Structure-constrained combination-based nonlinear association analysis between incomplete multimodal imaging and genetic data for biomarker detection of neurodegenerative diseases

Xiumei Chen , Tao Wang , Haoran Lai , Xiaoling Zhang , Qianjin Feng , Meiyan Huang

PII: S1361-8415(22)00070-6
DOI: https://doi.org/10.1016/j.media.2022.102419
Reference: MEDIMA 102419

To appear in: *Medical Image Analysis*

Received date: 2 October 2021
Revised date: 15 February 2022
Accepted date: 10 March 2022

Please cite this article as: Xiumei Chen , Tao Wang , Haoran Lai , Xiaoling Zhang , Qianjin Feng , Meiyan Huang , Structure-constrained combination-based nonlinear association analysis between incomplete multimodal imaging and genetic data for biomarker detection of neurodegenerative diseases, *Medical Image Analysis* (2022), doi: https://doi.org/10.1016/j.media.2022.102419

Highlights:

- Latent imaging representation is learned to exploit inter- and intra-modal interactions.
- Structure constraints are combined to detect the correlations among SNPs and QTs.
- A nonlinear kernel-based method is used to analyze associations between QTs and SNPs.
- Our method is applied in the ND data to detect disease-related biomarkers.
- Modality-shared and modality-specified imaging and genetic biomarkers can be **detected.**

Structure-constrained combination-based nonlinear association analysis between incomplete multimodal imaging and genetic data for biomarker detection of neurodegenerative diseases

Xiumei Chen[a], Tao Wang[a], Haoran Lai[a], Xiaoling Zhang[a], Qianjin Feng[a,b,c,*], and Meiyan Huang[a,b,c,*]

[a] School of Biomedical Engineering, Southern Medical University, Guangzhou, 510515, China

[b] Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou, 510515, China

[c] Guangdong Province Engineering Laboratory for Medical Imaging and Diagnostic Technology, Southern Medical University, Guangzhou, 510515, China

* Corresponding authors

Xiumei Chen
E-mail: chenxiumei97@163.com

Tao Wang
E-mail: wangtao_9802@sina.com

Haoran Lai
E-mail: haoranlai@163.com

Xiaoling Zhang
E-mail: zhangxiaoling9911@163.com

Qianjin Feng
E-mail: fengqj99@smu.edu.cn

Meiyan Huang
E-mail: huangmeiyan16@163.com

Abstract: Multimodal imaging data are widely applied in imaging genetic studies to identify associations between imaging and genetic data for the biomarker detection of neurodegenerative diseases (NDs). However, the incomplete multimodal imaging data and complex relationships among imaging and genetic data make it difficult to effectively analyze associations between imaging and genetic data and accurately detect disease-related biomarkers. This study proposed a novel structure-constrained combination-based nonlinear association analysis method to exploit associations between incomplete multimodal imaging and genetic data for potential biomarker detection of NDs. Two types of structure constraints were used in imaging and genetic data. First, a parallel concatenated projection method with multiple constraints was adopted to handle missing data. Modality-shared and modality-specific information could be well captured to obtain latent imaging representations. A locality preserving constraint was applied to the imaging data for retaining structure information before and after projection. A connectivity penalty was also included to capture structure associations among latent imaging representations. Second, a group-induced graph self-expression constraint was incorporated into our method to exploit strong structure correlations among inter- and intra-group of genetic data. Finally, a nonlinear kernel-based method was used to explore the complex associations between latent imaging representations and genetic data for biomarker detection. A set of simulation data and two sets of real ND data, which were obtained from Alzheimer's disease neuroimaging initiative and Parkinson's progression markers initiative databases, were applied to assess the effectiveness of our method. High accuracy of biomarker detection was achieved. Moreover, the identification of disease-related biomarkers was confirmed in previous studies. Therefore, our method may provide a novel way to gain insights into the pathological mechanism of NDs and early prediction of these diseases.

Keywords: Neurodegenerative diseases, Incomplete multimodal imaging data, Structure-constrained combination, Nonlinear imaging genetics.

1. Introduction

Neurodegenerative diseases (NDs) are a heterogeneous group of disorders that are characterized by the progressive degeneration and death of nerve cells (Lei et al., 2020). Among the NDs, Alzheimer's disease (AD) and Parkinson's disease (PD) are two most common types that occur among the elderly (Adeli et al., 2016; Sperling et al., 2011). To date, the underlying pathogenesis of most NDs (e.g., AD and PD) remains unclear, and an effective treatment for the disease is lacking. Fortunately, previous studies demonstrate that disease-related biomarkers detected from different neuroimages can be applied to describe the structural and functional changes of the brain regions during the progression of NDs, which contributes to their early prediction and thus slows down their progression (Huang et al., 2021b; Huang et al., 2019). However, the pathological mechanisms of NDs are difficult to be interpreted using imaging biomarkers alone (Wachinger et al., 2018). Some reports indicate that NDs are linked with genetic factors (Scheltens et al., 2021). Therefore, biomarkers exploited from genetic data may provide pathological interpretation for NDs. However, directly associating genetic variants with pathological behaviors, such as disease status, may result in inaccurate or even incorrect detection results due to genes that cannot correctly encode pathological behaviors (Bi et al., 2017). To solve this problem, imaging quantitative traits (QTs) can be used as endophenotypes to construct an indirect correlation between genetic variants and pathological behaviors. Detecting NDs' biomarkers from imaging QTs and genetic variants is a potential means to gain insights into the underlying pathological mechanism of the disease and early prediction of this disease.

Recently, brain imaging genetics have been applied to analyze associations between imaging QTs and genetic variants (such as single nucleotide polymorphisms, SNPs) and detect disease-associated biomarkers. The QTs obtained by using different imaging technologies can be used to measure the brain from different perspective and might provide complementary information of the brain. For example, the structural magnetic resonance imaging (sMRI) can reveal structural abnormalities of the brain. Some changes can be observed in the sMRI of patients with NDs, such as atrophy in the hippocampus and medial temporal lobe of AD patient compared with those of normal control (NC) (Huang et al., 2015; Salvatore et al., 2014), and structural changes in middle frontal gyrus and superior

temporal gyrus of PD patients (Sheng et al., 2014). Diffusion-weighted tensor imaging (DTI) scans can detect the degeneration of the dopaminergic neurons, whose loss results in PD (Mishra et al., 2019). Moreover, the positron-emission tomography (PET) scans can capture functional abnormalities of the brain. For instance, the accumulation of β-amyloid protein in the human body can be discovered by using PET, and more β-amyloid proteins can be found in AD patients than those in NC (Jagust and Mormino, 2011; Marcus et al., 2014). Therefore, combining multimodal imaging QTs could help in effectively identifying NDs' biomarkers (i.e., QTs and SNPs).

However, multimodal imaging data have data missing problem due to imaging quality and high cost. When the missing samples are removed, and only the complete modal data are used in the analysis of imaging genetics, some useful information may be lost. Therefore, it is important to handle the missing data problem to employ more samples to train a more reliable model. Some studies (Huang et al., 2021a; Zhou et al., 2019b) demonstrated that the association between QTs and SNPs is a complex many-to-many relationship. For example, an SNP is probably associated with multiple imaging QTs or one imaging QT is associated with multiple SNPs. In this case, the association between QTs and SNPs is difficult to analyze with a linear model. Therefore, how to improve the performance of biomarker detection for NDs by solving the abovementioned problems is an attractive issue in this work.

1.1. Related works

Existing relevant works with regard to the abovementioned problems include multi-task regression, missing data handling, and imaging genetic analysis with a nonlinear model.

1.1.1. Multi-task regression

Multi-task regression for association analysis between QTs and SNPs can be applied to simultaneously identify the disease-related SNPs and QTs (Du et al., 2021; Kim et al., 2020; Wang et al., 2012). Wang et al. (Wang et al., 2012) proposed a group-sparse multi-task regression and feature selection (G-SMuRFS) method, which considered each QT as a response variable (i.e., a learning task), and formulated a multitask regression framework to identify the relationship between QTs and SNPs. Moreover, the SNP effects

of group and individual levels are taken into account in the G-SMuRFS method. Although the G-SMuRFS method can be used to simultaneously identify the disease-related SNPs and QTs, the correlation information among QTs is discarded, which may lead to the decrease in the detection performance of QTs. Du et al. (Du et al., 2021) proposed a multi-task sparse canonical correlation analysis (MTSCCA) that uses complementary information carried by different imaging multimodal data to identify bi-multivariate associations between SNPs and multimodal imaging QTs. However, the abovementioned methods are applied to separately explore the associations between SNPs and imaging QTs of each modality; these approaches ignore the underlying associations among different modalities. Therefore, modality-shared biomarkers may remain undiscovered when these methods are used. Recently, Kim et al. (Kim et al., 2020) proposed a joint connectivity-based SCCA (JCB-SCCA) for incorporating biological prior information to identify associations between SNPs and QTs, where the inter- and intra-modal information were included. Although the information of SNPs and QTs were incorporated into the SCCA-based models, and promising results were obtained by using these methods, they removed missing data and only applied complete data in the multimodal data for association analysis.

1.1.2. Missing data handling

Multimodal data have the missing data issue (i.e., not all the samples have complete multimodal data) due to imaging quality and high cost. To address this issue, two commonly used approaches were proposed in previous studies (Candès and Recht, 2009; Hastie et al., 2015; Schneider, 2001; Thung et al., 2014; Zhu et al., 2011): 1) discard the samples with missing data, or 2) impute the missing data. Most existing methods discard samples with at least one missing multimodal data and perform imaging genetic studies to identify associations between SNPs and QTs based on the remaining multimodal data. This approach discards much of the available information, which might result in the decrease in the detection performance of imaging genetics. Besides, imputation methods for missing data are also widely used, which estimate missing values based on the available data by using specific imputation techniques. Schneider et al. (Schneider, 2001) used expectation maximization (EM) algorithm to estimate the mean and the covariance matrix of an incomplete dataset and fill in missing values with imputed values. Hastie el al. (Hastie et

al., 2015) proposed a singular value decomposition (SVD) to impute missing values. Although the missing data issue can be solved by using these methods, unnecessary noise may be introduced, diminishing the detection performance (Zhou et al., 2019a). Accordingly, the correlations across multiple modalities cannot be effectively exploited by using these methods. To solve this problem, Zhou et al. (Zhou et al., 2019a) proposed a projection method based on latent representation learning to consider the associations between inter- and intra-modal data and handle missing data. The data used in this method are divided into two parts to utilize all available data. One part consists of complete modal data, which is projected into a common latent space to learn the common latent imaging representation. The other part consists of incomplete modal data, which is projected into a modality-specific latent space to learn the modality-specific latent imaging representation. Then, the common latent imaging representation is cascaded with all modality-specific latent imaging representations to form a feature representation in the latent space. The missing data issue can be solved with this procedure, and correlations among different modality data can be exploited.

1.1.3. Imaging genetic analysis with a nonlinear model

As reported in previous studies (Wang et al., 2018), analyzing the complex associations between SNPs and QTs is difficult by using a simple linear model, using the nonlinear model can alleviate such a difficulty to an extent. Kernel canonical correlation analysis (KCCA) is a typical nonlinear model. KCCA maps input features provided from multiple views into a common space, and employs the kernel method to maximally capture nonlinear associations between multiple views (Yoshida et al., 2017). However, carrying out feature selection or capturing important canonical components is difficult for KCCA, so it is mainly applied in feature fusion and classification (Yoshida et al., 2017). Additionally, additive-based methods have been presented to explore the complex nonlinear associations between SNPs and QTs (Huang et al., 2021a; Yin et al., 2012). Yin et al. (Yin et al., 2012) proposed a group sparse additive model (GroupSpAM) to identify the association between SNPs and QTs, where each additive component is a smooth function of a single SNP. Consequently, the nonlinear effect of SNP can be incorporated into the association model to improve the detection performance. However, the GroupSpAM method only uses a single QT data and rich information of multiple brain

regions is ignored. Huang et al. (Huang et al., 2021a) proposed a temporal group sparse regression and additive model (T-GSRAM) to identify disease related QTs and SNPs simultaneously. The T-GSRAM method assumed that the effects of each SNP on QT are regarded as a smooth function of time, where the smooth function can be a specified parametric form (such as a polynomial of a variable) or nonlinear transformation. However, these nonlinear image genetic studies applied only unimodal data for association analysis, and the possible association information prior to multimodal data is ignored, and could only be applied to the complete multimodal data.

Four main challenges are involved in imaging genetics. First, some traditional methods merely analyze single SNP and single QT correlation without considering the group information of SNP and correlation among QTs. Second, most existing methods often use imaging QTs extracted from a single modality, whereas some multimodal methods simply concentrate on features from each modality without regarding the inter- and intra-modal correlations. Third, the missing data issue is common in multimodal settings; hence, how to train a more reliable model by using all available subjects is highly essential. Fourth, most existing methods for association analysis of imaging genetics assume that a linear correlation exists between SNP and QT. However, the association between SNP and QT is complex, and this complex association is difficult to detect with only a simply linear model.

1.2. Overview of the proposed method

In this study, a novel structure-constrained combination-based nonlinear association analysis (ScCNAA) is proposed to analyze associations between incomplete multimodal imaging and genetic data for potential biomarker detection of NDs. The contributions of this study are three-fold.

First, for incomplete multimodal imaging data, a two-stage projection strategy was applied to explore the associations between QTs and SNPs for biomarker detection. In the first stage, a parallel concatenated projection method combined with several constraints was applied to handle missing data. Modality-shared and modality-specific information can be well captured to obtain latent imaging representations. All samples were divided into two parts: samples with complete multimodal data and those with incomplete multimodal data. The samples with complete multimodal data are used to learn common

latent imaging representations (i.e., correlation among different modalities). The samples with incomplete multimodal data were applied to learn an independent latent imaging representation (i.e., modality-specific) for each modality. The projection from different modalities to common and specific latent imaging representations is expected to efficiently exploit inter- and intra-modal interactions, respectively. In contrast with Zhou's work (Zhou et al., 2019a), we assumed that some modality-shared and modality-specific QTs and SNPs can be detected by using all modalities or specific modality data, respectively. The independent latent imaging representation of each modality is concatenated in parallel with the common latent imaging representations to generate a new latent imaging representation for each modality (Fig. 1). Subsequently, a locality preserving constraint was incorporated into the proposed method to retain structure information before and after the first stage projection. Moreover, brain connectivity can be used for imaging QTs to measure the degree of coherence or collaboration between different brain regions connected by dissecting fiber bundles or functional connections. Accordingly, a connectivity constraint was applied in the proposed method to explore the structure associations among QTs, and an l_{21} norm was added to exploit the individual information of QTs. In the second stage, the latent imaging representations were projected into an association space. After the two-stage projection in imaging QTs, SNPs were also projected into the association space by using a nonlinear kernel-based method. Therefore, the association analysis between QTs and SNPs can be performed in the association space. With the first stage of projection, modality-shared and modality-specific information can be included in the latent imaging representations. Therefore, exploring the associations between the latent imaging representations and the SNPs can be used to detect the modality-shared and modality-specific biomarkers for QTs. Moreover, the second projection of QTs and the projection of SNPs are inspired by the idea of CCA (Hardoon and Shawe-Taylor, 2011). Unlike CCA, the proposed method considered the nonlinear association between QTs and SNPs was taken into consideration, given by their complex relationship. Moreover, instead of using nonlinear projections in QTs and SNPs, only a nonlinear projection was used in SNPs in the proposed method. The reasons were two-fold: First, model complexity and parameter number would be increased considerably if nonlinear projection was used in QTs because a two-stage projection strategy was used in

QTs. Second, nonlinear projection was used in SNPs, supposing that the nonlinear association between QTs and SNPs was adequate and could be used to explore the complex association between QTs and SNPs well.

Second, for genetic data, a novel group-induced graph self-expression constraint was adopted, which mainly consists of a group and graph self-expression constraints. The graph self-expression constraint was used to exploit associations among SNPs. Given that several SNPs in a gene carried the same inherent functions, a group constraint (i.e., G_{21} norm) was applied to the proposed method, which could be introduced as group information into the graph self-expression constraint and SNPs to learn gene-gene correlations and joint effects of SNPs in a gene. Therefore, most effective components in SNPs can be detected by using the proposed group-induced graph self-expression constraint. Moreover, an l_{21} norm was used to incorporate individual level information of SNPs into the proposed method. Unlike traditional group lasso (i.e., G_{21} and l_{21} norms), relationships among all SNPs are considered and strong correlations among SNPs within a group are retained in the proposed method by using the group-induced graph self-expression constraint, which contributes to the following SNP biomarker detection.

