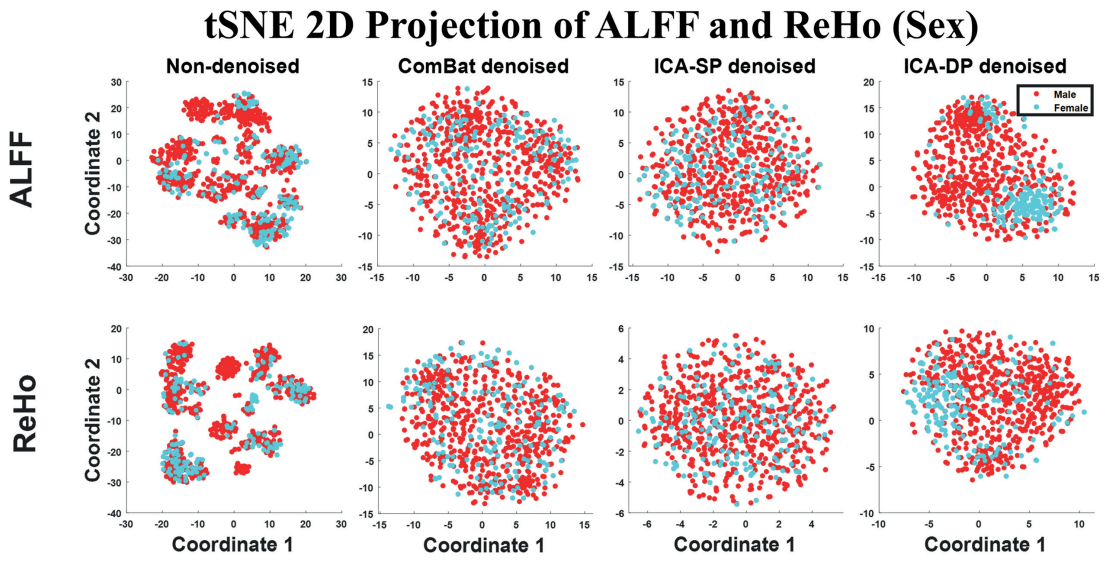


FIGURE 7
Dimension reduction visualization by t-SNE before and after denoising (Sex). No sex-clustered distribution before denoising, and there was a light sex-clustered tendency after denoising by ICA-DP.

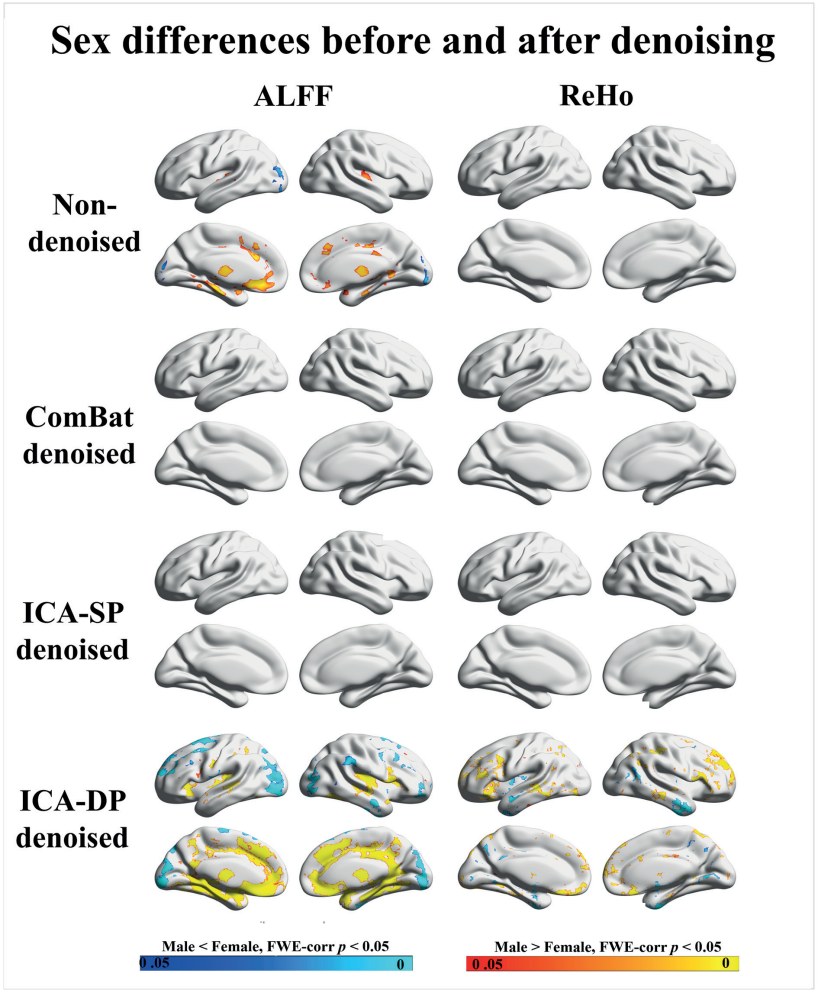


FIGURE 8
Sex differences before and after denoising. "Male < Female" refers to significantly greater amplitude in females, whereas "Male > Female" refers to significantly greater amplitude in males. The sex effects are enhanced by ICA-DP, while weakened by ICA-SP and ComBat.

Moreover, these large site differences may mask the examination of biological effects (Eshaghzadeh Torbati et al., 2021). 2) The results of the F-test indicate that the two modalities all show obvious site differences, and the impact is whole-brain.

Regarding denoising performance, both ICA-DP and ComBat methods can thoroughly remove site effects: (1) the denoised data no longer clusters by site; F-test results no longer have significantly correlated activation regions with site variables. The traditional ICA can only remove the effects of site effects to some extent and cannot eradicate it, because the traditional ICA method only removes the influence of pure noise components and does not deal with mixed components. If the proportion of mixed components in all components is relatively large, or the site effects in mixed components are relatively apparent, the denoising ability of the traditional ICA method will be greatly discounted; on the contrary, the ICA method will have better denoising effect.

In addition to evaluating the performance of the three methods in removing site effects, it is equally important to evaluate their ability to preserve biological signals. To this end, we define age, sex, and group difference (ASD/HC) as variables of interest. The results show that our ICA-DP method effectively removes site effects while also enhancing the examination of biological signals, including the effects of age, sex, and group difference (ASD/HC). The other two methods reduced the examination of these biological effects. Our method's enhancement of biological signals is due to the fact that for each noise component identified, we first regress out the influence of biological signals and then use it for denoising so that the proportion of physiological signals in the denoised data is relatively large and it is easier to detect brain regions that are related to signals through statistical tests. From another perspective, this might result in other variables, in which we are not interested, not being well preserved (Hao et al., 2023). The other limitation of the proposed harmonization method is that when the noise variable is strongly related to the signal variable, ICA-DP could not eliminate the intersection effects related to both site and signal variables.

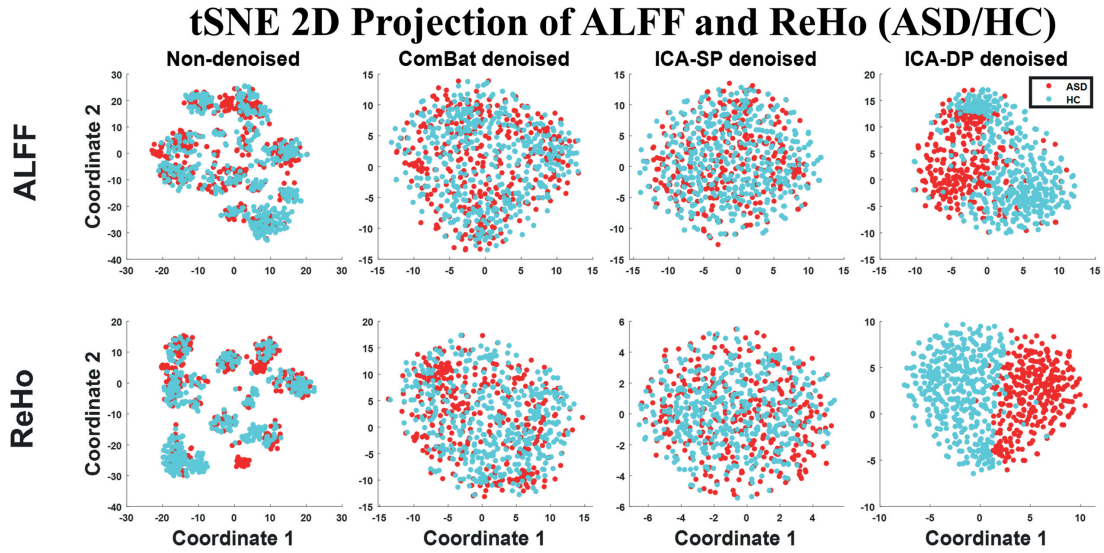


FIGURE 9
Dimension reduction visualization by t-SNE before and after denoising (ASD vs. HC). The data points of the non-denoised data were randomly distributed, and only the data points after denoising by ICA-DP could be divided into two groups according to the group difference (ASD/HC).

In 2010, Biswal et al. conducted a study on age and sex in a large sample of fMRI data from 35 sites. They also reported site effects. Although they did not remove the site effect in their study and just utilized sites as covariates in a generalized linear model (GLM), they still identified some brain regions that were significantly correlated with age and sex in the ALFF. In some of our results (age and sex effect of ALFF), we also found activation regions that highly overlap with Biswal's results. Because we used different datasets and different sample sizes, our results are highly overlapping, but not exactly the same. In addition, we believe that if we apply our method to their dataset and remove the site effect with ICA-DP, more similar activation regions related to age and sex will be founded.

Regarding the statistical results of group differences between individuals with autism spectrum disorder (ASD) and healthy controls (HC), our research identified similar brain regions that have been highlighted in previous studies. For example, in patients with ASD, increased ALFF in the Temporal Lobe and Frontal Lobe while decreased ALFF in the Occipital Lobe was also found compared to HC (Wang et al., 2023). Participants with ASD also showed increased ReHo in the Frontal Lobe and decreased ReHo in the Temporal and Frontal Lobe compared to HC (Paakki et al., 2010; Canario et al., 2021; Wang et al., 2023).

To the best of our knowledge, there are only a few studies focused on the harmonization of multiple sites for ALFF and ReHo to reveal their associations with age, sex, and group difference (ASD/HC). Thus, we are cautious in interpreting the results until the same results can be repeated on a large sample dataset from a single center.

## 5. Conclusion

The combination of multi-site MRI data has the potential to increase the statistical power and improve the reliability and reproducibility of neuroimaging research. However, the analysis of MRI data is often confounded by site effects. Removing these site effects is a critical step in the process of multi-site data fusion. In addition, preserving signals of interest is a major concern when applying any denoising strategy. ICA-SP and ComBat reduced associations with age and sex.

In contrast, our ICA-DP method has proven to be effective in removing site effects and preserving biological variability. With our ICA-DP method, multi-site fMRI data can be harmonized, thus allowing for more robust and accurate analysis. This approach can significantly enhance the validity of neuroimaging research, and we believe it will be a valuable tool for future studies.

## Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: http://fcon_1000.projects.nitrc.org/indi/abide/abide_II.html.

## Author contributions

HX wrote the manuscript with comments from TK, HL, LN, and FC. YZ and DZ downloaded the data and preprocessed them. YH contributed to the guidance of methods. HX carried out the data analysis. All authors read and approved the final manuscript.

## Funding

This work was supported by STI 2030–Major Projects 2022ZD0211500, Science and Technology Planning Project of

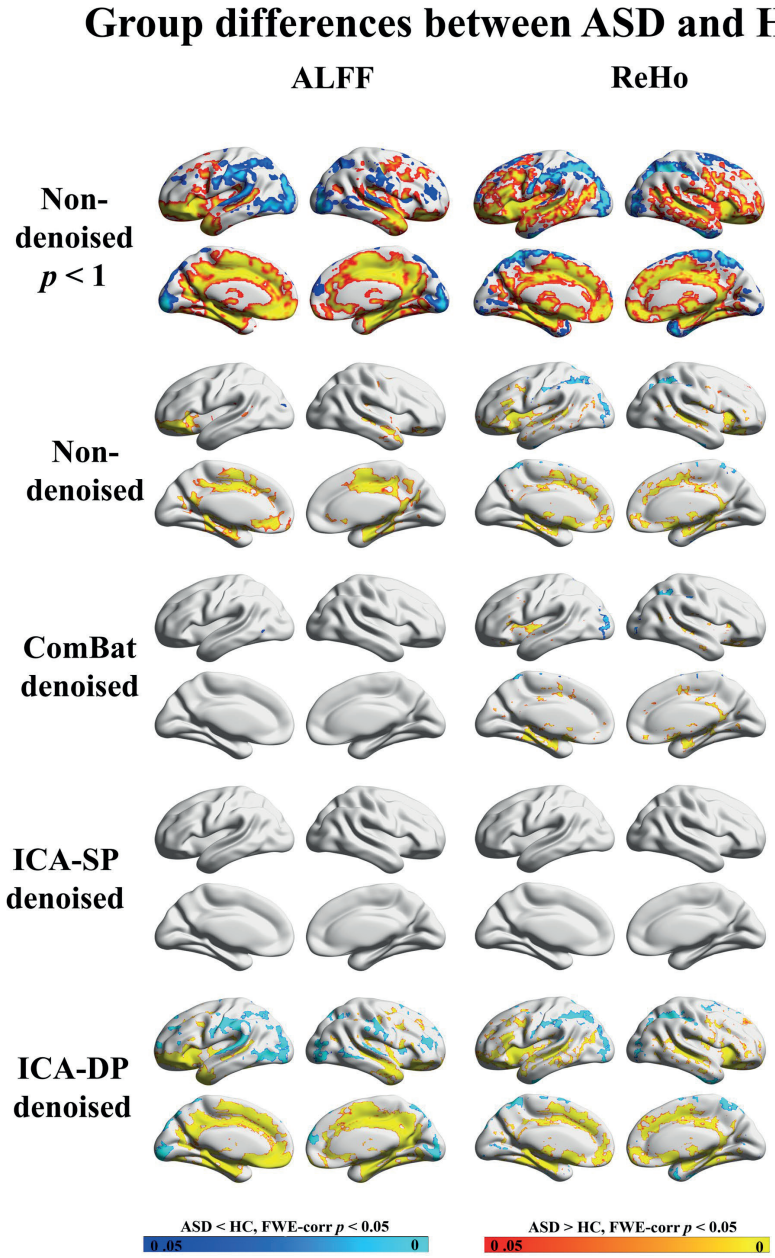


FIGURE 10
Group differences between ASD and HC before and after denoising. No significant differences in regions were found from the data denoised by ICA-SP, and fewer regions were found from the data denoised by ComBat, while ICA-DP could increase the significance of the regions related to ASD/HC. FWE-corr $p < 1$ was shown for non-denoised data to indicate that the regions tested from ICA-DP denoised data were not reintroduced artifacts.