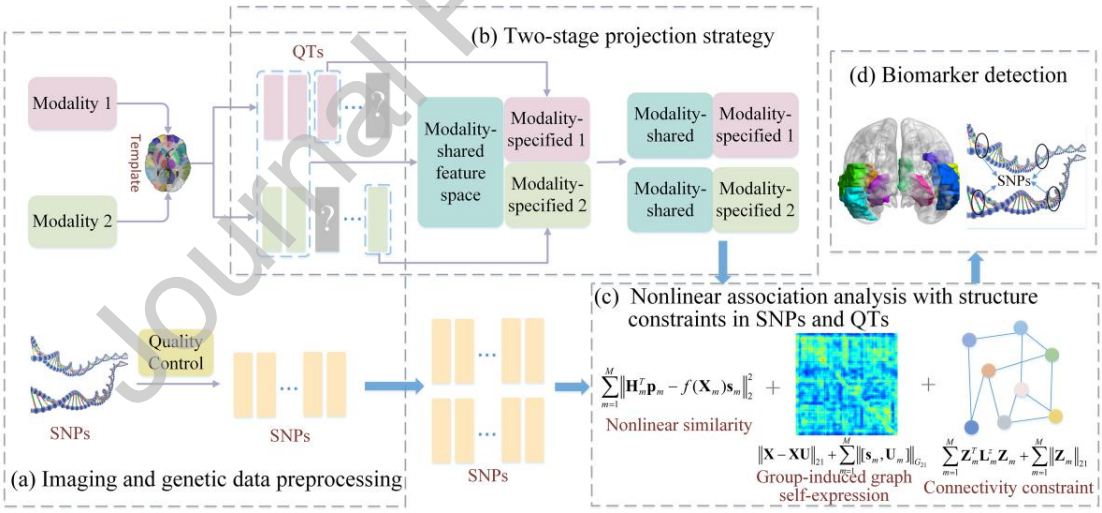


Fig. 1. Flowchart of the proposed ScCNAA

Finally, the complex association between QTs and SNPs is difficult to be effectively analyzed by using linear models. A nonlinear kernel-based method was applied in the proposed method to explore the complex association between QTs and SNPs, where the feature selection is performed via applying an l_2 norm and a kernel function to learn the

corresponding weights of SNPs. An optimization method is then applied to effectively solve the whole nonlinear model.

A set of simulation data is first applied to evaluate the performance of the proposed method, and an advanced biomarker detection accuracy is achieved. Moreover, two real datasets provided by Alzheimer's disease neuroimaging initiative (ADNI) and Parkinson's progression markers initiative (PPMI) are used to further assess the effectiveness of the proposed method. Based on previous reports, no research has combined parallel concatenated projection, self-information in a group of genetic data, and kernel-based method into imaging genetics to investigate the underlying information on incompletely multimodal imaging and genetic data and analyze the complex associations between imaging and genetic data for potential biomarker detection of NDs. The rest of this paper is organized as follows. Section 2 introduces the proposed ScCNAA. Section 3 describes the neuroimaging and SNP preprocessing steps and the simulation and real data experimental results. Section 4 provides the related discussion.

2. Methods

The proposed ScCNAA model (Fig. 1) mainly consists of four components. First, we preprocessed multimodal imaging and genetic data to obtain QTs and SNPs, respectively. Second, a two-stage projection strategy was applied in QTs. In the first stage, we used a parallel concatenated projection method to address missing data issue in the multimodal QTs. We first divided multimodal imaging QTs into two parts: complete multimodal QTs and incomplete multimodal QTs. Then, we projected the complete and incomplete multimodal QTs into latent imaging representations to obtain a common latent imaging representation and an independent latent imaging representation of each modality, respectively. The independent latent imaging representation of each modality was concatenated in parallel with the common latent imaging representations to generate a latent imaging representation for each modality. In the second stage, the latent imaging representations were projected into an association space. Third, SNPs were also projected into the association space by using a nonlinear kernel-based method. Therefore, we identified the associations between the QTs and SNPs in the association space combined with the structure constraints in SNPs and QTs. With these three steps, some

disease-related biomarkers, which includes QT and SNP biomarkers, can be finally achieved. The details of the proposed method are provided in the following subsections. Moreover, the code of ScCNAA is available at the coding sharing site (https://github.com/Meiyan88/ScCNAA).

2.1. Mathematical formulation

In this work, we write matrices, vectors and scalars as boldface uppercase, boldface lowercase, and italic letters, respectively. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be the SNPs, where n and p denote the sample number and the feature dimension of SNPs, respectively. Let $\mathbf{Y}_m \in \mathbb{R}^{q \times n_m}$ ($m = 1, \ldots, M$) represent the QTs of the *m*-th modality, where M is the modality number of imaging QTs; and n_m and q denote the sample number of the *m*-th modality QTs and the feature dimension of QTs, respectively. Moreover, $\|\mathbf{X}\|_F = \sqrt{\sum_i \sum_j x_{ij}^2}$ denotes its Frobenius norm. $\|\mathbf{X}\|_{2,1} = \sum_i \sqrt{\sum_j x_{ij}^2}$ denotes the l_{21} norm, where x_{ij} denotes the element of the *i*-th row and *j*-th column of $\mathbf{X}$, and the *i*-th row and *j*-th column of $\mathbf{X}$ are denoted as $\mathbf{x}^i$ and $\mathbf{x}_j$, respectively. The main notations used in this study are listed in Table 1.

Table 1. Main notations used in the proposed method.

Notation	Size	Description
$\mathbf{Y}_m^c$	$q \times n^c$	The imaging QTs of *m*-th modality for samples with complete multimodal QTs
$\mathbf{Y}_m^{\bar{c}}$	$q \times n_m^{\bar{c}}$	The imaging QTs of *m*-th modality for samples with incomplete multimodal QTs
$\mathbf{Z}_m$	$q \times h$	Projection matrix for the *m*-th modality QTs
$\mathbf{H}^c$	$h \times n^c$	Latent imaging representations for samples with complete multimodality QTs
$\mathbf{H}_m^{\bar{c}}$	$h \times n_m^{\bar{c}}$	Latent imaging representation of the *m*-th modality for samples with incomplete multimodality QTs
$\mathbf{E}_m$	$h \times n^c$	Sparse error matrix for the *m*-th modality QTs
$\mathbf{X}_m$	$n_m \times p$	SNP data corresponding to the *m*-th modality QTs
$\mathbf{U}_m$	$p \times p$	Graph self-expression matrix of SNP data corresponding to the *m*-th modality QTs
$\mathbf{s}_m$	$p \times 1$	Association matrix of SNP data corresponding to the *m*-th modality QT data
$\mathbf{p}_m$		Association matrix of the learned latent imaging representation of QTs
p		Feature dimension of SNPs
q		Feature dimension of QTs
n_m		Sample number of the *m*-th modality QTs
n^c		Sample number of complete multimodal QTs
$n_m^{\bar{c}}$		Sample number of the *m*-th modality for incomplete multimodal QTs.
h		Feature dimension of the latent imaging representations
M		Modality number of imaging QTs

γ, λ	Constraint parameters

2.2. ScCNAA model

A two-stage projection strategy was applied in QTs to explore the association between QTs and SNPs. In the first stage, we used a parallel concatenated projection method to address the missing data issue in the multimodal QTs. We first divided the multimodal QTs into two parts, namely, complete and incomplete multimodal data. Then, we projected complete multimodal QTs to a common latent imaging representation to learn shared features of all modalities. We also projected the remaining *m*-th modality data to modality specific latent space to learn specific features of *m*-th modality. Subsequently, the independent latent imaging representation of each modality was concatenated in parallel with the common latent imaging representation to form a new latent imaging representation for each modality. The proposed parallel concatenated projection method not only utilized all available samples to train a reliable model but also considered the inter- and intra-modal interactions. In the second stage, the latent imaging representations were projected into an association space. Afterward, the SNPs were also projected into the association space by using a nonlinear kernel-based model. Thus, the complex associations between QTs and SNPs can be well contemplated. Moreover, the structure constraints on SNPs and QTs were also added in the proposed method to incorporate structure and individual level information of the SNPs and QTs, respectively. Therefore, the proposed ScCNAA model can be defined as follows:

$$\begin{aligned}\min_{S,\mathbf{Z},\mathbf{H},\mathbf{E},P} & \frac{1}{2}\sum_{m=1}^{M}\left\|\mathbf{H}_m^T\mathbf{p}_m - f(\mathbf{X}_m)\mathbf{s}_m\right\|_2^2 + \gamma\sum_{m=1}^{M}\left\|\mathbf{E}_m\right\|_1 + \lambda\sum_{m=1}^{M} tr(\mathbf{Z}_m^T\mathbf{Y}_m\mathbf{L}_m(\mathbf{Z}_m^T\mathbf{Y}_m)^T) \\ & +\Omega(\mathbf{S}) + \Omega(\mathbf{Z}),\ \ s.t.\ \ \mathbf{Z}_m^T\mathbf{Y}_m = \mathbf{H}_m + \mathbf{E}_m, \forall m \in \{1, 2, \cdots, M\}; \mathbf{P}^T\mathbf{P} = \mathbf{I}.\end{aligned} \tag{1}$$

where $\mathbf{Y}_m = [\mathbf{Y}_m^c, \mathbf{Y}_m^{\bar{c}}] \in \mathbb{R}^{q\times(n^c+n_m^{\bar{c}})}$ denotes the imaging QTs of *m*-th modality, $n_m = n^c + n_m^{\bar{c}}$ denotes the sample number of the *m*-th modality; and n^c and $n_m^{\bar{c}}$ are the sample number of complete multimodal QTs data and incomplete *m*-th modal QTs data, respectively. $\mathbf{H}_m = [\mathbf{H}^c, \mathbf{H}_m^{\bar{c}}] \in \mathbb{R}^{h\times(n^c+n_m^{\bar{c}})}$ is the latent representation of the *m*-th modality, where *h* is the feature dimension of the latent imaging representation, $\mathbf{H}^c$ denotes common latent imaging representation for samples with complete multimodal QTs, and $\mathbf{H}_m^{\bar{c}}$ denotes the

independent latent imaging representation of the *m*-th modality for samples with incomplete multimodal QTs. $\mathbf{E}_m \in \mathbb{R}^{h \times n_m}$ is the sparse error matrix for the *m*-th modality QTs. $\mathbf{p}_m \in \mathbb{R}^{h \times 1}$ is an association matrix of the learned latent imaging representation of QTs. $\mathbf{Z}_m \in \mathbb{R}^{q \times h}$ is a projection matrix of the *m*-th modality QTs. $\mathbf{s}_m \in \mathbb{R}^{p \times 1}$ is an association matrix of SNP data corresponding to the *m*-th modality QTs. $\Omega(\mathbf{S})$ and $\Omega(\mathbf{Z})$ are the constraints for selecting relevant SNPs and imaging QTs by using prior information. f denotes nonlinear transformation to construct nonlinear associations with SNPs and QTs.

As mentioned by Zhou et al. (Zhou et al., 2018), if $\mathbf{H}^c$ is not really used in model calculation and analysis, then $\sum_{m=1}^{M} \left\| \mathbf{Z}_m^T \mathbf{Y}_m - \mathbf{H}^c \right\|_F^2$ can be simplified to $\left\| \mathbf{Z}_1^T \mathbf{Y}_1 - \mathbf{Z}_2^T \mathbf{Y}_2 \right\|_F^2$ when the $M = 2$. Specifically, $\left\| \mathbf{Z}_1^T \mathbf{Y}_1 - \mathbf{Z}_2^T \mathbf{Y}_2 \right\|_F^2$ is used to enforce the projections of each modality to as close as possible. Thus, the inter-correlations across different modalities can be captured (Shao et al., 2020). Similar to this idea, the inter-modality correlations can be exploited by projecting the complete multimodal imaging QTs into a common latent imaging representation. Moreover, a locality preserving constraint $\sum_{m=1}^{M} tr(\mathbf{Z}_m^T \mathbf{Y}_m \mathbf{L}_m (\mathbf{Z}_m^T \mathbf{Y}_m)^T)$ is added in Eq. (1) to maintain the structure information among neighborhood before and after projection, where $\mathbf{L}_m \in \mathbb{R}^{n_m \times n_m}$ $(\mathbf{L}_m = \mathbf{D}_m - \mathbf{C}_m)$ is the Laplace matrix and $\mathbf{D}_m$ is a diagonal matrix with the *i*-th diagonal element that represents the sum of the *i*-th row in $\mathbf{C}_m$. $\mathbf{C}_m$ is a similarity matrix for the *m*-th modality, whose (i, j)-th element is $\exp(-\left\| \mathbf{Y}_{m,:i} - \mathbf{Y}_{m,:j} \right\|_2^2 / \sigma)$, where $\mathbf{Y}_{m,:i}$ and $\mathbf{Y}_{m,:j}$ denotes the *i*-th column and *j*-th column of $\mathbf{Y}_m$, respectively; and $\sigma = 1$ is empirically set in this study.

2.2.1. QT constraints

Brain connectivity measures the degree of coherence or synergy between different brain regions connected by anatomical fiber bundles or functional associations. Therefore, a connectivity constraint was applied to incorporate neurological prior information, and can be formulated as follows:

$$P(\mathbf{Z}_m) = \mathbf{Z}_m^T \mathbf{L}_m^z \mathbf{Z}_m \tag{2}$$

where $\mathbf{L}_m^z \in \mathbb{R}^{q \times q}$ is the Laplacian matrix of connectivity matrix of $\mathbf{Y}_m$, which can be used to explore the structure correlations between two QT features. Moreover, Laplacian matrix can be calculated as $\mathbf{L}_m^z = \mathbf{D}_m^z - \mathbf{C}_m^z$, where $\mathbf{D}_m^z$ is a diagonal matrix with the *i*-th diagonal element that represents the sum of the *i*-th row in connectivity matrix $\mathbf{C}_m^z$. $\mathbf{C}_m^z$ can be obtained by calculating the Pearson correlation between the two features of $\mathbf{Y}_m$. Although the connectivity constraint is meaningful, there is a lack of feature selection at individual level. The model for a large number of imaging features is complex and difficult to comprehend because of the non-sparse results without feature selection. Therefore, sparse induction constraint is also necessary for imaging QTs, and an l_{21} norm was used on the imaging QTs:

$$\left\|\mathbf{Z}_m\right\|_{21} = \sum_{i=1}^{q} \sqrt{\sum_{j=1}^{h} z_{m,ij}^2} \tag{3}$$

Therefore, the constraint $\Omega(\mathbf{Z})$ of QTs can be defined as follows:

$$\Omega(\mathbf{Z}) = \sum_{m=1}^{M} \mathbf{Z}_m^T \mathbf{L}_m^z \mathbf{Z}_m + \sum_{m=1}^{M} \left\|\mathbf{Z}_m\right\|_{21} \tag{4}$$

2.2.2. SNP constraints

SNPs within a gene usually carry out the same genetic function. Moreover, linkage disequilibrium (Barrett et al., 2005) describes the non-random association between alleles at different loci, through which the SNPs in high linkage disequilibrium are linked together in meiosis. This information should be considered in a realistic modeling method. A group-induced graph self-expression constraint was applied in the proposed method to the explore correlations among the SNPs. This constraint, consisting of graph self-expression and a G_{21} norm, can be defined as follows:

$$\mathbf{P}(\mathbf{U}_m) = \sum_{m=1}^{M} \left\|\mathbf{X}_m - \mathbf{X}_m \mathbf{U}_m\right\|_{21} + \sum_{m=1}^{M} \left\|[\mathbf{s}_m, \mathbf{U}_m]\right\|_{G_{21}} = \sum_{m=1}^{M} \left\|\mathbf{X}_m - \mathbf{X}_m \mathbf{U}_m\right\|_{21} + \sum_{m=1}^{M} \sum_{g=1}^{K} \sqrt{\sum_{i \in g_k} \sum_{j=1}^{p+1} [\mathbf{s}_m, \mathbf{U}_m]_{ij}^2} \tag{5}$$

where $\sum_{m=1}^{M} \left\|\mathbf{X}_m - \mathbf{X}_m \mathbf{U}_m\right\|_{21}$ is a graph self-expression constraint, which can be used to exploit pair-wise structure correlations among SNPs. Moreover, $\mathbf{U}_m$ is a self-expression

matrix of SNP data corresponding to *m*-th modality QT data, where the value of u_{ij} (u_{ij} is the element of the *i*-th row and the *j*-th column of $\mathbf{U}_m$) measures the correlation between two samples, namely, $\mathbf{x}_i$ and $\mathbf{x}_j$. $\left\|[\mathbf{s}_m, \mathbf{U}_m]\right\|_{G_{21}}$ is a group sparsity constraint on $\mathbf{s}_m$ and $\mathbf{U}_m$ in the *m*-th modality to explore structure correlation among the inter-group of SNPs, where the SNPs are portioned into *K* groups $\mathcal{G} = \{g_k\}_{k=1}^{K}$.