Liaoning Provincial (nos. 2022JH2/10700002 and 2021JH1/10400049), National Natural Science Foundation of China [grant numbers 91748105 and 81471742], National Foundation in China [grant number JCKY 2019110B009], and National Institutes of Health [NIA RF1 AG078304].

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note



## References

Beer, J. C., Tustison, N. J., Cook, P. A., Davatzikos, C., Sheline, Y. I., Shinohara, R. T., et al. (2020). Longitudinal ComBat: a method for harmonizing longitudinal multi-scanner imaging data. *NeuroImage* 220:117129. doi: 10.1016/j.neuroimage.2020.117129

Bell, T. K., Godfrey, K. J., Ware, A. L., Yeates, K. O., and Harris, A. D. (2022). Harmonization of multi-site MRS data with ComBat. *NeuroImage* 257:119330. doi: 10.1016/j.neuroimage.2022.119330

Biswal, B. B., Mennes, M., Zuo, X. N., Gohel, S., Kelly, C., Smith, S. M., et al. (2010). Toward discovery science of human brain function. *Proc. Natl. Acad. Sci. U. S. A.* 107, 4734–4739. doi: 10.1073/pnas.0911855107

Cackowski, S., Barbier, E. L., Dojat, M., and Christen, T. (2021). comBat versus cycleGAN for multi-center MR images harmonization. *Proc. Mach. Learn. Res. Under Rev.* 2017, 1–11.

Canario, E., Chen, D., and Biswal, B. (2021). A review of resting-state fMRI and its use to examine psychiatric disorders. *Psychoradiology* 1, 42–53. doi: 10.1093/psyrad/kkab003

Cetin-Karayumak, S., Stegmayer, K., Walther, S., Szeszko, P. R., Crow, T., James, A., et al. (2020). Exploring the limits of ComBat method for multi-site diffusion MRI harmonization. BioRxiv [Preprint]. doi: 10.1101/2020.11.20.390120

Da-ano, R., Masson, I., Lucia, F., Doré, M., Robin, P., Alfieri, J., et al. (2020). Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. *Sci. Rep.* 10:10248. doi: 10.1038/s41598-020-66110-w

Di Martino, A., O'Connor, D., Chen, B., Alaerts, K., Anderson, J. S., Assaf, M., et al. (2017). Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Sci. Data* 4, 170010–170015. doi: 10.1038/sdata.2017.10

Eklund, A., Nichols, T. E., and Knutsson, H. (2016). Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. U. S. A.* 113, 7900–7905. doi: 10.1073/pnas.1602413113

Eshaghzadeh Torbati, M., Minhas, D. S., Ahmad, G., O'Connor, E. E., Muschelli, J., Laymon, C. M., et al. (2021). A multi-scanner neuroimaging data harmonization using RAVEL and ComBat. *NeuroImage* 245:118703. doi: 10.1016/j.neuroimage.2021.118703

Fortin, J. P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., et al. (2018). Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage* 167, 104–120. doi: 10.1016/j.neuroimage.2017.11.024

Fortin, J. P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., et al. (2017). Harmonization of multi-site diffusion tensor imaging data. *NeuroImage* 161, 149–170. doi: 10.1016/j.neuroimage.2017.08.047

Groves, A. R., Beckmann, C. F., Smith, S. M., and Woolrich, M. W. (2011). Linked independent component analysis for multimodal data fusion. *NeuroImage* 54, 2198–2217. doi: 10.1016/j.neuroimage.2010.09.073

Hao, Y., Xu, H., Xia, M., Yan, C., Zhang, Y., Zhou, D., et al. (2023). Site effects depth denoising and signal enhancement using dual-projection based ICA model. BioRxiv [Preprint]. doi: 10.1101/2023.04.26.538366

Jia, X. Z., Wang, J., Sun, H. Y., Zhang, H., Liao, W., Wang, Z., et al. (2019). RESTplus: an improved toolkit for resting-state functional magnetic resonance imaging data processing. *Sci. Bull.* 64, 953–954. doi: 10.1016/j.scib.2019.05.008

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

Li, H., Smith, S. M., Gruber, S., Lukas, S. E., Silveri, M. M., Hill, K. P., et al. (2020). Denoising scanner effects from multimodal MRI data using linked independent component analysis. *NeuroImage* 208:116388. doi: 10.1016/j.neuroimage.2019.116388

Maikusa, N., Zhu, Y., Uematsu, A., Yamashita, A., Saotome, K., Okada, N., et al. (2021). Comparison of traveling-subject and ComBat harmonization methods for assessing structural brain characteristics. *Hum. Brain Mapp.* 42, 5278–5287. doi: 10.1002/hbm.25615

Nichols, T. E., Das, S., Eickhoff, S. B., Evans, A. C., Glatard, T., Hanke, M., et al. (2017). Best practices in data analysis and sharing in neuroimaging using MRI. *Nat. Neurosci.* 20, 299–303. doi: 10.1038/nn.4500

Orlhac, F., Eertink, J. J., Cottereau, A. S., Zijlstra, J. M., Thieblemont, C., Meignan, M., et al. (2022). A guide to ComBat harmonization of imaging biomarkers in multicenter studies. *J. Nucl. Med. Off.* 63, 172–179. doi: 10.2967/jnumed.121.262464

Paakki, J. J., Rahko, J., Long, X., Moilanen, I., Tervonen, O., Nikkinen, J., et al. (2010). Alterations in regional homogeneity of resting-state brain activity in autism spectrum disorders. *Brain Res.* 1321, 169–179. doi: 10.1016/j.brainres.2009.12.081

Pinto, M. S., Paolella, R., Billiet, T., Van Dyck, P., Guns, P. J., Jeurissen, B., et al. (2020). Harmonization of brain diffusion MRI: concepts and methods. *Front. Neurosci.* 14:e00396. doi: 10.3389/fnins.2020.00396

Stein, C. K., Qu, P., Epstein, J., Buros, A., Rosenthal, A., Crowley, J., et al. (2015). Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat. *BMC Bioinform. 16*, 63–69. doi: 10.1186/s12859-015-0478-3

Wang, J., Rui, L., Jia, L., Yingzi, Y., Jing, M., Xiaohui, S., et al. (2023). Delineation of functional changes and associated cortical transcriptomic proles for autism spectrum disorders. *Research Square* [Preprint]. doi: 10.21203/rs.3.rs-2610086/v1

Xia, M., Wang, J., and He, Y. (2013). BrainNet Viewer: A network visualization tool for human brain connectomics. *PLoS ONE* 8:e68910. doi: 10.1371/journal.pone.0068910

Yan, C. G., Wang, X.Di, Zuo, X. N., and Zang, Y. F. (2016). DPABI: Data Processing & Analysis for (resting-state) brain imaging. *Neuroinformatics*, *14*, 339–351. doi: 10.1007/s12021-016-9299-4

Yang, J., Gohel, S., and Vachha, B. (2020). Current methods and new directions in resting state fMRI. *Clin. Imag.* 65, 47–53. doi: 10.1016/j.clinimag.2020.04.004