Although the strong structure correlations among all SNPs was introduced in the group-induced graph self-expression constraint, not all SNPs in the group are disease-associated SNPs, which may result in a lack of feature selection at the individual level. The SNPs related to disease are almost impossible to be in the same group. Moreover, a few SNPs within a particular group may be related to QT, whereas the remaining SNPs may be unrelated. Therefore, we added an l_1 norm to select an individual feature, which is defined as follows:

$$\left\|\mathbf{U}_m\right\|_1 = \sum_{i=1}^{p}\sum_{j=1}^{p}\left|u_{m;ij}\right| \tag{6}$$

Therefore, the constraint of SNPs $\Omega(\mathbf{S})$ can be defined as:

$$\Omega(S) = \sum_{m=1}^{M}\left\|\mathbf{X}_m - \mathbf{X}_m\mathbf{U}_m\right\|_{21} + \alpha_1\sum_{m=1}^{M}\left\|[\mathbf{s}_m, \mathbf{U}_m]\right\|_{21} + \alpha_2\sum_{m=1}^{M}\left\|\mathbf{U}_m\right\|_1 \tag{7}$$

2.3. Optimization

The objective function in Eq. (1) is not jointly convex with respect to all variables, we can update each of these variables by fixing other variables iteratively. The Augmented Lagrange Multiplier (ALM) (Lin et al., 2010) was applied to solve the optimization problem in Eq. (1). Thus, the Eq. (1) can be solved by minimizing the following ALM problem

$$\begin{aligned}\mathcal{L}(\mathbf{Z}_m, \mathbf{E}_m, \mathbf{Q}_m, \mathbf{S}, \mathbf{H}_m, \mathbf{p}_m) &= \frac{1}{2}\sum_{m=1}^{M}\left\|\mathbf{H}_m^T\mathbf{p}_m - f(\mathbf{X}_m)\mathbf{s}_m\right\|_2^2 + \sum_{m=1}^{M}\left\|\mathbf{X}_m - \mathbf{X}_m\mathbf{U}_m\right\|_{21} \\ &+ \gamma\sum_{m=1}^{M}\left\|\mathbf{E}_m\right\|_1 + \lambda\sum_{m=1}^{M}tr(\mathbf{Z}_m^T\mathbf{Y}_m\mathbf{L}_m(\mathbf{Z}_m^T\mathbf{Y}_m)^T) + \alpha_1\sum_{m=1}^{M}\left\|[\mathbf{s}_m, \mathbf{U}_m]\right\|_{21} + \alpha_2\left\|\mathbf{U}\right\|_1 \\ &+ \beta_1\sum_{m=1}^{M}\left\|\mathbf{Z}_m\right\|_{21} + \beta_2\sum_{m=1}^{M}\mathbf{Z}_m^T\mathbf{L}_m^z\mathbf{Z}_m + \frac{\eta}{2}\left\|\mathbf{p}_m\right\|_F^2 + \sum_m\Phi(\mathbf{Q}_m, \mathbf{Z}_m^T\mathbf{Y}_m - \mathbf{H}_m - \mathbf{E}_m).\end{aligned} \tag{8}$$

where $\Phi(\mathbf{Q}, \Omega) = \frac{\mu}{2}\left\|\Omega\right\|_F^2 + \left\langle\mathbf{Q}, \Omega\right\rangle$, $\left\langle\cdot,\cdot\right\rangle$ denotes the matrix inner product, μ is a positive

penalty scalar, and $\mathbf{Q}_m$ is a Lagrange multiplier. Given that the form of the nonlinear transformation of SNP was unknown, we introduced the kernels to represent the nonlinear transformation. $\mathbf{s}_m$ can be rewritten as $\mathbf{s}_m = f(\mathbf{X}_m)^T \mathbf{a}_m$ with $\mathbf{a}_m \in \mathbb{R}^{n_m \times 1}$. Thus, $f(\mathbf{X}_m)\mathbf{s}_m$ can be rewritten as $\mathbf{K}_m \mathbf{a}_m$, where $\mathbf{K}_m$ denotes kernel function. Eq. (1) can be rewritten as:

$$\begin{aligned}\mathcal{L}(\mathbf{Z}_m, \mathbf{E}_m, \mathbf{Q}_m, \mathbf{S}, \mathbf{H}_m, \mathbf{p}_m) &= \frac{1}{2}\sum_{m=1}^{M}\left\|\mathbf{H}_m^T\mathbf{p}_m - \mathbf{K}_m\mathbf{a}_m\right\|_2^2 + \sum_{m=1}^{M}\left\|\mathbf{X}_m - \mathbf{X}_m\mathbf{U}_m\right\|_{21} \\ &+ \lambda\sum_{m=1}^{M} tr(\mathbf{Z}_m^T\mathbf{Y}_m\mathbf{L}_m(\mathbf{Z}_m^T\mathbf{Y}_m)^T) + \gamma\sum_{m=1}^{M}\left\|\mathbf{E}_m\right\|_1 + \alpha_1\sum_{m=1}^{M}\left\|[f(\mathbf{X}_m)^T\mathbf{s}_m, \mathbf{U}_m]\right\|_{G_{21}} + \alpha_2\sum_{m=1}^{M}\left\|\mathbf{U}_m\right\|_1 \\ &+ \beta_1\sum_{m=1}^{M}\left\|\mathbf{Z}_m\right\|_{21} + \beta_2\sum_{m=1}^{M}\mathbf{Z}_m^T\mathbf{L}_m^z\mathbf{Z}_m + \frac{\eta}{2}\left\|\mathbf{p}\right\|_F^2 + \sum_m \Phi(\mathbf{Q}_m, \mathbf{Z}_m^T\mathbf{Y}_m - \mathbf{H}_m - \mathbf{E}_m).\end{aligned} \tag{9}$$

We needed to alternately update the variables (i.e., update one variable while fixing other variables) to find the minimize $\mathcal{L}$. Thus, we decomposed the above-mentioned optimization problem into the following subproblems.

2.3.1. Optimizing Z_m

Fixed all other variables except $\mathbf{Z}_m$ in Eq. (2), the expression can be rewritten as

$$\begin{aligned}\min_{\mathbf{Z}_m} &\lambda\sum_{m=1}^{M} tr(\mathbf{Z}_m^T\mathbf{Y}_m\mathbf{L}_m(\mathbf{Z}_m^T\mathbf{Y}_m)^T) + \beta_1\sum_{m=1}^{M}\left\|\mathbf{Z}_m\right\|_{21} \\ &+ \beta_2\sum_{m=1}^{M}\mathbf{Z}_m^T\mathbf{L}_m^z\mathbf{Z}_m + \sum_m \Phi(\mathbf{Q}_m, \mathbf{Z}_m^T\mathbf{Y}_m - \mathbf{H}_m - \mathbf{E}_m).\end{aligned} \tag{10}$$

Then, Eq. (3) can be simplified to

$$\begin{aligned}\min_{\mathbf{Z}_m} &\lambda tr(\mathbf{Z}_m^T\mathbf{Y}_m\mathbf{L}_m(\mathbf{Z}_m^T\mathbf{Y}_m)^T) + \beta_1\left\|\mathbf{Z}_m\right\|_{21} \\ &+ \beta_2\mathbf{Z}_m^T\mathbf{L}_m^z\mathbf{Z}_m + \frac{\mu}{2}\left\|\mathbf{Z}_m^T\mathbf{Y}_m - \mathbf{H}_m - \mathbf{E}_m + \mathbf{Q}_m/\mu\right\|_F^2.\end{aligned} \tag{11}$$

We can solve Eq (4) by taking the derivative with respect to $\mathbf{Z}_m$ and setting it to zero. We have

$$\lambda\mathbf{Y}_m\mathbf{L}_m\mathbf{Y}_m^T\mathbf{Z}_m + \beta_1\mathbf{D}_1\mathbf{Z}_m + \beta_2\mathbf{L}_m^z\mathbf{Z}_m + \frac{\mu}{2}(\mathbf{Y}_m\mathbf{Y}_m^T\mathbf{Z}_m - \mathbf{Y}_m(\mathbf{H}_m + \mathbf{E}_m - \mathbf{Q}_m/\mu)^T) = 0. \tag{12}$$

Thus, the $\mathbf{Z}_m$ is given as

$$\mathbf{Z}_m = (\lambda\mathbf{Y}_m\mathbf{L}_m\mathbf{Y}_m^T + \beta_1\mathbf{D}_1 + \beta_2\mathbf{L}_m^z + \frac{\mu}{2}\mathbf{Y}_m\mathbf{Y}_m^T)^{-1}(\frac{\mu}{2}\mathbf{Y}_m(\mathbf{H}_m + \mathbf{E}_m - \mathbf{Q}_m/\mu)^T). \tag{13}$$

where $\mathbf{D}_1 \in \mathbb{R}^{q\times q}$ is a diagnose matrix, and its *j*-th diagonal element is $1/\left\|Z_{m,j:}\right\|_2$.

2.3.2. Optimizing $\mathbf{E}_m$

All other variables, except $\mathbf{E}_m$ in Eq. (2), are fixed, the expression can be rewritten as

$$\begin{aligned} &\min_{\mathbf{E}_m} \gamma \sum_{m=1}^{M} \left\| \mathbf{E}_m \right\|_1 + \sum_m \Phi(\mathbf{Q}_m, \mathbf{Z}_m^T \mathbf{Y}_m - \mathbf{H}_m - \mathbf{E}_m) \\ &\Leftrightarrow \min_{\mathbf{E}_m} \frac{\gamma}{\mu} \left\| \mathbf{E}_m \right\|_1 + \frac{\mu}{2} \left\| \mathbf{E}_m - (\mathbf{Z}_m^T \mathbf{Y}_m - \mathbf{H}_m + \mathbf{Q}_m / \mu) \right\|_F^2 \end{aligned} \tag{14}$$

We can obtain the optimal $\mathbf{E}_m$ by

$$\mathbf{E}_m = S_{\frac{\gamma}{\mu}}[\mathbf{Z}_m^T \mathbf{Y}_m - \mathbf{H}_m + \mathbf{Q}_m / \mu] \tag{15}$$

where $S_{\frac{\gamma}{\mu}}$ is a soft threshold operator (Lin et al., 2010), and it can be defined as

$$S_{\frac{\gamma}{\mu}}[e] = \begin{cases} e - \gamma/\mu & if \ e > \gamma/\mu \\ e + \gamma/\mu & if \ e < \gamma/\mu \\ 0 & otherwise \end{cases} \tag{16}$$

2.3.3. Optimizing $\mathbf{H}_m^{\bar{c}}$

Fixed all other variables except $\mathbf{H}_m^{\bar{c}}$ in Eq. (2), Eq. (2) can be rewritten as

$$\min_{\mathbf{H}_m^{\bar{c}}} \frac{1}{2} \sum_{m=1}^{M} \left\| \mathbf{H}_m^{\bar{c}T} \mathbf{p}_m - \mathbf{K}_m^{\bar{c}} \mathbf{a}_m^{\bar{c}} \right\|_2^2 + \sum_m \Phi(\mathbf{Q}_m^{\bar{c}}, \mathbf{Z}_m^T \mathbf{Y}_m^{\bar{c}} - \mathbf{H}_m^{\bar{c}} - \mathbf{E}_m^{\bar{c}}) \tag{17}$$

We can obtain the optimal $\mathbf{H}_m^{\bar{c}}$ by

$$\mathbf{p}_m \mathbf{p}_m^T \mathbf{H}_m^{\bar{c}} - \mathbf{p}_m (\mathbf{K}_m^{\bar{c}} \mathbf{a}_m^{\bar{c}})^T + \mu \mathbf{H}_m^{\bar{c}} - \mu (\mathbf{Z}_m^T \mathbf{Y}_m^{\bar{c}} - \mathbf{E}_m^{\bar{c}} + \mathbf{Q}_m^{\bar{c}} / \mu) = 0 \tag{18}$$

Thus, the $\mathbf{H}_m^{\bar{c}}$ is given as

$$\mathbf{H}_m^{\bar{c}} = (\mathbf{p}_m \mathbf{p}_m^T + \mu \mathbf{I})^{-1} (P_m (\mathbf{K}_m^{\bar{c}} \mathbf{a}_m^{\bar{c}})^T + \mu (\mathbf{Z}_m^T \mathbf{Y}_m^{\bar{c}} - \mathbf{E}_m^{\bar{c}} + \mathbf{Q}_m^{\bar{c}} / \mu)) \tag{19}$$

2.3.4. Optimizing $\mathbf{H}^c$

Fixed all other variables except $\mathbf{H}^c$ in Eq. (2), the expression can be rewritten as follows:

$$\begin{aligned} &\min_{\mathbf{H}^c} \frac{1}{2} \sum_{m=1}^{M} \left\| \mathbf{H}^{cT} \mathbf{p}_m - \mathbf{K}_m^c \mathbf{a}_m^c \right\|_2^2 + \sum_m \Phi(\mathbf{Q}_m^c, \mathbf{Z}_m^T \mathbf{Y}_m^c - \mathbf{H}^c - \mathbf{E}_m^c) \\ &\Leftrightarrow \min_{\mathbf{H}^c} \frac{1}{2} \sum_{m=1}^{M} \left\| \mathbf{H}^{cT} \mathbf{p}_m - \mathbf{K}_m^c \mathbf{a}_m^c \right\|_2^2 + \frac{\mu}{2} \sum_{m=1}^{M} \left\| \mathbf{Z}_m^T \mathbf{Y}_m^c - \mathbf{H}^c - \mathbf{E}_m^c + \mathbf{Q}_m^c / \mu \right\|_F^2 \end{aligned} \tag{20}$$

The close-form solution is given as

$$\mathbf{H}^{c} = (\sum_{m} \mathbf{p}_{m}\mathbf{p}_{m}^{T} + \mu \sum_{m} \mathbf{I})^{-1} \sum_{m} (\mathbf{p}_{m}(\mathbf{K}_{m}^{c}\mathbf{a}_{m}^{c})^{T} + \mu(\mathbf{Z}_{m}^{T}\mathbf{Y}_{m}^{c} - \mathbf{E}_{m}^{c} + \mathbf{Q}_{m}^{c}/\mu)) \tag{21}$$

2.3.5. Optimizing P

Fixed all other variables except $\mathbf{P}$ in Eq. (2), Eq. (2) can be rewritten as

$$\min_{P} \frac{1}{2}\sum_{m=1}^{M} \left\|\mathbf{H}_{m}^{T}\mathbf{p}_{m} - \mathbf{K}_{m}^{c}\mathbf{a}_{m}^{c}\right\|_{2}^{2} + \frac{\eta}{2}\left\|\mathbf{P}\right\|_{F}^{2} \tag{22}$$

Hence, the close-form solution is given as

$$\mathbf{p}_{m} = (\mathbf{H}_{m}\mathbf{H}_{m}^{T} + \eta\mathbf{I})^{-1}\mathbf{H}_{m}\mathbf{K}_{m}^{c}\mathbf{a}_{m}^{c} \tag{23}$$

2.3.6. Optimizing A

Fixed all the other variables except **S** in Eq. (2), the expression can be rewritten as

$$\begin{aligned} &\min_{\mathbf{s}} \frac{1}{2}\sum_{m=1}^{M} \left\|\mathbf{H}_{m}^{T}\mathbf{p}_{m} - f(\mathbf{X}_{m})\mathbf{s}_{m}\right\|_{2}^{2} + \alpha_{1}\sum_{m=1}^{M}\left\|[\mathbf{s}_{m}, \mathbf{U}_{m}]\right\|_{G_{21}} \\ &\Leftrightarrow \min_{\mathbf{s}} \frac{1}{2}\left\|\mathbf{H}_{m}^{T}\mathbf{p}_{m} - f(\mathbf{X}_{m})\mathbf{s}_{m}\right\|_{2}^{2} + \alpha_{1}tr(\mathbf{s}_{m}^{T}\mathbf{F}\mathbf{s}_{m}) \end{aligned} \tag{24}$$

Let $\mathbf{s}_{m}^{'} = \mathbf{F}^{1/2}\mathbf{s}_{m}$, Eq. (17) can be rewritten as

$$\min_{\mathbf{s}} \frac{1}{2}\left\|\mathbf{H}_{m}^{T}\mathbf{p}_{m} - f(\mathbf{X}_{m})\mathbf{F}^{-1/2}\mathbf{s}_{m}^{'}\right\|_{2}^{2} + \alpha_{1}tr(\mathbf{s}_{m}^{'T}\mathbf{s}_{m}^{'}) \tag{25}$$