Zang, Y. F., Yong, H., Chao-Zhe, Z., Qing-Jiu, C., Man-Qiu, S., Meng, L., et al. (2007). Altered baseline brain activity in children with ADHD revealed by resting-state functional MRI. *Brain Dev. 29*, 83–91. doi: 10.1016/j.braindev.2006.07.002

Zang, Y., Jiang, T., Lu, Y., He, Y., and Tian, L. (2004). Regional homogeneity approach to fMRI data analysis. *NeuroImage 22*, 394–400. doi: 10.1016/j.neuroimage.2003.12.030

# III

# ENHANCING PERFORMANCE OF LINKED INDEPENDENT COMPONENT ANALYSIS: INVESTIGATING THE INFLUENCE OF SUBJECTS AND MODALITIES

by

Huashuai Xu, Tommi Kärkkäinen, Huanjie Li, and Fengyu Cong, 2023

In 2023 International Conference on Computers, Information Processing and Advanced Education (CIPAE), pp. 726-732

# Enhancing Performance of Linked Independent Component Analysis: Investigating the Influence of Subjects and Modalities

Huashuai Xu
Faculty of Information Technology
University of Jyväskylä
Jyväskylä, Finland
huaxu@student.jyu.fi

Huanjie Li
School of Biomedical Engineering,
Dalian University of Technology,
Dalian, China
hj_li@dlut.edu.cn

School of Artificial Intelligence,
Faculty of Electronic
Information and Electrical
Engineering
Key Laboratory of Integrated
Circuit and Biomedical
Electronic System
Dalian University of Technology
Dalian, China
Faculty of Information Technology
University of Jyväskylä
Jyväskylä, Finland
cong@dlut.edu.cn

Tommi Kärkkäinen
Faculty of Information Technology
University of Jyväskylä
Jyväskylä, Finland
tka@jyu.fi

Fengyu Cong
School of Biomedical Engineering,
Faculty of Electronic
Information and Electrical
Engineering

***Abstract*—In recent years, neuroimaging studies have increasingly been acquiring multiple modalities of data. The benefit of integrating multiple modalities through fusion lies in its ability to combine the unique strengths of each modality when analyzed collectively, as opposed to examining each one individually. In 2011, Adrian R. Groves proposed the Linked independent component analysis (LICA) method, which simultaneously models and discovers common features across multiple modalities. LICA has emerged as a powerful technique for analyzing multivariate data, particularly in neuroimaging and biomedical signal processing. The performance of LICA can be affected by the number of subjects and modalities. However, the detailed influence of the number of subjects and modalities on its performance remains an open question. In this study, we test the effects of the number of subjects and modalities on the performance of LICA using both simulated multimodal MRI data and the real multimodal MRI datasets from Autism Brain Imaging Data Exchange II (ABIDE II). Simulated data were utilized to evaluate the influence of subjects and modalities' variabilities. Real multi-site MRI data were used to demonstrate the advantages of multimodal fusion in identifying site-related components and removing site effects. Based on the simulation results, we found that increasing the number of modalities and subjects can improve the results when LICA can not recover the spatial maps or subject courses well. The correlation among subject courses from various modalities, the number of modalities, and the choice of components for decomposition all affect the linking performance of LICA. Our results from real-world datasets also demonstrated the advantages of multimodal fusion by LICA: 1) identify more site-related components; 2) remove more site effects.**

***Keywords— LICA, multimodal, multi-site, site effects***

## I. INTRODUCTION

In recent years, it has been common for neuroimaging studies to acquire multiple modalities of data from the same individual using different imaging techniques [1]. Different forms of brain data can provide various perspectives on both the function and structure of the brain. For instance, functional magnetic resonance imaging (fMRI) measures cerebral activity through the observation of alterations in blood circulation and oxygen levels; structural MRI provides high-resolution images of the brain's structure, and captures the differences in tissue properties, such as gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF), to create detailed, three-dimensional images of the brain's anatomy [2], [3]. These data are usually analyzed separately, and the joint information among these modalities can not be used. One primary incentive for employing multimodal fusion is to capitalize on the interconnected data offered by diverse imaging methods. This can be particularly beneficial for discerning trends of correlated alterations across multiple types, when they exist. Multimodal fusion refers to using a common symmetric model that explains different sorts of data [4]. Since integrating information from differing data types can lead to more accurate and comprehensive insights, efficient and appropriate multimodal fusion methods are necessary for the development of neuroimaging studies.

Several methods have been developed to address the challenges of multimodal data fusion, including but not limited to Linked Independent Component Analysis (LICA), Joint Independent Component Analysis (jICA), and multimodal Canonical Correlation Analysis (mCCA). More details can be seen in Sui's reviewing paper [2].

jICA is a multivariate data fusion method that extends traditional ICA to integrate information from multiple modalities. The main goal of jICA is to identify shared independent components across the different datasets while accounting for their inherent relationships [5]. jICA concatenates the datasets from different modalities along a specific dimension (usually subjects or time points). The concatenated data is then subjected to ICA to extract the joint independent components that capture the shared information across the modalities. By incorporating the inter-modality relationships, jICA allows for identifying common latent factors and investigating their effects on each modality. jICA has the assumption that the number of independent components is equal across modalities. However, different modalities may have different spatial source histograms. So

jICA has the potential loss of modality-specific information. Since different modalities may have different noise levels and a different number of voxels, jICA will be dominated by the largest-variance modalities or the modalities with the most voxels if the scaling is mismatched. So jICA requires the same resolution and smoothing instead of optimized values for all modalities. mCCA is a data fusion technique that extends traditional CCA to handle multiple modalities simultaneously. The main goal of mCCA is to identify linear combinations (called canonical variates, CVs) from each dataset, such that the correlations among these canonical variates are maximized across all modalities [6], [7]. mCCA has the capability to identify common patterns and underlying latent factors across multiple modalities. However, it also has some limitations, such as the assumption of linear relationships between the datasets and the reliance on the covariance structure, which might not capture complex, nonlinear relationships.

In 2011, Adrian R. Groves proposed the Linked independent component analysis (LICA) method [8]. LICA is an advanced multivariate data fusion technique that allows researchers to identify common features across multiple modalities. It is similar to independent component analysis (ICA), a method for identifying and separating independent sources in a single dataset. In addition, LICA extends the basic principles of ICA to allow for data integration from multiple modalities. Unlike jICA, which concatenates datasets and extracts a single set of independent components (ICs) representing shared information, LICA operates by linking the ICs of each dataset through a common set of mixing coefficients. This approach allows LICA to model both the shared and modality-specific information while accounting for the relationships between the datasets.

Compared with jICA and mCCA, LICA offers several advantages: 1) flexibility: modalities can potentially have completely different units, signal-to-noise ratios (SNR), voxel counts, spatial smoothing levels, and intensity distributions; 2) specificity: LICA automatically determines the optimal weighting of each modality, and also can detect single-modality structured components when present. These qualities make LICA a promising approach for the integration of multimodal data.

Since being proposed, LICA has been widely used in neuroimaging studies [1], [9]. Adrian R. Groves used LICA to analyze the age effects from multiple modalities [1]. Li proposed a denoising method [9] for multimodal imaging measures that implemented LICA as a novel approach to remove site effects from multi-study data, and the novel method showed more effective performance at removing site-related effects compared with the conventional single-modality ICA denoising methods. The performance of LICA is affected by the number of subjects and modalities. However, few studies have investigated the influence of these parameters. To show the best power of LICA in neuroscience studies, in this study, we test the effects of the number of subjects and modalities on the performance of LICA using both simulated multimodal MRI data and the real multimodal MRI datasets from Autism Brain Imaging Data Exchange II (ABIDE II). Simulated data were employed to assess the impact of variability in subjects and modalities, while real multi-site MRI data were utilized to showcase the benefits of multimodal fusion in terms of identifying site-related components and mitigating site effects.

## II. Methods

### A. Study data

#### 1) Simulated data

To assess the effect of the number of subjects on the LICA results, we used three different component numbers of 10, 20, and 30, and for each choice of component numbers (all the simulated components are spatial independent), the number of subjects varied from 40 to 200 (40, 50, 60, 70, 80, 90, 100, 150, 200). The criterion for evaluating LICA is calculating the correlation between the spatial maps and subject courses generated from LICA and those used in the simulation. The simulated MRI data, including the different number of sources and subjects, were generated in this study. For each subject, the data were generated by computing spatial maps and one set of ground truth subject courses. Each component map was multiplied by the corresponding subject course, and then they were added together to obtain the simulated MRI data for each subject. The spatial maps (shown in Fig. 1) were obtained by combining different areas of the standard brain template (https://github.com/ThomasYeoLab/CBIG/tree/master/stable_projects/brain_parcellation/Schaefer2018_LocalGlobal/Parcellations/ ).

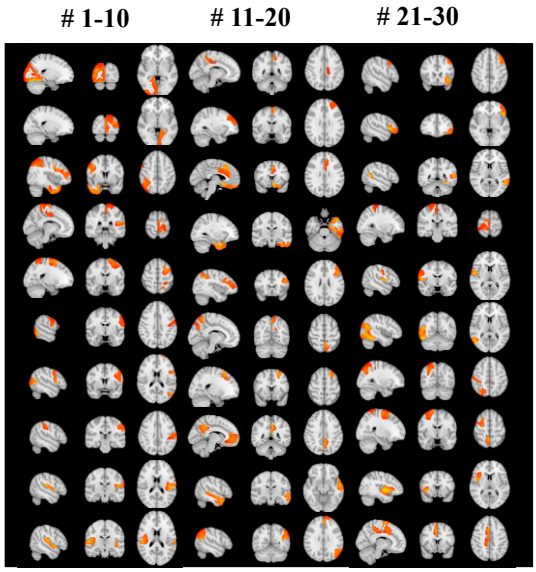


Fig. 1. 30 independent brain spatial maps used to simulate MRI data.

We assessed the effect of the number of modalities on the LICA results from two perspectives of view: 1) We input the simulated data mentioned above repeatedly (regarded as Pseudo multimodal data) as different modalities into LICA to evaluate the influence of the number of modalities; 2) one interesting multimodal component was set firstly by defining a signal variable, which is related to one component in each modality at different levels (see Table 1) and regarded the other nine components as non-interested components. This multimodal component from LICA was used to evaluate the effects of the number of subjects and modalities on LICA. The number of subjects we used here was 100.

TABLE I. There is one component related to the signal variable in each modality, and the correlation coefficient ranges from 0.4 to 0.9, with 0.1 intervals. The corresponding correlation coefficients among subject courses from different modalities can be seen in the right column.

| Correlation between signal and subject courses | Correlation among subject courses from different modalities |
|---|---|
| 0.4 | 0.16 |
| 0.5 | 0.25 |
| 0.6 | 0.36 |
| 0.7 | 0.49 |
| 0.8 | 0.64 |
| 0.9 | 0.81 |

*2) Multi-site MRI data from ABIDE II*

A total of 309 participants were sourced from five distinct locations, using various equipment manufacturers such as Siemens, Philips, and GE. This data was obtained from the Autism Brain Imaging Data Exchange II (ABIDE II) dataset, accessible through [10]. We omitted images that contained noticeable artifacts, exhibited significant head movement (exceeding the size of a single voxel), or lacked comprehensive brain scans. Following rigorous quality control, we included functional MRI data for 309 subjects in our analysis—comprising 91 individuals with Autism Spectrum Disorder (ASD) and 218 Healthy Controls (HC). Relevant scanning and demographic variables, like repetition time (TR), echo time (TE), flip angle (FA), voxel dimensions, as well as ASD/HC categorization, gender, and age, are outlined in Table 2.

For each subject, six modalities of MRI data were generated, including grey matter (GM), cortical thickness (CT), pial surface area (PSA), and three functional MRI outcomes: regional homogeneity (ReHo), amplitude of low frequency fluctuation (ALFF), fractional amplitude of low frequency fluctuations (fALFF). In this study, to match the number of subjects in each site, we utilized ALFF, fALFF, and ReHo to test LICA in terms of multimodal data fusion.

TABLE II. Parameters for Scanning and Demographic Details for Multi-Site ABIDE II Dataset. The data were collected from 5 different sites: Erasmus University Medical Center (EMC), ETH Zürich (ETH), Georgetown University (GU), Indiana University (IU), Kennedy Krieger Institute (KKI)

| Sites | Scan-ners | TR/TE (ms) | FA (deg-ree) | Voxel Size | ASD/ HC | Gen-der | Age |
|---|---|---|---|---|---|---|---|
| EMC | GE MR750 | 2000/ 30 | 85 | 3.6*3.6 *4.0 | 14/13 | 22/5 | 8.39 ±1.03 |
| ETH | Philips Achieva | 2000/ 25 | 90 | 3*3 *3 | 7/22 | 29/0 | 23.36 ±4.59 |
| GU | Siemens TriTim | 2000/ 30 | 90 | 3*3 *3 | 27/41 | 46/22 | 10.89 ±1.62 |
| IU | Siemens TriTim | 813/ 28 | 60 | 3.4*3.4 *3.4 | 18/19 | 28/9 | 24.62 ±7.59 |
| KKI | Philips Achieva | 2500/ 30 | 75 | 3*3 *3 | 25/123 | 89/59 | 10.37 ±1.27 |

In this study, the site differences are defined as noise variables, and group differences (ASD/HC), age, and gender are regarded as signal variables. We used data from two to five sites to validate the LICA method.

## B. Data preprocessing

We preprocessed the initial fMRI datasets using FSL FEAT, which involved the exclusion of the first six volumes, alongside procedures for motion rectification and spatial standardization to the conventional MNI coordinate system. Utilizing DPABI [11], two distinct functional parameters, namely ALFF (Amplitude of Low Frequency Fluctuations) and ReHo (Regional Homogeneity), were derived from the cleansed fMRI information. For ReHo, spatial refinement was carried out post-calculation, using a Full Width at Half Maximum (FWHM) value of 6 mm. In contrast, for ALFF, this spatial smoothing process was executed prior to the actual calculation [12].

## C. Run LICA and post analysis

We downloaded the LICA tool from FSL website (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLICA), and then put it under Matlab (https://uk.mathworks.com/ ) folder after unpacking. About the setting of the number of components, 1) for the simulated data, we just set the number of components as same as in the simulation since we knew the ground truth; 2) for the real MRI data, since LICA can automatically determine the number of components that are needed to describe the data optimally, we just preset a large number of components, then allow LICA to downweight and eliminate the weak components gradually.