Then, let $\mathbf{s}_{m}^{'} = f(\mathbf{X}_{m})^{T}\mathbf{a}_{m}$, Eq. (18) can be rewritten as

$$\min_{A} \frac{1}{2}\left\|\mathbf{H}_{m}^{T}\mathbf{p}_{m} - \mathbf{K}_{m}\mathbf{a}_{m}\right\|_{2}^{2} + \alpha_{1}tr(\mathbf{K}_{m}^{T}\mathbf{a}_{m}\mathbf{K}_{m}) \tag{26}$$

Therefore, the close-form solution is given as

$$\mathbf{a}_{m} = (\mathbf{K}_{m} + \alpha_{1}\mathbf{I})^{-1}\mathbf{H}_{m}\mathbf{p}_{m} \tag{27}$$

2.3.7. Optimizing $\mathbf{U}_m$

Fixed all the other variables except $\mathbf{U}_{m}$ in Eq. (2), Eq. (2) can be rewritten as

$$\min_{\mathbf{u}_{m}} \sum_{m=1}^{M}\left\|\mathbf{X}_{m} - \mathbf{X}_{m}\mathbf{U}_{m}\right\|_{21} + \alpha_{1}\sum_{m=1}^{M}\left\|[f(\mathbf{X}_{m})^{T}\mathbf{s}_{m}, \mathbf{U}_{m}]\right\|_{G_{21}} + \alpha_{2}\sum_{m=1}^{M}\left\|\mathbf{U}_{m}\right\|_{1} \tag{28}$$

The close-form solution is given as

$$\mathbf{U}_{m} = (\mathbf{X}_{m}^{T}\mathbf{D}_{2}\mathbf{X}_{m} + \alpha_{1}\bar{\mathbf{D}}_{2} + \alpha_{2}\tilde{\mathbf{D}}_{2})^{-1}\mathbf{X}_{m}^{T}\mathbf{D}_{2}\mathbf{X}_{m} \tag{29}$$

where $\mathbf{D}_2 \in \mathbb{R}^{n\times n}$ is a diagnose matrix, and its *i*-th diagonal element is $1/\left\|\mathbf{X}_{m;i} - \mathbf{X}_{m;i}\mathbf{U}_m\right\|_2$. $\bar{\mathbf{D}}_2$ is a block diagonal matrix with the *k*-th block being $\mathbf{I}_k/2\left\|\mathbf{U}_{m;k}\right\|_F$, and $\mathbf{I}_k$ is an identify matrix that has the same size as the *k*-th group. The grouping information can be given on the basis of genes. $\tilde{\mathbf{D}}_2$ is diagonal matrix, where $1/\left|u_{ij}\right|$ is the *j*-th diagonal element in $\mathbf{D}_2$.

2.3.8. Optimizing Q_m

The multipliers $\mathbf{Q}_m (m = 1, 2, \cdots, M)$ can be updated by

$$\mathbf{Q}_m := \mathbf{Q}_m + \mu(\mathbf{Z}_m^T\mathbf{Y}_m - \mathbf{H}_m - \mathbf{E}_m) \tag{30}$$

The detailed steps for optimizing the objective function described in Eq. (1) are summarized in Algorithm 1.

Algorithm 1.

1.**Input:** $\mathbf{X} \in R^{n\times p}$, $\mathbf{Y}_m \in R^{q\times n_m}$, $m = 1, \cdots, M$, $\lambda, \gamma, \alpha_1, \alpha_2, \beta_1, \beta_2, \eta, h$

2. **Initialize:** Initialize $\mathbf{Z}_m, \mathbf{E}_m, \mathbf{Q}_m, \mathbf{s}_m, \mathbf{H}_m, \mathbf{P}, \mathbf{a}_m$, $\varepsilon = 10^{-5}$, $\mu = 10^{-4}$, $\max_\mu = 10^6$

3. **While** *not convergence* **do**

4. Update $\mathbf{Z}_m$ according to Eq. (13);

5. Update $\mathbf{E}_m$ according to Eq. (15);

6. Update $\mathbf{H}_m^{\bar{c}}$ according to Eq. (19);

7. Update $\mathbf{H}^c$ according to Eq. (21);

8. Update $\mathbf{p}_m$ according to Eq. (23);

9. Update $\mathbf{a}_m$ according to Eq. (27);

10. Update $\mathbf{U}_m$ according to Eq. (29);

11. Update $\mathbf{Q}_m$ according to Eq. (30);

11. Update the parameter μ via $\mu = \min(\rho\mu, \max_\mu)$;

12. **end while**

Output: $\mathbf{Z}_m, \mathbf{E}_m, \mathbf{Q}_m, \mathbf{a}_m, \mathbf{H}_m, \mathbf{P}$

2.4. Convergence analysis

We have the following theorems regarding the ScCNAA algorithm. We first introduced Lemma 1 and 2 described by (Du et al., 2016; Nie et al., 2010).

Lemma1: For the following inequality that holds for two nonzero vectors $\bar{\mathbf{z}}$ and $\mathbf{z}$, we have

$$\left\|\bar{\mathbf{z}}\right\|_2 - \frac{\left\|\bar{\mathbf{z}}\right\|_2^2}{2\left\|\mathbf{z}\right\|_2} \le \left\|\mathbf{z}\right\|_2 - \frac{\left\|\mathbf{z}\right\|_2^2}{2\left\|\mathbf{z}\right\|_2} \tag{31}$$

Lemma2: For any real number $\bar{z}$ and any nonzero real number z, we have

$$\left\|\bar{z}\right\|_1 - \frac{\left\|\bar{z}\right\|_1^2}{2\left\|z\right\|_1} \le \left\|z\right\|_1 - \frac{\left\|z\right\|_1^2}{2\left\|z\right\|_1} \tag{32}$$

Proof: The proof is obvious. Given Lemma 1, $\left\|\bar{z}\right\|_1 = \left\|\bar{z}\right\|_2$ and $\left\|z\right\|_1 = \left\|z\right\|_2$.

1) Theorem 1: Algorithm 1 decreases the objective value in each iteration in ScCNAA.

Proof: We denote the updated $\mathbf{Z}_m$, $\mathbf{E}_m$, $\mathbf{Q}_m$, $\mathbf{a}_m$, $\mathbf{H}_m$, $\mathbf{U}_m$ and $\mathbf{P}$ as $\bar{\mathbf{Z}}_m$, $\bar{\mathbf{E}}_m$, $\bar{\mathbf{Q}}_m$, $\bar{\mathbf{a}}_m$, $\bar{\mathbf{H}}_m$, $\bar{\mathbf{U}}_m$ and $\bar{\mathbf{P}}$. We first proved that the objective decreases after updating $\mathbf{Z}_m$. Thus, we should prove that the following inequation is satisfied:

$$\begin{aligned}
&\lambda\sum_{m=1}^{M} tr(\bar{\mathbf{Z}}_m^T\mathbf{Y}_m\mathbf{L}_m(\bar{\mathbf{Z}}_m^T\mathbf{Y}_m)^T) + \beta_1\sum_{m=1}^{M} tr(\bar{\mathbf{Z}}_m^T\mathbf{D}_1\bar{\mathbf{Z}}_m)\\
&+\beta_2\sum_{m=1}^{M}\bar{\mathbf{Z}}_m^T\mathbf{L}_m^z\bar{\mathbf{Z}}_m + \frac{\mu}{2}\left\|\bar{\mathbf{Z}}_m^T\mathbf{Y}_m - \mathbf{H}_m - \mathbf{E}_m + \mathbf{Q}_m/\mu\right\|_F^2\\
&\le \lambda\sum_{m=1}^{M} tr(\mathbf{Z}_m^T\mathbf{Y}_m\mathbf{L}_m(\mathbf{Z}_m^T\mathbf{Y}_m)^T) + \beta_1\sum_{m=1}^{M} tr(\mathbf{Z}_m^T\mathbf{D}_1\mathbf{Z}_m)\\
&+\beta_2\sum_{m=1}^{M}\mathbf{Z}_m^T\mathbf{L}_m^z\mathbf{Z}_m + \frac{\mu}{2}\left\|\mathbf{Z}_m^T\mathbf{Y}_m - \mathbf{H}_m - \mathbf{E}_m + \mathbf{Q}_m/\mu\right\|_F^2
\end{aligned} \tag{33}$$

Based on definitions of $\mathbf{D}_1$, Eq. (33) can be rewritten as

$$\begin{aligned}
&\lambda\sum_{m=1}^{M} tr(\bar{\mathbf{Z}}_m^T\mathbf{Y}_m\mathbf{L}_m(\bar{\mathbf{Z}}_m^T\mathbf{Y}_m)^T) + \beta_1\sum_{m=1}^{M}\sum_{i=1}^{q}\frac{\left\|\bar{\mathbf{Z}}_m\right\|_2^2}{2\left\|\mathbf{Z}_m\right\|_2}\\
&+\beta_2\sum_{m=1}^{M}\bar{\mathbf{Z}}_m^T\mathbf{L}_m^z\bar{\mathbf{Z}}_m + \frac{\mu}{2}\left\|\bar{\mathbf{Z}}_m^T\mathbf{Y}_m - \mathbf{H}_m - \mathbf{E}_m + \mathbf{Q}_m/\mu\right\|_F^2\\
&\le \lambda\sum_{m=1}^{M} tr(\mathbf{Z}_m^T\mathbf{Y}_m\mathbf{L}_m(\mathbf{Z}_m^T\mathbf{Y}_m)^T) + \beta_1\sum_{m=1}^{M}\sum_{i=1}^{q}\frac{\left\|\mathbf{Z}_m\right\|_2^2}{2\left\|\mathbf{Z}_m\right\|_2}\\
&+\beta_2\sum_{m=1}^{M}\mathbf{Z}_m^T\mathbf{L}_m^z\mathbf{Z}_m + \frac{\mu}{2}\left\|\mathbf{Z}_m^T\mathbf{Y}_m - \mathbf{H}_m - \mathbf{E}_m + \mathbf{Q}_m/\mu\right\|_F^2
\end{aligned} \tag{34}$$

Applying Lemma 1 twice to Eq. (34), we have

$$\begin{aligned}
&\lambda \sum_{m=1}^{M} tr(\bar{\mathbf{Z}}_m^T \mathbf{Y}_m \mathbf{L}_m (\bar{\mathbf{Z}}_m^T \mathbf{Y}_m)^T) + \beta_1 \sum_{m=1}^{M} \left\| \bar{\mathbf{Z}}_m \right\|_{21} \\
&+ \beta_2 \sum_{m=1}^{M} \bar{\mathbf{Z}}_m^T \mathbf{L}_m^z \bar{\mathbf{Z}}_m + \frac{\mu}{2} \left\| \bar{\mathbf{Z}}_m^T \mathbf{Y}_m - \mathbf{H}_m - \mathbf{E}_m + \mathbf{Q}_m / \mu \right\|_F^2 \\
&\le \lambda \sum_{m=1}^{M} tr(\mathbf{Z}_m^T \mathbf{Y}_m \mathbf{L}_m (\mathbf{Z}_m^T \mathbf{Y}_m)^T) + \beta_1 \sum_{m=1}^{M} \left\| \mathbf{Z}_m \right\|_{21} \\
&+ \beta_2 \sum_{m=1}^{M} \mathbf{Z}_m^T \mathbf{L}_m^z \mathbf{Z}_m + \frac{\mu}{2} \left\| \mathbf{Z}_m^T \mathbf{Y}_m - \mathbf{H}_m - \mathbf{E}_m + \mathbf{Q}_m / \mu \right\|_F^2
\end{aligned} \tag{35}$$

Therefore, the objective value decreases when $\mathbf{Z}_m$ is updated.

2) Theorem2: We can also prove that the objective value decreases in each iteration when $\mathbf{E}_m$ is updated, i.e., we should prove that the following inequation is satisfied:

$$\begin{aligned}
&\frac{\gamma}{\mu} \sum_{i=1}^{h} \frac{\left\| \bar{\mathbf{e}}_m \right\|_1^2}{\left\| \mathbf{e}_m \right\|_1} + \frac{\mu}{2} \left\| \bar{\mathbf{E}}_m - (\mathbf{Z}_m^T \mathbf{Y}_m - \mathbf{H}_m + \mathbf{Q}_m / \mu) \right\|_F^2 \\
&\le \frac{\gamma}{\mu} \sum_{i=1}^{h} \frac{\left\| \mathbf{e}_m \right\|_1^2}{\left\| \mathbf{e}_m \right\|_1} + \frac{\mu}{2} \left\| \mathbf{E}_m - (\mathbf{Z}_m^T \mathbf{Y}_m - \mathbf{H}_m + \mathbf{Q}_m / \mu) \right\|_F^2
\end{aligned} \tag{36}$$

We applied Lemma 2 twice to Eq. (36); Thus, we have

$$\begin{aligned}
&\frac{\gamma}{\mu} \left\| \bar{\mathbf{E}}_m \right\|_1 + \frac{\mu}{2} \left\| \bar{\mathbf{E}}_m - (\mathbf{Z}_m^T \mathbf{Y}_m - \mathbf{H}_m + \mathbf{Q}_m / \mu) \right\|_F^2 \\
&\le \frac{\gamma}{\mu} \left\| \mathbf{E}_m \right\|_1 + \frac{\mu}{2} \left\| \mathbf{E}_m - (\mathbf{Z}_m^T \mathbf{Y}_m - \mathbf{H}_m + \mathbf{Q}_m / \mu) \right\|_F^2
\end{aligned} \tag{37}$$

Therefore, the objective value decreases when $\mathbf{E}_m$ is updated.

3) Similarly, we can prove that the objective also decreases with each update of $\mathbf{Q}_m$, $\mathbf{a}_m$, $\mathbf{H}_m$, $\mathbf{U}_m$, and $\mathbf{P}$.

The proof is completed by combining conclusions 1), 2), and 3).

3. Experiments

3.1. Simulation study

3.1.1. Experimental setup

We generated two sets of simulation data by using different methods to evaluate the performance of our proposed ScCNAA (e.g., linear and nonlinear associations between SNPs and QTs). First, linkage disequilibrium (LD) blocks defined by Haploview (Barrett

et al., 2005) and PLINK (Purcell et al., 2007) are applied to generate SNP sets. To calculate the LD blocks, *n* subjects were simulated randomly by combining the haplotypes of HapMap CEU subjects. Thus, the LD blocks could be determined by using PLINK based on these subjects. Then, we randomly selected 100 blocks, and combined the haplotype of HapMap CEU subjects in each block to form genotype variables for these subjects. We randomly selected 10 SNPs in each block and had 1000 SNPs for each subject. Subsequently, we applied the following two cases to generate **Y**:

Case 1:

$$\mathbf{Y}\mathbf{V} = \mathbf{X}\mathbf{S} + \xi \tag{38}$$

Case 2:

$$\mathbf{Y}\mathbf{V} = e^{(\mathbf{X})}\mathbf{S} + \xi \tag{39}$$

where $\xi \sim N(0, \sigma^2)$ is the error term and X is the simulated SNP data as described above. We randomly generated sparse matrices $\mathbf{V} \in \mathbb{R}^{q \times M}$ and $\mathbf{S} \in \mathbb{R}^{p \times M}$ (i.e., the disease-related QTs or SNPs were set to nonzero values, whereas the others were set to zeros), where *p* and *q* are the feature dimensions of SNP and QT with values of 1000 and 500, respectively. The modality number *M* is set to two. Afterward, we simulated **Y** by using Eqs. (38) and (39).

The constraint parameters for each method should be fine-tuned during experiments. In this work, we employed nested five-fold cross-validation method to select the optimal parameters in different methods. In particular, the parameters in the inner loop that generate the minimum mean root mean square error (RMSE) values will be selected as the optimal parameters (i.e., $\mathrm{RMSE} = \frac{1}{5}\sum_{i=1}^{5}\sum_{m=1}^{M}\sqrt{(\mathbf{Y}_{i;m}\mathbf{v}_{i;m} - \mathbf{X}_{i;m}\mathbf{s}_{i;m})^2}$, where $\mathbf{X}_{i;m}$ and $\mathbf{Y}_{i;m}$ are the *i*-th validation sets in the inner loop). Then, the external loop calculates the final results based on the optimal parameters obtained from the inner loop. The ScCNAA method would take a significant amount of time if we simultaneously tune eight hyper-parameters (i.e., $\lambda, \gamma, \eta, \alpha_1, \alpha_2, \beta_1, \beta_2$, and h). In this case, we tuned parameters step by step. At each time, two parameters were tuned while the other six parameters were fixed. We tuned $\lambda, \gamma, \eta, \alpha_1, \alpha_2, \beta_1$ and β_2 from a moderate interval $10^i (i = -5, -4, \cdots, 4, 5)$, and h from [1, 10, 50, 80, 100, 150, 200, ⋯, 450, 500]. During the experiments, the ALM algorithm will be stopped when $\max|\mathbf{V}^{iter+1} - \mathbf{V}^{iter}| \leq \varepsilon$ and

$\max|\mathbf{s}^{iter+1} - \mathbf{s}^{iter}| \leq \varepsilon$ are satisfied, where ε is the predefined tolerable error and set to 10^{-5} empirically. In the simulation study, the area under the curve (AUC) was used as the quantitative metric to assess the biomarker detection performance in different methods. The AUC can be obtained by calculating the area under of the receiver operating characteristic curve, which uses the values of false positive rate (FPR) as the *x*-axis and values of true positive rate (TPR) as the *y*-axis. Moreover, the TPR and FPR can be defined as follows:

$$TPR = \frac{TP}{TP + FN} \tag{40}$$

$$FPR = \frac{FP}{TN + FP} \tag{41}$$

where TP, FP, TN, and FN are true positive, false positive, true negative, and false negative, respectively.