The subject courses for each component were assessed for relationships with site and signal using the Pearson correlation coefficient. As the differences between sites are categorical in nature, direct correlation coefficient calculations between categorical and numerical variables are not feasible. To assess the significance levels of subject trajectories and site-specific variables, we employed Analysis of Variance (ANOVA) methods.

# III. Results

We first show the results from simulated data, including the influence of the number of subjects and modalities. Then, we show the results from the real-world fMRI outcomes from the ABIDE II dataset, including the signal- and site-related components and denoising results from a single modality and multiple modalities.

### A. *Results from simulated data*

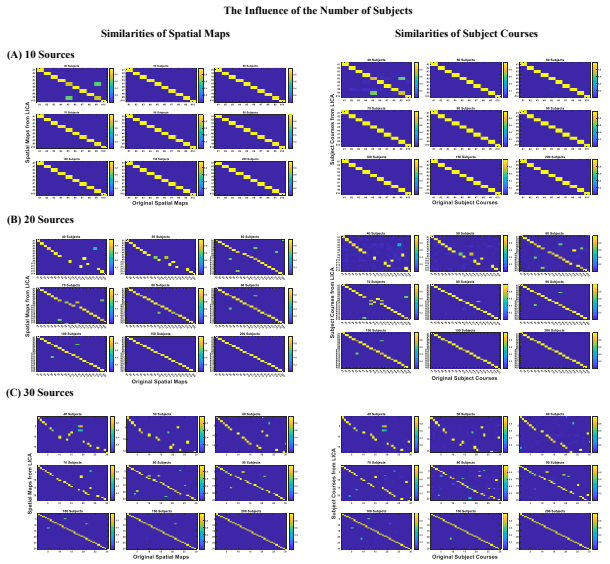


Fig. 2. The influence of the number of subjects on recovering spatial maps and subject loadings of LICA. (A) When there are 10 independent components, LICA can recover all the components from the data when the number of subjects is larger than 40, and there are two correlated components that should be independent when the number of subjects is 40; (B) when there are 20 independent components, LICA can recover all the components from the data with more than 80 subjects, and can get all totally independent components when the number of subjects is more than 150; (C) when there are 30 independent components, LICA can recover all the components from the data with more than 100 subjects, and can get can all totally independent components when the number of subjects is more than 200.

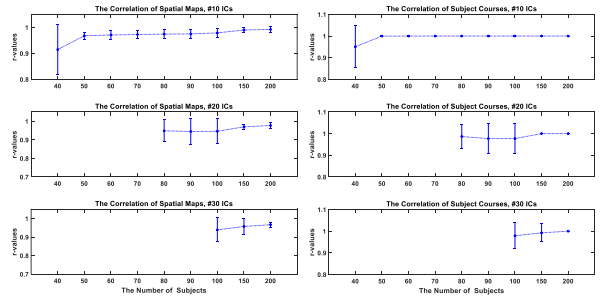


Fig. 3. The influence of the number of subjects on recovering spatial maps and subject loadings of LICA. The r-values were obtained by computing the mean values of correlation coefficients between the components identified by LICA and used in the simulated data. LICA can recover all the components only when the number of subjects is large enough: a minimum of 40 subjects were required to recover 10 components, 80 subjects for 20 components, and 100 subjects for 30 components.

Figs. 2 and 3 showed the influence of the number of subjects on recovering spatial maps and subject loadings of LICA. When the number of independent components is 10, all the simulated components can be recovered by LICA. However, two components are correlated when the number of subjects is not large enough (say 40). In other words, the 10 components generated by LICA are not as totally independent as those used in the simulation. As the number of subjects increases, all 10 simulated independent components can be recovered from LICA. When the number of independent components is 20, LICA can only recover parts of the simulated components when the number of subjects is less than 80. Specifically, 14 components can be recovered when the number of subjects is 40 and 19 components can be recovered with 60 subjects. In addition, LICA can recover all 20 components when the number of subjects is more than 80 and recover all the totally independent components when the number of subjects is more than 150. The same trend can be observed when the number of independent components is 30. The numbers of subjects needed to recover all the 30 components and all the independent components are 100 and 200, respectively.

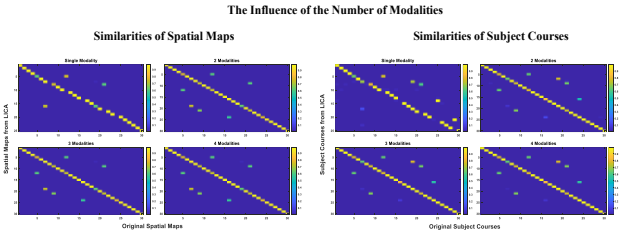


Fig. 4. The influence of the number of modalities on recovering spatial maps and subject loadings when the number of sources is 30 and the number of subjects is 90. LICA can recover more simulated multimodal components from multiple modalities than a single modality.

Figs. 4 and 5 show the influence of the number of modalities from two different perspectives of view. From Fig. 4, when the number of sources is 30 and the number of subjects is 90, LICA can only recover 25 components from a single modality. To increase the number of modalities, we repeatedly added the same data into LICA. In this way, Pseudo multimodal data can be obtained. LICA can recover all 30 components from multiple modalities but can not get 30 totally independent multimodal components even with 4 modalities.

Fig. 5 shows the influence of modalities from another perspective of view. The biggest highlight of the LICA method is its linking function. However, few studies have focused on the linking performance of LICA. In this study, we explore how the correlation among subject courses from different modalities affects the linking performance of LICA. To better focus on the linking performance of LICA, we selected 10 components and 100 subjects, which have proved that LICA can fully recover all components in this case. The results show that the signal-related components from each modality can not always be linked, although they are all related to the signal variable and correlated to each other. Specifically, linking performance can not be realized when the signal-related components among modalities are weakly correlated (correlation coefficient < 0.16). As the correlation increases, the linking function of LICA starts to play a role (correlation coefficient = 0.25 for three and four modalities and 0.49 for two modalities). When the correlation coefficients are large but not large enough (less than 0.81), the choices of the number of components can affect the linking performance. For example, when the correlation coefficient of the signal-related components among multiple modalities is 0.25, LICA can recover either one linked multimodal component or three modality-specific components, and the results strongly depend on the choice of the number of components. When the correlation coefficients are large enough (more than 0.81), the choices of the number of components can not affect the linking performance anymore. LICA can always obtain the multimodal

component and significantly increase the correlation with the signal variable.

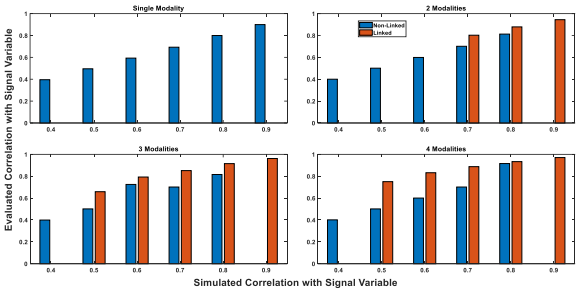


Fig. 5. The influence of the number of modalities on the linking performance. The signal-related components from multiple modalities can not be linked when they are weakly correlated (less than 0.16) and can start to be linked with the increase of the correlation (larger than 0.25). As the number of modalities increases, LICA begins to link the related components earlier, and the correlation between the corresponding components and signal variable also becomes stronger.