We conducted five sets of experiments on the simulation data to evaluate fully the performance of our proposed ScCNAA. In these experiments, LD block information was applied to define SNP groups on the simulation data. We initially set *M*, *n*, *p*, and *q* to 2, 500, 1000, and 500, respectively. The numbers of the missing samples of each modality was 10% of all sample number unless otherwise specified. In the first set of experiments, we varied parameters with different values to evaluate the influence of different parameters on the proposed method. In the second set of experiments, we removed one of the constraints at each time to investigate the effectiveness of using different constraints on ScCNAA. In the third set of experiments, two experiments were performed to investigate the advantages of the proposed method in the handling missing data issue. First, we varied the number of missing samples in different modalities from 10%, 20%, and 30% of the total samples to assess the effects of missing samples with different number on ScCNAA. Second, we compared the ScCNAA with the following four missing data handling methods: 1) Zero imputation, in which the missing values were set to zeros after all the features are normalized. 2) EM imputation. 3) SVD imputation. 4) Protection method proposed in (Zhou et al., 2019a). In the fourth set of experiments, we performed the first and the second stage projection separately to assess the effectiveness of the two-stage projection strategy used in QTs. From the first to fourth sets of experiments, we generated

simulation data by using a linear model in Case 1. In the fifth set of experiments, we fully compared the proposed ScCNAA method with three state-of-the-art methods, namely, G-SMuRFS, MTSCCA, and JCB-SCCA, on the two datasets generated by Cases 1 and 2, respectively. Moreover, the objective functions of these three comparison methods and our proposed method are provided in Table 2.

Table 2. Objective function of all comparison methods.

Method	Objective functions
G-SMuRFS	$\min_{\mathbf{W}} \sum_{i=1}^{n} \left\| \mathbf{W}^T \mathbf{X} - \mathbf{Y} \right\|_F^2 + \gamma_1 \left\| \mathbf{W} \right\|_{G_{2,1}} + \gamma_2 \left\| \mathbf{W} \right\|_{2,1}$
MTSCCA	$\min_{\mathbf{u}_j, \mathbf{v}_j} \sum_{m=1}^{M} -\mathbf{u}_m^T \mathbf{X}^T \mathbf{Y}_m \mathbf{v}_m \quad s.t. \left\| \mathbf{X} \mathbf{u}_m \right\|_2^2 = 1, \left\| \mathbf{Y}_m \mathbf{v}_m \right\|_2^2 = 1, \left\| \mathbf{U} \right\|_{G_{2,1}} \le a, \left\| \mathbf{V} \right\|_{2,1} \le b, \forall m.$
JCB-SCCA	$\min_{u,v} -\sum_{m=1}^{M} \frac{1}{2} \mathbf{u}^T \mathbf{X}^T \mathbf{Y}_m \mathbf{v}_m + \frac{\lambda_1}{2} \mathbf{u}^T \mathbf{L}_u \mathbf{u} + \beta_1 \left\| \mathbf{u} \right\|_1 + \frac{\lambda_2}{2} \sum_{m=1}^{M} \mathbf{v}_m^T \mathbf{L}_v^m \mathbf{v}_m + \beta_1 \sum_{m=1}^{M} \left\| \mathbf{v}_m \right\|_1 + \tau \sum_{m<m'} \left\| \mathbf{v}_m - \mathbf{v}_{m'} \right\| \quad s.t. \left\| \mathbf{u} \right\|_2^2 \le 1, \left\| \mathbf{V} \right\|_F^2 \le 1$
ScCNAA	$\min_{S,\mathbf{Z},\mathbf{H},\mathbf{E},P} \frac{1}{2} \sum_{m=1}^{M} \left\| \mathbf{H}_m^T \mathbf{p}_m - f(X_m) \mathbf{s}_m \right\|_2^2 + \gamma \sum_{m=1}^{M} \left\| \mathbf{E}_m \right\|_1 + \sum_{m=1}^{M} \mathbf{Z}_m^T \mathbf{L}_m^z \mathbf{Z}_m + \lambda \sum_{m=1}^{M} tr(\mathbf{Z}_m^T \mathbf{Y}_m \mathbf{L}_m (\mathbf{Z}_m^T \mathbf{Y}_m)^T) + \sum_{m=1}^{M} \left\| \mathbf{Z}_m \right\|_{21} + \sum_{m=1}^{M} \left\| \mathbf{X}_m - \mathbf{X}_m \mathbf{U}_m \right\|_{21} + \alpha_2 \sum_{m=1}^{M} \left\| \mathbf{U}_m \right\|_1 + \alpha_1 \sum_{m=1}^{M} \left\| [s_m, \mathbf{U}_m] \right\|_{21} \quad s.t. \ \mathbf{Z}_m^T \mathbf{Y}_m = \mathbf{H}_m + \mathbf{E}_m, \forall m \in \{1, 2, \cdots, M\}; \mathbf{P}^T \mathbf{P} = \mathbf{I}$

3.1.2. Influence of parameters on the simulation data

In this section, we studied the influence of hyper-parameters (i.e., λ, γ, η, α_1, α_2, β_1, β_2, and h) on the proposed method. Specially, we set the values of λ, γ, η, α_1, α_2, β_1, and β_2 in the range of a moderate interval 10^i $(i = -5, -4, \cdots, 0, \cdots, 4, 5)$ and h in the range of (1, 10, 50, 80, 100, 150, 200, ⋯, 450, 500). Two parameters were tuned, while the other six parameters were fixed. Fig. 2 shows the AUC of SNPs achieved by our proposed method with different parameters in Case 1. Fig. 2 demonstrates that the AUC values fluctuated greatly when fixing six parameters and tuning α_1, α_2 or β_1, β_2. The AUC values of SNPs slightly varied when fixing the six parameters and tuning λ and h. The experimental results demonstrate that the proposed method obtains better detection performance when

the values of $\alpha_1, \alpha_2, \beta_1,$ and β_2 fell in [10^1, 10^3], [10^5, 10^{-3}], [10^2, 10^3], and [10^4, 10^5], respectively.

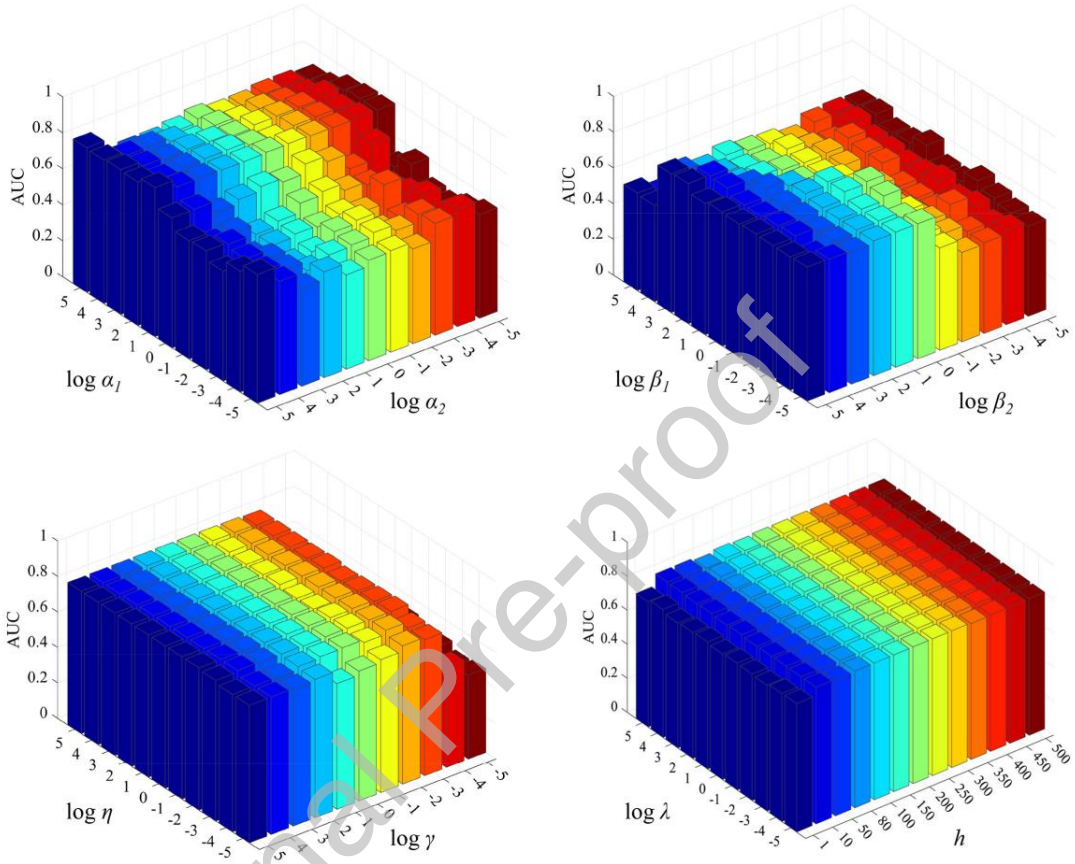


Fig. 2. The AUC results of different parameter setting (i.e., $\lambda, \gamma, \eta, \alpha_1, \alpha_2, \beta_1$ and β_2) in ScCNAA.

3.1.3. Effectiveness of the constraints

To assess the effectiveness of using different constraints, we implemented and compared different variants of our proposed method. The experimental results are shown in Table 3. The following observations are obtained: 1) The detection AUC values of SNPs and QTs decrease when the graph self-expression constraint is discarded, indicating that the correlation included among the intra-group of SNPs can help improve the detection performance on SNPs and QTs. 2) The AUC values of SNP and QT detection decrease when the G_{21} norm is removed, implying that the correlation among the inter-group of SNPs plays an important role in detecting SNPs and QTs. 3) The AUC values of QT detection fluctuate when connectivity penalty and l_{21} norms are removed, indicating that

considering structure correlations among QTs and individual level information of QTs can help induce considerable information to improve detection the performance on QTs. 4) Our proposed method outperforms all variants of the proposed method, indicating that the constraints utilized in our method are useful for detecting SNPs and QTs.

Table 3. The AUC of removed one of the constraints at each time in ScCNAA.

Constraint of SNPs			Constraint of QTs			AUC	
Graph self-expression	G_{21}	l_{21}	connectivity	l_{21}	Local preserving constraint	SNPs	QTs
--	√	√	√	√	√	0.80±0.02	0.75±0.12
√	--	√	√	√	√	0.72±0.06	0.84±0.01
√	√	--	√	√	√	0.80±0.02	0.97±0.01
√	√	√	--	√	√	0.75±0.02	0.52±0.04
√	√	√	√	--	√	0.71±0.06	0.62±0.03
√	√	√	√	√	--	0.81±0.02	0.95±0.08
√	√	√	√	√	√	**0.83±0.02**	**0.99±0.01**

3.1.4. Effectiveness of missing data handling

Table 4 lists the AUC values of ScCNAA in different numbers of missing samples. Table 4 shows that a slight fluctuation is observed in the AUC values of SNP and QT detection in different numbers of missing samples, demonstrating the stability of the proposed method in handling missing data with different numbers. Moreover, we compared the proposed method with the four missing data handling methods and report the AUC results in Table 5. Table 5 illustrates that our proposed method outperforms four missing data handling methods. On the one hand, this finding is obtained probably because our proposed method does not involve any imputation and avoid imputation errors that may affect detection performance based on imputation methods. On the other hand, this situation may also result from the use of latent imaging representations when performing a correlation analysis, which includes minimal noisy information. Moreover, inter- and intra-modality correlations are considered. In addition, we performed paired *t*-test between the proposed method and other comparison methods. The results indicate that the AUC values between the ScCNAA and all compared methods are significantly different.

Table 4. AUC values of SNPs and QTs under different missing samples in the ScCNAA.

AUC	10%	20%	30%
SNP	**0.83±0.02**	0.82±0.01	0.81±0.01
QT	**0.99±0.01**	0.97±0.01	0.97±0.01

Table 5. AUC of different methods and their *p*-values of paired *t*-tests comparing the different missing data handling methods with ScCNAA for SNPs and QTs.

AUC	Zeros	EM	SVD	Zhou	ScCNAA
SNP	0.5±0.20	0.62±0.17	0.56±0.02	0.77±0.02	**0.83±0.02**
p-value	0.006	0.029	0.002	0.001	--
QT	0.51±0.34	0.91±0.07	0.89±0.01	0.96±0.01	**0.99±0.01**
p-value	0.014	0.040	0.006	0.040	--

3.1.5. Effectiveness of two-stage projection strategy

In this section, the first and the second stage projection used in QTs were conducted separately to assess the effectiveness of the two-stage projection strategy on ScCNAA performance. For the first stage projection-based method (denoted as Stage_1), the second stage projection used in the proposed method was ignored, i.e., $\mathbf{p}_m$ was removed, whereas $\mathbf{Z}_m$ was removed for the second stage projection-based method (denoted as Stage_2). All constrains of QTs and SNPs used in the proposed method were retained in Stage_1 and Stage_2 methods for a fair comparison. Moreover, the objective function in Eq. (8) can be revised as Eqs. (42) and (43) for Stage_1 and Stage_2, respectively.

$$\begin{aligned}\mathcal{L}(\mathbf{Z}_m,\mathbf{E}_m,\mathbf{Q}_m,\mathbf{S},\mathbf{H}_m) &= \frac{1}{2}\sum_{m=1}^{M}\left\|\mathbf{H}_m^T - f(\mathbf{X}_m)\mathbf{S}_m\right\|_2^2 + \sum_{m=1}^{M}\left\|\mathbf{X}_m - \mathbf{X}_m\mathbf{U}_m\right\|_{21} \\ &+ \gamma\sum_{m=1}^{M}\left\|\mathbf{E}_m\right\|_1 + \lambda\sum_{m=1}^{M} tr(\mathbf{Z}_m^T\mathbf{Y}_m\mathbf{L}_m(\mathbf{Z}_m^T\mathbf{Y}_m)^T) + \alpha_1\sum_{m=1}^{M}\left\|[\mathbf{S}_m,\mathbf{U}_m]\right\|_{21} + \alpha_2\sum_{m=1}^{M}\left\|\mathbf{U}_m\right\|_1 \\ &+ \beta_1\sum_{m=1}^{M}\left\|\mathbf{Z}_m\right\|_{21} + \beta_2\sum_{m=1}^{M}\mathbf{Z}_m^T\mathbf{L}_m^z\mathbf{Z}_m + \sum_m \Phi(\mathbf{Q}_m,\mathbf{Z}_m^T\mathbf{Y}_m - \mathbf{H}_m - \mathbf{E}_m).\end{aligned} \tag{42}$$

$$\begin{aligned}\mathcal{L}(\mathbf{p}_m,\mathbf{S}) &= \frac{1}{2}\sum_{m=1}^{M}\left\|\mathbf{Y}_m^T\mathbf{p}_m - f(\mathbf{X}_m)\mathbf{s}_m\right\|_2^2 + \sum_{m=1}^{M}\left\|\mathbf{X}_m - \mathbf{X}_m\mathbf{U}_m\right\|_{21} + \alpha_1\sum_{m=1}^{M}\left\|[\mathbf{s}_m,\mathbf{U}_m]\right\|_{21} \\ &+ \alpha_2\sum_{m=1}^{M}\left\|\mathbf{U}_m\right\|_1 + \lambda\sum_{m=1}^{M} tr(\mathbf{p}_m^T\mathbf{Y}_m\mathbf{L}_m(\mathbf{p}_m^T\mathbf{Y}_m)^T) + \beta_1\sum_{m=1}^{M}\left\|\mathbf{p}_m\right\|_{21} + \beta_2\sum_{m=1}^{M}\mathbf{p}_m^T\mathbf{L}_m^z\mathbf{p}_m\end{aligned} \tag{43}$$

Table 6 shows that the detection AUC of SNPs and QTs decreased when the first or the second stage projection was removed in the proposed method. In particular, compared with our proposed method, the AUC values of Stage_1 (the second stage projection is removed) on SNPs and QTs are decreased considerably. The first stage projection mainly aims to deal with the missing data issue, achieve latent imaging representations (i.e., $\mathbf{H}_m$), and explore modality-shared and modality-specific information associated with QTs. Moreover, the latent imaging representations were projected into an association space (i.e., $\mathbf{H}_m\mathbf{p}_m$) in the second stage. Therefore, the association space was discarded when the second stage projection was removed (i.e., $\mathbf{p}_m$ is removed). This scenario may lead to the enlarged

distance between $\mathbf{H}_m$ and SNPs, and thus results in poor detection AUC on QTs and SNPs. On the contrast, the inter-modality correlation within QTs was discarded when the first stage projection is removed (Stage_2 method), and each modality QT was projected into the association space to perform association analysis between QTs and SNPs. Moreover, modality-shared information and modality-specific information were included in each modality QT. In this case, modality-shared information may be reused when the association between multimodal QTs and SNPs was analyzed in Stage_2 method, possibly resulting in information redundancy in the association coefficient of SNPs (i.e., **S**). Therefore, compared with that of the proposed method, the AUC value of SNPs is decreased in Stage_2.