## *B. Results from the ABIDE II dataset*

### *1) Linking performance*

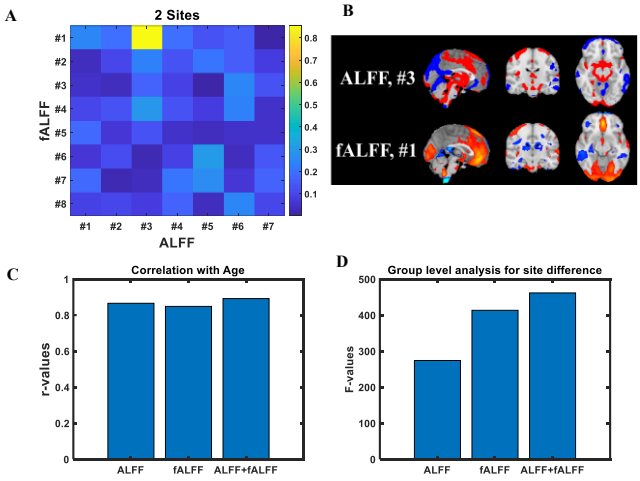


Fig. 6. Components from LICA based on a single modality (ALFF, fALFF) and two modalities. (A) The correlation of subject courses among all the components generated by LICA on ALFF and fALFF; (B) The spatial maps of the significantly related components from ALFF and fALFF; (C) The correlation with age from a single modality and two modalities; (D) The correlation with site difference from a single modality and two modalities

Fig. 6 shows the linking performance of LICA using two-site data from ABIDE II. We first used LICA to analyze ALFF and fALFF, respectively, and then fused the two modalities with LICA. LICA can generate seven components with ALFF data and eight components with fALFF data. Among them, the subject courses of one ALFF component (#3) and one fALFF component (#1) were highly correlated (Fig. 6A). The corresponding spatial maps are shown in Fig. 6B, which are related to both age and site variables. When LICA was used to merge ALFF and fALFF data, these two correlated components were linked together as one multimodal component, and the correlation coefficients of the multimodal components with age and site variables were increased.

### *2) Denoising results*

TABLE III. Results from LICA decomposition. Numbers marked in red in parentheses represent pure site-related components, and in black represent mixed components (related to both site difference and signal variables)

| Data | Subjects | Coms from ALFF | Coms from ALFF+fALFF | Coms from ALFF+fALFF+ReHo |
|---|---|---|---|---|
| 2 Sites | 56 | 7(0+1) | 11(0+2) | 15(0+3) |
| 3 Sites | 124 | 13(0+2) | 23(2+3) | 28(5+9) |
| 4 Sites | 161 | 17(3+3) | 23(6+3) | 32(9+8) |
| 5 Sites | 309 | 42(13+5) | 46(21+8) | 66(23+7) |

Table 3 shows the LICA decomposition results. The results have the same tendency as those from simulated data: with more subjects and/or modalities, LICA can obtain more components. In this study, only pure site-related components (numbers in red) were used to regress out from the original non-denoised data for the goal of site denoising.

Fig. 7 shows the tSNE-2D projection of ALFF data from various sites (3, 4, and 5) before and after site effects denoising. The tSNE can project the data into two vectors, which can be regarded as the two dominant features of the data. The data points of the non-denoised data show site-clustered distribution as all the centers have their own specific cluster areas, while this site-clustered distribution decreases after being denoised by LICA based on a single modality and even disappears by LICA based on multiple modalities. Fig. 8 shows the group-level analysis for site effects. The non-denoised ALFF data were globally affected by the site effects. Though all the LICA (based on a single modality and multiple modalities) methods can remove parts of the site effects, the methods can not alleviate the whole site effects. In addition, LICA based on multiple modalities can remove more site effects than LICA based on a single modality(FWE-corrected $p < 0.05$).

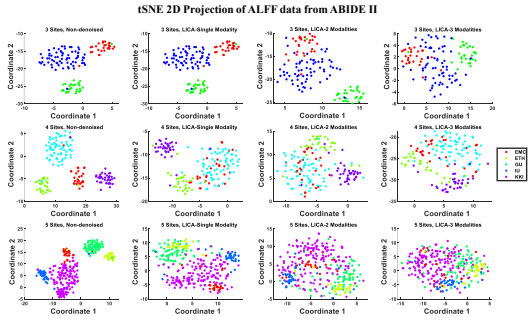


Fig. 7. Dimension reduction visualization by t-SNE before and after denoising. The site-cluster distribution before denoising indicated the site effects, and it decreased when the data points were randomly distributed after denoising.

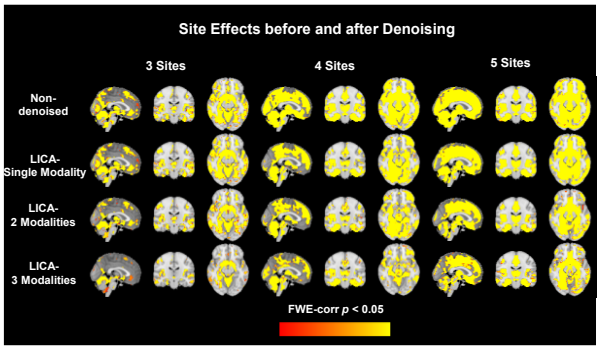


Fig. 8. Group-level analysis for site effects before and after denoising. As the number of sites increases, more significantly different regions are related to sites. LICA based on multiple modalities can remove more site effects than LICA based on a single modality

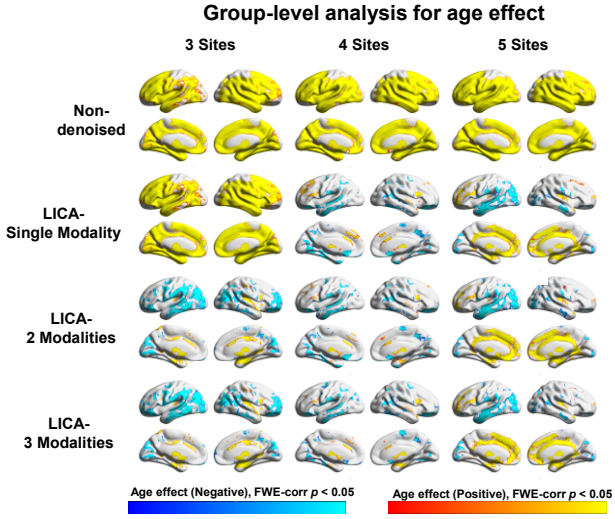


Fig. 9. Group-level analysis for age effects before and after denoising. "Positive" refers to significantly increasing amplitude with increasing age, whereas "Negative" refers to significantly increasing amplitude with decreasing age.

Fig. 9 displays the results of a group-level examination focused on the impact of age. To more effectively isolate age effects without the influence of ASD, we limited our analysis to healthy individuals. This group-level study employed a generalized linear model featuring a one-way ANOVA approach, with age as the factor and gender as a covariate. Negative age effects were not observable in the original, non-denoised data, due to the presence of site-specific influences. However, after applying various denoising techniques to remove these site effects, the concealed negative age impacts became evident.

## IV. Discussions

In this study, we test the performance of LICA using simulated and multimodal data from ABIDE II. Simulated data were used to test the influence of the number of subjects and modalities, and multimodal MRI data from ABIDE II were used to evaluate the linking and denoising performance.