Table 6. AUC values of SNPs and QTs after removing one of two-stage projection in the ScCNAA.

AUC	SNP	*p*-value	QT	*p*-value
Stage_1	0.64±0.08	0.004	0.58±0.06	8.8×10^{-5}
Stage_2	0.64±0.16	0.03	0.94±0.01	0.01
ScCNAA	**0.83±0.02**	--	**0.99±0.01**	--

3.1.6. Comparison with previous studies

In this experiment, we compared the proposed ScCNAA method with three state-of-the-art methods on two datasets generated by Cases 1 and 2. First, three comparison methods were performed on the complete multimodal data discarding missing data (i.e., n = 400) because these methods can only be conducted on complete data. By contrast, the proposed method was performed on the incomplete multimodal data. Then, to make a further comparison between the proposed method and three methods, the three comparison methods were performed on the complete multimodal data with n = 500. Meanwhile, the proposed method still performed on the incomplete multimodal data. The results are listed in Tables 7 and 8. Our proposed method achieves the best detection performance among different methods (paired t-test p-value < 0.05), indicating that potential structure correlations among QTs and SNPs can be well captured by using the proposed method, and thus detection performance can be improved. The performance of the three comparison methods on the complete multimodal data with n = 500 is higher than that on the complete multimodal data discarding missing data (i.e., n = 400). The bi-multivariate analytical methods (i.e., MTSCCA and JCB-SCCA) outperform with G-SMuRFS method. Moreover, the detection performance of JCB-SCCA is better than

that of MTSCCA, indicating that considering inter- and intra-modal correlations and structure correlations among QTs may improve detection performance. Our proposed method outperforms JCB-SCCA, which implies that our proposed method is more flexible in analyzing the complex association between SNPs and QTs.

Table 7. AUC of different methods and their *p*-values of paired *t*-tests comparing the different methods with ScCNAA for SNP and QT detection, where three comparison methods were performed on the complete multimodal data discarding missing data (i.e., $n = 400$). "--" in the table indicates that only single QT is used in the corresponding method, or no self of paired *t*-test in ScCNAA.

	AUC	G-SMuFS	MTSCCA	JCB-SCCA	ScCNAA
Case 1	SNP	0.65±0.03	0.75±0.01	0.74±0.01	**0.83±0.01**
	p-value	6.30×10^{-7}	1.50×10^{-6}	6.50×10^{-5}	--
	QT	--	0.78±0.03	0.87±0.01	**0.99±0.01**
	p-value	--	1.70×10^{-5}	1.20×10^{-5}	--
Case 2	SNP	0.57±0.003	0.63±0.02	0.70±0.01	**0.77±0.02**
	p-value	4.90×10^{-6}	0.01	0.002	--
	QT	--	0.70±0.03	0.90±0.01	**1.00±0.00**
	p-value	--	2.10×10^{-7}	0.01	--

Table 8. AUC of different methods and their *p*-values of paired *t*-tests comparing the different methods with ScCNAA for SNP and QT detection, where three comparison methods were performed on the complete multimodal data with $n = 500$. "--" in the table indicates that only single QT is used in the corresponding method, or no self of paired *t*-test in ScCNAA.

	AUC	G-SMuFS	MTSCCA	JCB-SCCA	ScCNAA
Case 1	SNP	0.68±0.01	0.79±0.01	0.78±0.02	**0.83±0.01**
	p-value	1.38×10^{-7}	0.04	0.008	--
	QT	--	0.81±0.03	0.90±0.01	**0.99±0.01**
	p-value	--	4.01×10^{-7}	2.35×10^{-9}	--
Case 2	SNP	0.62±0.01	0.72±0.02	0.70±0.03	**0.77±0.02**
	p-value	0.002	0.01	0.001	**--**
	QT	--	0.77±0.03	0.91±0.03	**1.00±0.00**
	p-value	--	7.8×10^{-5}	0.01	**--**

3.2. Real data study

3.2.1. Data preprocessing on the real data

In this article, the genetic data and multimodal brain imaging data were obtained from the ADNI1 database (www.adni.loni.usc.edu) and PPMI database (https://www.ppmi-info.org/). ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and

non-profit organizations, as a $60 million, 5-year public–private partnership. The primary goal of ADNI has been to test whether serial MRI, PET and other biological markers are useful in clinical trials of mild cognitive impairment (MCI) and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. ADNI subjects aged 55 to 90 from over 50 sites across the US and Canada participated in the research and more detailed information is available at www.adni-info.org. Moreover, PPMI is the first internationally recognized observational study created to identify and validate biomarkers for prediction of PD progression. For up-to-date information on the study, visit www.ppmi-info.org.

From ADNI dataset, T1-weighted MRI and fluorodeoxyglucose PET (FDG-PET) images were used in this study. The scanning parameters of 1.5T MRI images can be found in a previous study (Jack Jr et al., 2008). For the PD, we utilized the T1-weighted MRI and DTI images obtained from the PPMI database. The details of multimodal imaging data in the AD and PD datasets are listed in Tables 9 and 10, respectively.

Table 9. Participant characteristic of sMRI and PET imaging QTs in ADNI dataset.

	sMRI			PET		
	NC	MCI	AD	NC	MCI	AD
Num (n)	198	346	164	90	175	81
Gender (M/F)	108/90	233/123	90/74	57/33	118/57	49/32
Age at Baseline (mean±SD)	76.4±4.9	75.2±7.3	75.6±7.6	75.92±4.8	75.5±7.1	75.8±7.2

Table 10. Participant characteristic of sMRI and DTI imaging QTs in PPMI dataset.

	sMRI			DTI		
	NC	SWEDD	PD	NC	SWEDD	PD
Num (n)	145	50	317	54	31	138
Gender (M/F)	95/50	32/18	205/112	34/20	20/11	82/56
Age at Baseline (mean±SD)	60.2±11.9	60.8±10.0	61.5±10.0	61.3±11.5	56.8±12.6	63.4±10.8

All MRI data were processed under the following steps to extract region of interest (ROI) based features: a) using MIPAV software for anterior commissure and posterior commissure correction; b) image intensity inhomogeneity correction by applying N3 algorithm (Sled et al., 1998); c) skull stripping (HD-BET, https://github.com/MIC-DKFZ/HD-BET); d) registering all images to Montreal Neurological Institute (MNI) space by using advanced normalization tools (Avants et al., 2011; Tustison et al., 2014); e) tissue segmentation by using Atropos algorithm to obtain

four tissues: GM, WM, ventricle, and CSF; f) using the automated anatomical label (AAL) atlas of MNI space to label 90 ROIs; g) computing the gray matter tissue volume of each ROI in the MNI space. Besides, for each subject, we first aligned PET images to their corresponding T1-weighted MRI by using affine registration, and then computed the average PET intensity value of each ROI as PET feature. Thus, we had a 90-dimensional ROI-based feature from both the MRI and PET data, respectively.

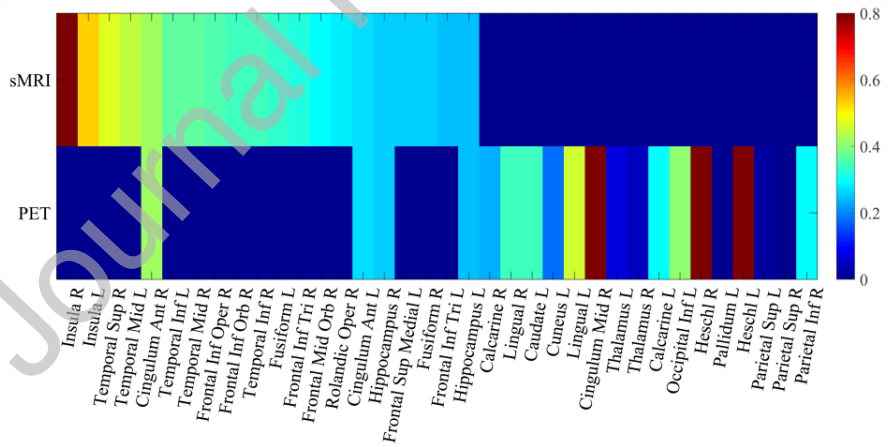
For DTI data, each subject contains 65 original format images where the b0 image does not activate the diffusion gradient, whereas the other 64 images have different gradient directions. The DTI data were preprocessed by the following steps: a) dcm2niix tool (https://www.nitrc.org/frs/?group_id=152) was applied to convert images into a 4-D image and generate a b-vector file and a b-value file indicating each gradient direction and its scalar value, respectively; b) eddy correct command of FMRIB Software Library (FSL) (Jenkinson et al., 2012) was used to correct the eddy current distortion on the 4-D image; c) BET algorithm of FSL was used to preform skull-stripping on b0 image; d) the difiti command of FSL tool and files that have been generated to calculate fractional anisotropy (FA); e) b0 image was aligned to MNI space by using affine registration, and the transformation matrix was applied in FA; f) the AAL atlas was used to calculate the mean tissue density of each region of FA and then the corresponding 90 dimensional ROI-based features could be obtained.

For SNP data provided by ADNI (Saykin et al., 2010) and PPMI datasets (Marek et al., 2018), our quality control procedures include (i) call rate check per subject and SNP marker, (ii) gender check, (iii) sibling pair identification, (iv) the Hardy–Weinberg equilibrium test, (v) marker removal by the minor allele frequency, and (vi) population stratification. The second line preprocessing steps include removal of SNPs with (i) more than 5% missing values, (ii) minor allele frequencies of below 5%, and (iii) Hardy–Weinberg equilibrium $P < 10^{-6}$. The remaining missing genotype variables were imputed as the modal value. After implementing these procedures, 708 subjects in ADNI datasets and 512 subjects in PPMI datasets remained for the subsequent analysis. Finally, a global sure independence screening procedure presented in our previous study (Huang et al., 2015) was applied to select the candidate SNPs, and led to 3000 SNPs for ADNI and PPMI datasets, respectively. After that, the ANNOVAR (Wang et al., 2010) was used to annotate

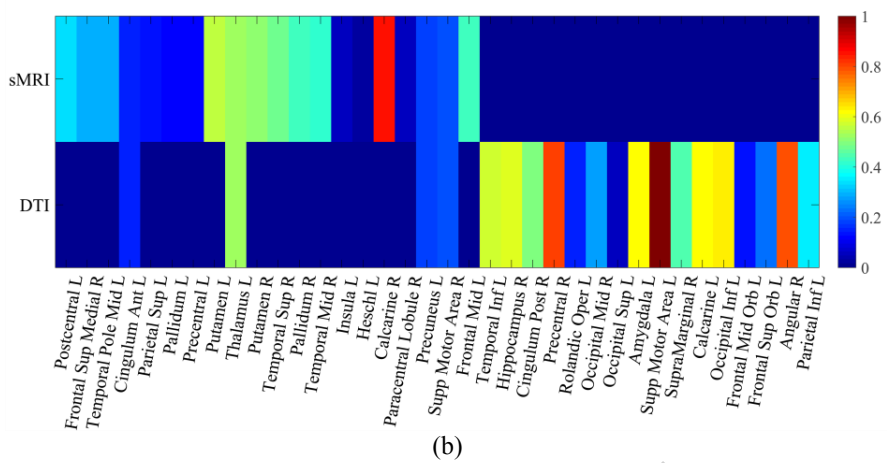
gene corresponding to candidate SNPs. Therefore, gene information was used to group SNPs on the real data.

3.2.2. Biomarker detection

In this section, our goal was to detect the potential biomarkers associated with NDs. In real data, our parameter selection strategy was consistent with that in the simulation data. We averaged the obtained sparse weights across fivefold to ensure stable selection. The top 20 imaging QTs with the highest absolute average $\mathbf{z}_m$ of each modality for AD and PD are shown in Figs. 3 (a) and (b), respectively. The full and short names of 90 QTs were displayed in Table A.1. Top 20 imaging QTs for each modality were selected first, and then the union set of the selected QTs from two modality were listed in Fig. 3. Moreover, we visualize the top 10 important brain regions of each modality from AD and PD datasets in Fig. 4. From Figs 3 and 4, the following QT biomarkers are exploited among the top 20 QTs in each modality for PD: the pallidus, putamen, and thalamus are related to early PD (Garg et al., 2015); precentral is related to PD with diphasic dyskinesia (Zhi et al., 2019); supplementary motor area, precuneus, and hippocampus are associated with PD (Foo et al., 2017; Owens-Walton et al., 2019; Shin et al., 2017).



(a)



(b)

Fig. 3. Heatmaps of weights of the top 20 imaging QTs selected from each modality for (a) AD and (b) PD, respectively, by using the proposed method. Two rows corresponding to two modalities of imaging QTs.

For AD, the following QT biomarkers are exploited among the top 20 QTs in each modality: the hippocampal is associated with cognitive impairment in AD (De Leon et al., 1997; Poulin et al., 2011); insula is related to MCI (Firbank et al., 2021); the anterior cingulate and paracingulate gyri reduces glucose metabolism rate in AD (Jiang et al., 2018); the volume atrophy of thalamus appeared in AD (Low et al., 2019); the cuneus, pallidum, superior temporal gyrus, and middle temporal gyrus are also confirmed to be associated with AD. In Figs. 3 and 4, modality-shared and modality-specific QTs can be identified for each modality of PD/AD. For instance, the bilateral hippocampus and anterior cingulate and paracingulate gyri can be identified, indicating that these QTs extracted from sMRI and PET are potentially AD-related biomarkers. We can also observe that several QTs can be identified by specific imaging technology, for example, the volume atrophy of the hippocampus is observed in the MCI stage by using sMRI scan (Poulin et al., 2011).

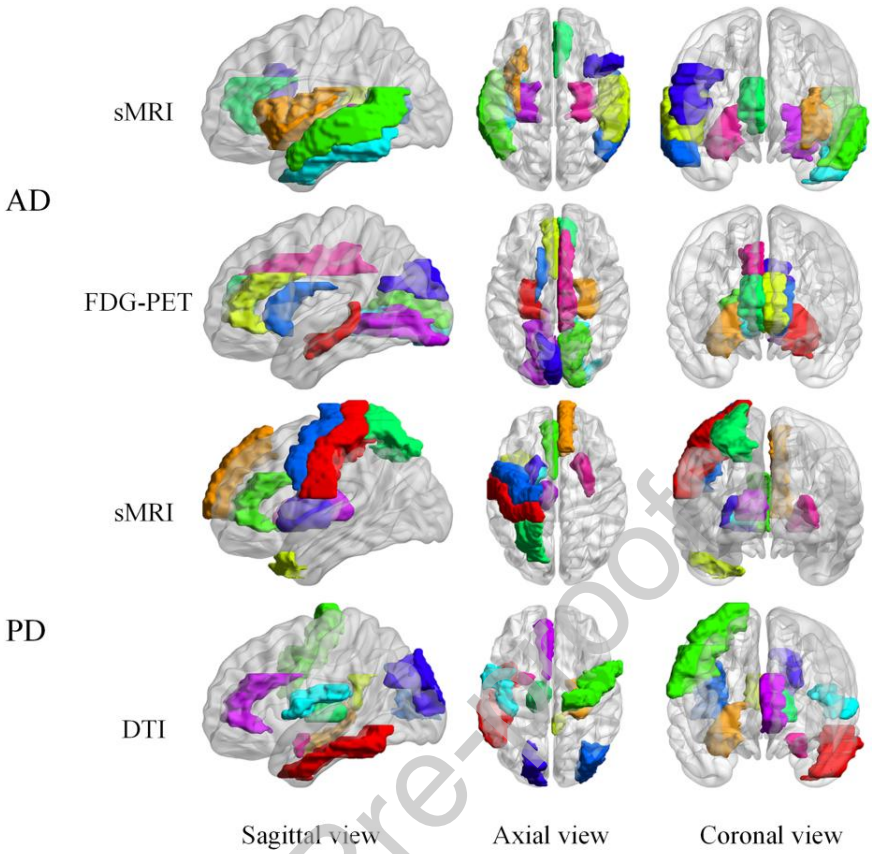


Fig. 4. Top 10 important brain regions and overall distribution from the sMRI (first row) and FDG-PET imaging data of AD (second row), and sMRI (third row) and DTI imaging data of PD (fourth row).

We also identified the relevant top SNPs by using the proposed method. Figs. 5 (a) and (b) show the weights of the selected top 20 SNPs of each modality for AD and PD, respectively. Top 20 SNPs for each modality were selected first, and then the union set of the selected SNPs from two modality were listed in Fig. 5. For PD, the following genes are detected: variants in SNCA gene is related to PD (Campêlo et al., 2017); HMMR, TMPRSS2, and HLA-DRA genes are associated with PD (Salles-Gándara et al., 2020; Zhang et al., 2017); CDKAL1 is linked to bipolar disorder (Haljas et al., 2018); and SNCA gene is connected with the thalamus (Kim et al., 2017), which is related to early PD. For AD, the following genes are detected: CDH13 variants may increase AD risk (Liu et al., 2018); NRXN3 gene variants is associated with schizophrenia, and might increase neuron inflammation in AD (Hishimoto et al., 2019); polymorphisms within ASTN2 gene are connected with age at onset of AD (Wang et al., 2015); ANK3 gene is linked to late-onset AD (Morgan et al., 2007); LRP1B and SLC1A3 genes are associated with AD (Bi et al.,

2019; Zhao et al., 2021); and MACROD2 gene is related to autism (Jahanshad et al., 2013).

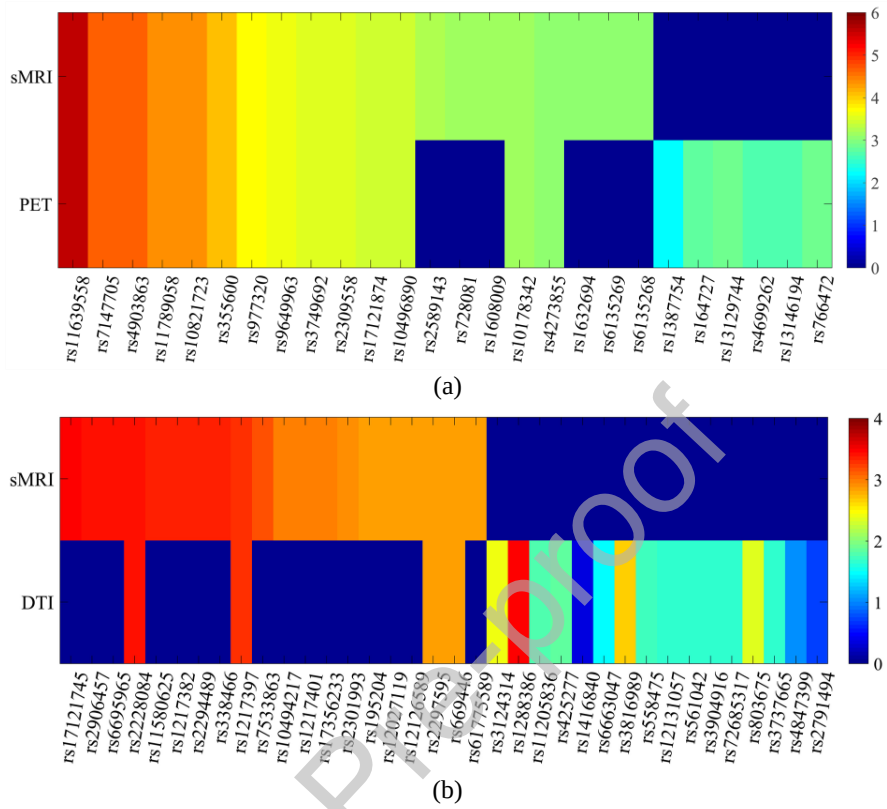


Fig. 5. Heatmaps of weights of the top 20 SNPs selected from each modality for (a) AD and (b) PD, respectively, by using the proposed method. Two rows corresponding to two modalities of imaging QTs.

3.2.3. Comparison with previous studies

We compared the performance of ScCNAA on real data with that of three other state-of-art methods to evaluate the effectiveness of the proposed method. Given that the biomarkers associated with AD/PD in QTs and SNPs for real data have no ground truth, RMSE was used for the performance evaluation, and the results are shown in Table 11. We observe that the proposed method obtained the lowest RMSE values, which indicates that our proposed method outperforms other state-of-art methods. Moreover, the RMSE values of SCCA-based methods (i.e., MTSCCA and JCB-SCCA) are lower than those of G-SMuRFS, which may be because SCCA-based methods consider the structure information among QTs, thereby improving the detection performance.

Table 11. RMSE of different methods on ADNI and PPMI datasets. "--" in the table denotes that a non-*p* value is obtained.

	Method	G-SMuFS	MTSCCA	JCB-SCCA	ScCNAA
ADNI datasets	RMSE	4.3±0.25	0.13±0.002	0.05±0.015	**0.025±0.003**
	p-value	2.8×10^{-10}	4.1×10^{-11}	6.4×10^{-9}	--
PPMI datasets	RMSE	5.2±0.21	0.16±0.017	0.08±0.012	**0.045±0.018**
	p-value	1×10^{-11}	5×10^{-6}	0.004	--

4. Discussion

In this article, a novel ScCNAA method is introduced to analyze the associations between SNPs and multimodal QTs and uncover the genetic basis of the brain structure, function, and disorder associated with NDs. A parallel concatenated projection method is applied without imputing missing data to handle missing data, where modality-shared and modality-specific information can be well captured to obtain latent imaging representations. Connectivity analysis is suitable for exploring complex interconnected network, such as brain connectivity can be used to represent the degree of neuronal fiber connection between two brain regions. Therefore, the structure constraints are applied in the proposed method to explore the structure correlation among SNPs and QTs, and select interconnected QTs or SNPs that are consistent with biological prior knowledge. An l_{21} norm is used in our ScCNAA model to exploit individual information among SNPs and QTs. We also assume that the effects of SNPs on QTs are regarded as a nonlinear function to deal with the complex associations between SNPs and QTs. Moreover, the proposed ScCNAA method is validated on the simulation and real ND datasets, and high performance of biomarker detection is achieved by using the proposed method. Therefore, the proposed method can help to understand the underlying pathological mechanism of NDs.

4.1. Comparison with different missing data handling methods

We compared our proposed method with three imputation methods and a projection method on data simulated by using Case 1. In the QT and SNP detection, the detection AUC value of the proposed ScCNAA was higher than the projection method, thereby indicating that much information on modality-shared and modality-specific QT and SNP biomarkers can be exploited by using the projection approach for handling missing data proposed by our method than that proposed by the Zhou's work. Meanwhile, the detection

AUC values of the projection-based methods were higher than those of the three imputation methods. Consequently, some noise information is introduced by using the imputation method, which may result in decreased detection performance. Moreover, all available data can be used in the projection-based method to train a reliable model thus improving the detection performance.

4.2. Comparison with the previous methods

We compare the proposed method with three state-of-the-art methods using simulation data and two real ND datasets. In the simulation data, the three comparison methods can only be conducted on complete multimodal data, so we first performed three comparison methods on the complete multimodal data discarding missing data (i.e., n = 400). To further evaluate the effectiveness of our method, we then performed the three comparison methods on the complete multimodal data with n = 500, and conducted the proposed method on the multimodal data with missing data. Table 8 shows that the detection AUC value of our proposed method is still higher than those of the other methods, suggesting that the effectiveness of the proposed ScCNAA in dealing with missing data and fully leveraging both multimodal data by using a parallel concatenated projection method with multiple constraints. In Tables 7 and 8, the detection AUC of QTs and SNPs of our method shows the best results on simulation data generated by using Cases 1 and 2, followed by JCB-SCCA, MTSCCA, and G-SMuRFS. Compared with G-SMuRFS, improved detection performance is achieved by using MTSCCA, which may be because individual level information on QTs is included in MTSCCA. Improvements in the JCB-SCCA may contribute to structure correlation in the QTs is captured by using the connectivity penalty. Further improvement of detection performance is observed in ScCNAA, which benefit to inherent correlations among inter- and intra-multimodal QTs as well as correlated data structures within SNPs are captured by using the proposed method.

In the real data, Table 11 shows that the RMSE values of the JCB-SCCA and MTSCCA methods are lower than that of G-SMuRFS (the pair t-test p-values < 0.05), which indicates that the effectiveness by including structure information among QTs on analyzing correlations between SNPs and QTs. Moreover, the RMSE value of JCB-SCCA is lower than that of MTSCCA (the pair t-test p-values are 2.2×10^{-10} and 2.3×10^{-5} in the ADNI and PPMI datasets, respectively), which suggests that the usefulness of exploiting

the information among inter- and intra-modal data on analyzing correlations between SNPs and QTs. In comparison with the JCB-SCCA method, the group-induced graph self-expression constraint is introduced in the proposed method to learn strong structure correlations among inter- and intra-group of SNPs. Accordingly, the lowest RMSE value is obtained by the proposed ScCNAA.

4.3. biomarker detection of NDs

The biomarker detection results on real ND data illustrate that a gene including some SNPs are detected by using the proposed method, which may be because the strong structure correlations among inter- and intra-group of SNPs are considered in the proposed method. These genes detected by the proposed method have been demonstrated to be related to NDs based on some previous studies (Gustaw-Rothenberg et al., 2010; Rosas et al., 2020; Yang et al., 2021). Moreover, brain regions related to NDs are discovered by using the proposed method, such as hippocampus and thalamus etc. Figs. 3 and 5 show that the modality shared and specific SNPs and QTs can be detected, which contributes to the parallel concatenated projection combined with structure constraints used in the proposed method.

4.4. Limitations and future work

In this research, we compared the proposed method with the multimodal regression (G-SMuRFS) and SCCA-based methods (MTSCCA and JCB-SCCA) on the simulation data sets and real AD data. Although, competitive results are obtained by using the proposed method, several technical issues are still needed to be addressed in our future research. First, we only used small samples, which may lead to the overfitting problem for various penalized regression methods. Therefore, more samples should be included in our experiments in future. Second, QTs are usually changed slowly over time as a disorder progress. Complementary information can be provided by using multimodal QTs. Therefore, integrating imaging data from different time points and different modality QTs into a framework may provide rich information to detect biomarkers and improve detection performance.

Declaration of interests

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

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 81601562, No. 81974275, and No. U1501256), the Science and Technology Planning Project of Guangzhou (No. 201904010417), the Guangdong Basic and Applied Basic Research Foundation (No. 2021A1515012011), and the Science and Technology Planning Project of Guangdong Province (No. 2015B010131011). Data collection and sharing for this project was funded by ADNI (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). Data used in the preparation of this article were obtained from the PPMI database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org. PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including AbbVie, Avid Radiopharmaceuticals, Biogen Idec, Bristol-Meyers Squibb, Covance, GE Healthcare, Genentech, Glaxo-SmithKline, Eli Lilly and Company, Lundbeck, Merck, Meso Scale Diagnostics, Pfizer Inc., Piramal Imaging, Roche CNS group, Servier, UCB and Golub Capital.

Appendix A. Full and short names of QTs

Table A.1. Full and short names of 90 QTs in ADNI1 and PPMI datasets.

QT	Short Name	Full Name	QT	Short Name	Full Name
1/2	Precentral L/R	Precentral Left/Right	47/78	Lingual L/R	Lingual gyrus Left/Right
3/4	Frontal Sup L/R	Superior frontal gyrus Left/Right	49/50	Occipital Sup L/R	Superior occipital gyrus Left/Right
5/6	Frontal Sup Orb L/R	Orbital superior frontal gyrus Left/Right	51/52	Occipital Mid L/R	Middle occipital gyrus Left/Right

7/8	Frontal Mid L/R	Middle frontal gyrus Left/Right	53/54	Occipital Inf L/R	Inferior occipital gyrus Left/Right
9/10	Frontal Mid Orb L/R	Orbital middle frontal gyrus Left/Right	55/56	Fusiform L/R	Fusiform gyrus Left/Right
11/12	Frontal Inf Oper L/R	Opercular inferior frontal gyrus Left/Right	57/58	Postcentral L/R	Postcentral Left/Right
13/14	Frontal Inf Tri L/R	Triangular inferior frontal gyrus Left/Right	59/60	Parietal Sup L/R	Superior parietal gyrus Left/Right
15/16	Frontal Inf Orb L/R	Orbital inferior frontal gyrus Left/Right	61/62	Parietal Inf L/R	Inferior parietal gyri Left/Right
17/18	Rolandic Oper L/R	Rolandic operculum Left/Right	63/64	SupraMarginal L/R	SupraMarginal gyrus Left/Right
19/20	Supp Motor Area L/R	Supplementary motor area Left/Right	65/66	Angular L/R	Angular gyrus Left/Right
21/22	Olfactory L/R	Olfactory cortex Left/Right	67/68	Precuneus L/R	Precuneus Left/Right
23/24	Frontal Sup Medial L/R	Superior frontal medial gyrus Left/Right	69/70	Paracentral Lobule L/R	Paracentral Lobule Left/Right
25/26	Frontal Mid Orb L/R	Medial orbital superior frontal gyrus Left	71/72	Caudate L/R	Caudate nucleus Left/Right
27/28	Rectus L/R	Gyrus rectus Left/Right	73/74	Putamen L/R	Putamen Left/Right
29/30	Insula L/R	Insula Left/Right	75/76	Pallidum L/R	Pallidum Left/Right
31/32	Cingulum Ant L/R	Anterior cingulate and paracingulate gyri Left/Right	77/78	Thalamus L/R	Thalamus Left/Right
33/34	Cingulum Mid L/R	Medican cingulate and paracingulate gyri Left/Right	79/80	Heschl L/R	Heschl gyrus Left/Right
35/36	Cingulum Post L/R	Posterior cingulate gyrus Left/Right	81/82	Temporal Sup L/R	Superior temporal gyrus Left/Right
37/38	Hippocampus L/R	Hippocampus Left/Right	83/84	Temporal Pole Sup L/R	Temporal pole: Superior temporal gyrus Left/Right
39/40	ParaHippocampal L/R	ParaHippocampal Left/Right	85/86	Temporal Mid L/R	Middle temporal gyrus Left/Right
41/42	Amygdala L/R	Amygdala Left/Right	87/88	Temporal Pole Mid L/R	Temporal pole: Middle temporal gyrus Left/Right
43/44	Calcarine L/R	Calcarine fissure Left/Right	89/90	Temporal Inf L/R	Inferior temporal gyrus Left/Right
45/46	Cuneus L/R	Cuneus Left/Right			

References

Adeli, E., Shi, F., An, L., Wee, C.-Y., Wu, G., Wang, T., Shen, D., 2016. Joint feature-sample selection and robust diagnosis of Parkinson's disease from MRI data. NeuroImage 141, 206-219.

Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C., 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. NeuroImage 54, 2033-2044.

Barrett, J.C., Fry, B., Maller, J., Daly, M.J., 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21, 263-265.

Bi, X.-a., Cai, R., Wang, Y., Liu, Y., 2019. Effective Diagnosis of Alzheimer's Disease via Multimodal Fusion Analysis Framework. Frontiers in Genetics 10.

Bi, X., Yang, L., Li, T., Wang, B., Zhu, H., Zhang, H., 2017. Genome-wide mediation analysis of psychiatric and cognitive traits through imaging phenotypes. Human Brain Mapping 38, 4088-4097.

Campêlo, C.L.C., Cagni, F.C., de Siqueira Figueredo, D., Oliveira Jr., L.G., Silva-Neto, A.B., Macêdo, P.T., Santos, J.R., Izídio, G.S., Ribeiro, A.M., de Andrade, T.G., de Oliveira Godeiro, C., Silva, R.H., 2017. Variants in SNCA Gene Are Associated with Parkinson's Disease Risk and Cognitive Symptoms in a Brazilian Sample. Frontiers in aging neuroscience 9.

Candès, E.J., Recht, B., 2009. Exact Matrix Completion via Convex Optimization. Foundations of Computational Mathematics 9, 717.