Regarding the number of components obtained by LICA, it's important to note that LICA is a Bayesian form of ICA, setting it apart from traditional ICA approaches like FastICA. LICA integrates dimensionality reduction directly into the ICA procedure by utilizing automatic relevance determination (ARD) priors on the components themselves. [13], [14]. During the interaction, eliminated components (or part-components) are eradicated from the model, avoiding additional inference on these zero-weight spatial maps. So, the number of components obtained by LICA has an upper limit, and the expected number of components may not be achieved sometimes.

There are two ways to obtain more reliable components with LICA: increasing the number of subjects and modalities. The approximate relationship between the number of subjects and the number of components is a four-fold relationship. That is to say, LICA can obtain 10 components (if they exist) from the data with more than 40 subjects and 20 components from the data with more than 80 subjects.

The linking function is the most prominent highlight of the LICA method. LICA can automatically determine the optimal weighting of each modality and detect single-modality structured components when present. This study profoundly evaluates how correlation among subject courses from various modalities affects the linking performance. In general, the realization of the linking function requires certain conditions. It can only be realized when components from two or more modalities are highly correlated. In addition, the number of modalities and the choice of component setting also influence the linking performance.

When the defined signal-related components from each modality are weakly correlated (0.16), the linking function can not be realized even when we increase the number of modalities and choose the smaller numbers for decomposition. As the correlation among subject courses from different modalities increases, the linking function plays a role via the increase of modalities and appropriate composition settings. For example, when the correlation coefficient = 0.25, via setting 10 as the number for composition, the linking performance can be realized from three and four modalities but not from two modalities, which means the number of modalities influences the linking function. Also, if we set 11 as the number for decomposition, LICA can obtain modality-specific components rather than shared ones. When the correlation coefficients are large enough (> 0.81), the number of modalities and choices of the number of components can not affect the linking performance anymore. LICA can always obtain the shared component and the correlation of the multimodal component generated from LICA with the signal variable is larger than that of the single modal component (Fig. 5).

Results from real MRI data (Fig. 6) can also confirm the conclusions above. The strongly correlated components from various modalities will be linked and generate a shared component more related to age and site variables.

Regarding the evaluation results of denoising, since our research only focuses on pure site-related components, and those mixed components usually contribute to the larger proportion of site effects, so all methods cannot alleviate the whole site effects. LICA based on multiple modalities can remove more site effects than LICA based on a single modality.

## V. Conclusions

In conclusion, when LICA can not recover the spatial maps or subject courses well, both increasing the number of modalities and the number of subjects can improve the

results. The factors affecting the linking function include the correlation among subject courses from various modalities, the number of modalities, and the choice of components for decomposition. Our results from real-world datasets also demonstrated the advantages of multimodal fusion by LICA: 1) identify more site-related components; 2) remove more site effects and find age effects masked by site effects.

## ACKNOWLEDGMENT

This work was supported by STI 2030 - Major Projects 2022ZD0211500, Science and Technology Planning Project of Liaoning Provincial (no. 2022JH2/10700002 and 2021JH1/10400049), National Natural Science Foundation of China [grant numbers 91748105 & 81471742], National Foundation in China [grant number JCKY 2019110B009], and the scholarship from China Scholarship Council(No. 201806060167).

## REFERENCES

[1] A. R. Groves *et al.*, "Benefits of multimodal fusion analysis on a large-scale dataset: Life-span patterns of inter-subject variability in cortical morphometry and white matter microstructure," *Neuroimage*, vol. 63, no. 1, pp. 365–380, 2012, doi: 10.1016/j.neuroimage.2012.06.038.

[2] J. Sui, T. Adali, Q. Yu, J. Chen, and V. D. Calhoun, "A review of multivariate methods for multimodal fusion of brain imaging data," *J Neurosci Methods*, vol. 204, no. 1, pp. 68–81, 2012, doi: 10.1016/j.jneumeth.2011.10.031.

[3] J. Sui, Q. Yu, H. He, G. D. Pearlson, and V. D. Calhoun, "A Selective Review of Multimodal Fusion Methods in Schizophrenia," *Front Hum Neurosci*, vol. 6, no. February, pp. 1–11, 2012, doi: 10.3389/fnhum.2012.00027.

[4] K. J. Friston, "Modalities, modes, and models in functional neuroimaging," *Science (1979)*, vol. 326, no. 5951, pp. 399–403, 2009, doi: 10.1126/science.1174521.

[5] V. D. Calhoun, T. Adali, N. R. Giuliani, J. J. Pekar, K. A. Kiehl, and G. D. Pearlson, "Method for multimodal analysis of independent source differences in schizophrenia: Combining gray matter structural and auditory oddball functional data," *Hum Brain Mapp*, vol. 27, no. 1, pp. 47–62, 2006, doi: 10.1002/hbm.20166.

[6] N. M. Correa, Y. O. Li, T. Adali, and V. D. Calhoun, "Canonical correlation analysis for feature-based fusion of biomedical imaging modalities and its application to detection of associative networks in Schizophrenia," *IEEE Journal on Selected Topics in Signal Processing*, vol. 2, no. 6, pp. 998–1007, 2008, doi: 10.1109/JSTSP.2008.2008265.

[7] N. M. Correa, T. Adalı, Y. Li, and V. D. Calhoun, "Canonical correlation analysis for data fusion and group inferences: examining applications of medical imaging data," *IEEE Signal Process Mag*, no. July, pp. 39–50, 2010.

[8] A. R. Groves, C. F. Beckmann, S. M. Smith, and M. W. Woolrich, "Linked independent component analysis for multimodal data fusion," *Neuroimage*, vol. 54, no. 3, pp. 2198–2217, 2011, doi: 10.1016/j.neuroimage.2010.09.073.

[9] H. Li *et al.*, "Denoising scanner effects from multimodal MRI data using linked independent component analysis," *Neuroimage*, vol. 208, no. September 2019, p. 116388, 2020, doi: 10.1016/j.neuroimage.2019.116388.

[10] A. Di Martino *et al.*, "Enhancing studies of the connectome in autism using the autism brain imaging data exchange II," *Sci Data*, vol. 4, pp. 1–15, 2017, doi: 10.1038/sdata.2017.10.

[11] C. G. Yan, X. Di Wang, X. N. Zuo, and Y. F. Zang, "DPABI: Data Processing & Analysis for (Resting-State) Brain Imaging," Neuroinformatics, vol. 14, no. 3, pp. 339–351, Jul. 2016, doi: 10.1007/s12021-016-9299-4.

[12] X. Z. Jia et al., "RESTplus: an improved toolkit for resting-state functional magnetic resonance imaging data processing," Science Bulletin, vol. 64, no. 14. Elsevier B.V., pp. 953–954, Jul. 30, 2019. doi: 10.1016/j.scib.2019.05.008.

[13] C. M. Bishop, "Variational principal components," IEE Conference Publication, vol. 1, no. 470, pp. 509–514, 1999, doi: 10.1049/cp:19991160.

[14] R. Choudrey and S. Roberts, "Flexible Bayesian independent component analysis for blind source separation," 2001. [Online]. Available: http://inc2.ucsd.edu/ica2001/047-choudrey.pdf

# IV

# HARMONIZATION OF MULTI-SITE MRI DATA WITH DUAL-PROJECTION BASED LINKED ICA MODEL

by

Huashuai Xu, Yuxing Hao, Yunge Zhang, Dongyue Zhou, Tommi Kärkkäinen, Lisa D. Nickerson, Huanjie Li, and Fengyu Cong, 2023

To be submitted

Request a copy from the author.