De Leon, M.J., George, A.E., Golomb, J., Tarshish, C., Convit, A., Kluger, A., De Santi, S., Mc Rae, T., Ferris, S.H., Reisberg, B., Ince, C., Rusinek, H., Bobinski, M., Quinn, B., Miller, D.C., Wisniewski, H.M., 1997. Frequency of hippocampal formation atrophy in normal aging and Alzheimer's disease. Neurobiology of Aging 18, 1-11.

Du, L., Huang, H., Yan, J., Kim, S., Risacher, S.L., Inlow, M., Moore, J.H., Saykin, A.J., Shen, L., Initiative, A.s.D.N., 2016. Structured sparse canonical correlation analysis for brain imaging genetics: an improved GraphNet method. Bioinformatics 32, 1544-1551.

Du, L., Liu, K., Yao, X., Risacher, S.L., Han, J., Saykin, A.J., Guo, L., Shen, L., 2021. Multi-Task Sparse Canonical Correlation Analysis with Application to Multi-Modal Brain Imaging Genetics. IEEE/ACM Transactions on Computational Biology and Bioinformatics 18, 227-239.

Firbank, M.J., Durcan, R., O'Brien, J.T., Allan, L.M., Barker, S., Ciafone, J., Donaghy, P.C., Hamilton, C.A., Lawley, S., Roberts, G., Taylor, J.-P., Thomas, A.J., 2021. Hippocampal and insula volume in mild cognitive impairment with Lewy bodies. Parkinsonism & Related Disorders 86, 27-33.

Foo, H., Mak, E., Chander, R.J., Ng, A., Au, W.L., Sitoh, Y.Y., Tan, L.C.S., Kandiah, N., 2017. Associations of hippocampal subfields in the progression of cognitive decline related to Parkinson's disease. NeuroImage: Clinical 14, 37-42.

Garg, A., Appel-Cresswell, S., Popuri, K., McKeown, M.J., Beg, M.F., 2015. Morphological alterations in the caudate, putamen, pallidum, and thalamus in Parkinson's disease. Frontiers in neuroscience 9.

Gustaw-Rothenberg, K., Lerner, A., Bonda, D.J., Lee, H.G., Zhu, X., Perry, G., Smith, M.A., 2010. Biomarkers in Alzheimer's disease: past, present and future. Biomarkers in medicine 4, 15-26.

Haljas, K., Amare, A.T., Alizadeh, B.Z., Hsu, Y.H., Mosley, T., Newman, A., Murabito, J., Tiemeier, H., Tanaka, T., van Duijn, C., Ding, J., Llewellyn, D.J., Bennett, D.A., Terracciano, A., Launer, L., Ladwig, K.H., Cornelis, M.C., Teumer, A., Grabe, H., Kardia, S.L.R., Ware, E.B., Smith, J.A., Snieder, H., Eriksson, J.G., Groop, L., Räikkönen, K., Lahti, J., 2018. Bivariate Genome-Wide Association Study of Depressive Symptoms With Type 2 Diabetes and Quantitative Glycemic Traits. Psychosomatic medicine 80, 242-251.

Hardoon, D.R., Shawe-Taylor, J., 2011. Sparse canonical correlation analysis. Machine Learning 83, 331-353.

Hastie, T., Mazumder, R., Lee, J.D., Zadeh, R., 2015. Matrix completion and low-rank SVD via fast alternating least squares. The Journal of Machine Learning Research 16, 3367-3402.

Hishimoto, A., Pletnikova, O., Lang, D.L., Troncoso, J.C., Egan, J.M., Liu, Q.-R., 2019. Neurexin 3 transmembrane and soluble isoform expression and splicing haplotype are associated with neuron inflammasome and Alzheimer's disease. Alzheimers Research & Therapy 11.

Huang, M., Chen, X., Yu, Y., Lai, H., Feng, Q., 2021a. Imaging Genetics Study Based on a Temporal Group Sparse Regression and Additive Model for Biomarker Detection of Alzheimer's Disease. IEEE Transactions on Medical Imaging 40, 1461-1473.

Huang, M., Lai, H., Yu, Y., Chen, X., Wang, T., Feng, Q., 2021b. Deep-gated recurrent unit and diet network-based genome-wide association analysis for detecting the biomarkers of Alzheimer's disease. Medical Image Analysis 73, 102189.

Huang, M., Nichols, T., Huang, C., Yu, Y., Lu, Z., Knickmeyer, R.C., Feng, Q., Zhu, H., Alzheimer's Disease Neuroimaging, I., 2015. FVGWAS: Fast voxelwise genome wide association analysis of large-scale imaging genetic data. NeuroImage 118, 613-627.

Huang, M., Yu, Y., Yang, W., Feng, Q., Initiative, A.s.D.N., 2019. Incorporating spatial–anatomical similarity into the VGWAS framework for AD biomarker detection. Bioinformatics 35, 5271-5280.

Jack Jr, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L. Whitwell, J., Ward, C., 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine 27, 685-691.

Jagust, W.J., Mormino, E.C., 2011. Lifespan brain activity, β-amyloid, and Alzheimer's disease. Trends Cogn Sci 15, 520-526.

Jahanshad, N., Rajagopalan, P., Hua, X., Hibar, D.P., Nir, T.M., Toga, A.W., Jack, C.R., Jr., Saykin, A.J., Green, R.C., Weiner, M.W., Medland, S.E., Montgomery, G.W., Hansell, N.K., McMahon, K.L., de Zubicaray, G.I., Martin, N.G., Wright, M.J., Thompson, P.M., Alzheimer's Dis Neuroimaging, I., 2013. Genome-wide scan of healthy human connectome discovers SPON1 gene variant influencing dementia severity. Proceedings of the National Academy of Sciences of the United States of America 110, 4768-4773.

Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M., 2012. Fsl. NeuroImage 62, 782-790.

Jiang, J., Sun, Y., Zhou, H., Li, S., Huang, Z., Wu, P., Shi, K., Zuo, C., Neuroimaging Initiative, A.s.D., 2018. Study of the Influence of Age in ¹⁸F-FDG PET Images Using a Data-Driven Approach and Its Evaluation in Alzheimer's Disease. Contrast Media & Molecular Imaging 2018, 3786083.

Kim, M., Kim, J., Lee, S.-H., Park, H., 2017. Imaging genetics approach to Parkinson's disease and its correlation with clinical score. Scientific Reports 7, 46700.

Kim, M., Won, J.H., Youn, J., Park, H., 2020. Joint-Connectivity-Based Sparse Canonical Correlation Analysis of Imaging Genetics for Detecting Biomarkers of Parkinson's Disease. IEEE Transactions on Medical Imaging 39, 23-34.

Lei, B., Zhao, Y., Huang, Z., Hao, X., Zhou, F., Elazab, A., Qin, J., Lei, H., 2020. Adaptive sparse learning using multi-template for neurodegenerative disease diagnosis. Medical Image Analysis 61, 101632.

Lin, Z., Chen, M., Ma, Y., 2010. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. arXiv preprint arXiv:1009.5055.

Liu, F.F., Zhang, Z., Chen, W., Gu, H.Y., Yan, Q.J., 2018. Regulatory mechanism of microRNA-377 on CDH13 expression in the cell model of Alzheimer's disease. European review for medical and pharmacological sciences 22, 2801-2808.

Low, A., Mak, E., Malpetti, M., Chouliaras, L., Nicastro, N., Su, L., Holland, N., Rittman, T., Rodríguez, P.V., Passamonti, L., Bevan-Jones, W.R., Jones, P.P.S., Rowe, J.B., O'Brien, J.T., 2019. Asymmetrical atrophy of thalamic subnuclei in Alzheimer's disease and amyloid-positive mild cognitive impairment is associated with key clinical features. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 11, 690-699.

Marcus, C., Mena, E., Subramaniam, R.M., 2014. Brain PET in the diagnosis of Alzheimer's disease. Clin Nucl Med 39, e413-e426.

Marek, K., Chowdhury, S., Siderowf, A., Lasch, S., Coffey, C.S., Caspell‐Garcia, C., Simuni, T., Jennings, D., Tanner, C.M., Trojanowski, J.Q., 2018. The Parkinson's progression markers initiative (PPMI)‐establishing a PD biomarker cohort. Annals of clinical translational neurology 5, 1460-1477.

Mishra, V.R., Sreenivasan, K.R., Zhuang, X., Yang, Z., Cordes, D., Walsh, R.R., 2019. Influence of analytic techniques on comparing DTI-derived measurements in early stage Parkinson's disease. Heliyon 5, e01481-e01481.

Morgan, A.R., Turic, D., Jehu, L., Hamilton, G., Hollingworth, P., Moskvina, V., Jones, L., Lovestone, S., Brayne, C., Rubinsztein, D.C., Lawlor, B., Gill, M., O'Donovan, M.C., Owen, M.J., Williams, J., 2007. Association studies of 23 positional/functional candidate genes on chromosome 10 in late-onset Alzheimer's disease. American Journal of Medical Genetics Part B-Neuropsychiatric Genetics 144B, 762-770.

Nie, F., Huang, H., Cai, X., Ding, C., 2010. Efficient and robust feature selection via joint ℓ2, 1-norms minimization. Advances in neural information processing systems 23.

Owens-Walton, C., Jakabek, D., Power, B.D., Walterfang, M., Velakoulis, D., Van Westen, D., Looi, J.C., Shaw, M., Hansson, O.J.P.o., 2019. Increased functional connectivity of thalamic subdivisions in patients with Parkinson's disease. PloS one 14, e0222002.

Poulin, S.P., Dautoff, R., Morris, J.C., Barrett, L.F., Dickerson, B.C., 2011. Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. Psychiatry Research: Neuroimaging 194, 7-13.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics 81, 559-575.

Rosas, I., Martínez, C., Clarimón, J., Lleó, A., Illán-Gala, I., Dols-Icardo, O., Borroni, B., Almeida, M.R., van der Zee, J., Van Broeckhoven, C., 2020. Role for ATXN1, ATXN2, and HTT intermediate repeats in frontotemporal dementia and Alzheimer's disease. Neurobiology of aging 87, 139-e131.

Salles-Gándara, P., Rojas-Fernandez, A., Salinas-Rebolledo, C., Milan-Sole, A., 2020. The potential role of SARS-COV-2 in the pathogenesis of Parkinson's Disease. Frontiers in neurology 11, 1044.

Salvatore, C., Cerasa, A., Castiglioni, I., Gallivanone, F., Augimeri, A., Lopez, M., Arabia, G., Morelli, M., Gilardi, M.C., Quattrone, A., 2014. Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and Progressive Supranuclear Palsy. Journal of neuroscience methods 222, 230-237.

Saykin, A.J., Shen, L., Foroud, T.M., Potkin, S.G., Swaminathan, S., Kim, S., Risacher, S.L., Nho, K., Huentelman, M.J., Craig, D.W., 2010. Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans. Alzheimer's & dementia 6, 265-273.

Scheltens, P., De Strooper, B., Kivipelto, M., Holstege, H., Chételat, G., Teunissen, C.E., Cummings, J., van der Flier, W.M., 2021. Alzheimer's disease. The Lancet 397, 1577-1590.

Schneider, T., 2001. Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values. Journal of Climate 14, 853-871.

Shao, W., Wang, T., Sun, L., Dong, T., Han, Z., Huang, Z., Zhang, J., Zhang, D., Huang, K., 2020. Multi-task multi-modal learning for joint diagnosis and prognosis of human cancers. Medical Image Analysis 65, 101795.

Sheng, K., Fang, W., Su, M., Li, R., Zou, D., Han, Y., Wang, X., Cheng, O., 2014. Altered Spontaneous Brain Activity in Patients with Parkinson's Disease Accompanied by Depressive Symptoms, as Revealed by Regional Homogeneity and Functional Connectivity in the Prefrontal-Limbic System. Plos One 9.

Shin, J.H., Shin, S.A., Lee, J.-Y., Nam, H., Lim, J.-S., Kim, Y.K., 2017. Precuneus degeneration and isolated apathy in patients with Parkinson's disease. Neuroscience Letters 653, 250-257.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE transactions on medical imaging 17, 87-97.

Sperling, R.A., Aisen, P.S., Beckett, L.A., Bennett, D.A., Craft, S., Fagan, A.M., Iwatsubo, T., Jack, C.R., Kaye, J., Montine, T.J., Park, D.C., Reiman, E.M., Rowe, C.C., Siemers, E., Stern, Y., Yaffe, K., Carrillo, M.C., Thies, B., Morrison-Bogorad, M., Wagster, M.V., Phelps, C.H., 2011. Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia 7, 280-292.

Thung, K.H., Wee, C.Y., Yap, P.T., Shen, D., 2014. Neurodegenerative disease diagnosis using incomplete multi-modality data via matrix shrinkage and completion. NeuroImage 91, 386-400.

Tustison, N.J., Cook, P.A., Klein, A., Song, G., Das, S.R., Duda, J.T., Kandel, B.M., van Strien, N., Stone, J.R., Gee, J.C., Avants, B.B., 2014. Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. NeuroImage 99, 166-179.

Wachinger, C., Nho, K., Saykin, A.J., Reuter, M., Rieckmann, A., Alzheimer's Disease Neuroimaging, I., 2018. A Longitudinal Imaging Genetics Study of Neuroanatomical Asymmetry in Alzheimer's Disease. Biol Psychiatry 84, 522-530.

Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S.L., Saykin, A.J., Shen, L., For the Alzheimer's Disease Neuroimaging, I., 2012. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. Bioinformatics 28, 229-237.

Wang, K.-S., Tonarelli, S., Luo, X., Wang, L., Su, B., Zuo, L., Mao, C., Rubin, L., Briones, D., Xu, C., 2015. Polymorphisms within ASTN2 gene are associated with age at onset of Alzheimer's disease. Journal of Neural Transmission 122, 701-708.

Wang, K., Li, M., Hakonarson, H., 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research 38, e164-e164.

Wang, X., Chen, H., Yan, J., Nho, K., Risacher, S.L., Saykin, A.J., Shen, L., Huang, H., ADNI, f.t., 2018. Quantitative trait loci identification for brain endophenotypes via new additive model with random networks. Bioinformatics 34, i866-i874.

Yang, J., Ma, Q., Dincheva, I., Giza, J., Jing, D., Marinic, T., Milner, T.A., Rajadhyaksha, A., Lee, F.S., Hempstead, B.L., 2021. SorCS2 is required for social memory and trafficking of the NMDA receptor. Molecular Psychiatry 26, 927-940.

Yin, J., Chen, X., Xing, E.P., 2012. Group Sparse Additive Models. Proc Int Conf Mach Learn 2012, 871-878.

Yoshida, K., Yoshimoto, J., Doya, K., 2017. Sparse kernel canonical correlation analysis for discovery of nonlinear interactions in high-dimensional data. Bmc Bioinformatics 18.

Zhang, M., Mu, H., Shang, Z., Kang, K., Lv, H., Duan, L., Li, J., Chen, X., Teng, Y., Jiang, Y., Zhang, R., 2017. Genome-wide pathway-based association analysis identifies risk pathways associated with Parkinson's disease. Neuroscience 340, 398-410.

Zhao, X., Yao, H., Li, X., 2021. Unearthing of Key Genes Driving the Pathogenesis of Alzheimer's Disease via Bioinformatics. Frontiers in Genetics 12.

Zhi, Y., Wang, M., Yuan, Y.S., Shen, Y.T., Ma, K.W., Gan, C.T., Si, Q.Q., Wang, L.N., Cao, S.W., Zhang, K.Z., 2019. The increased gray matter volumes of precentral gyri in Parkinson's disease patients with diphasic dyskinesia. Aging 11, 9661-9671.

Zhou, T., Liu, M., Thung, K.-H., Shen, D., 2019a. Latent Representation Learning for Alzheimer's Disease Diagnosis With Incomplete Multi-Modality Neuroimaging and Genetic Data. IEEE Transactions on Medical Imaging 38, 2411-2422.

Zhou, T., Thung, K.-H., Liu, M., Shen, D., 2018. Brain-wide genome-wide association study for Alzheimer's disease via joint projection learning and sparse regression model. IEEE Transactions on Biomedical Engineering 66, 165-175.

Zhou, T., Thung, K.H., Liu, M., Shen, D., 2019b. Brain-Wide Genome-Wide Association Study for Alzheimer's Disease via Joint Projection Learning and Sparse Regression Model. IEEE Transactions on Biomedical Engineering 66, 165-175.

Zhu, X., Zhang, S., Jin, Z., Zhang, Z., Xu, Z., 2011. Missing Value Estimation for Mixed-Attribute Data Sets. IEEE Transactions on Knowledge and Data Engineering 23, 110-121.

Graphical abstract