Computer Methods and Programs in Biomedicine 232 (2023) 107450

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

A multi-task SCCA method for brain imaging genetics and its application in neurodegenerative diseases

Xin Zhang[a,1], Yipeng Hao[a,1], Jin Zhang[b], Yanuo Ji[a], Shihong Zou[a], Shijie Zhao[b], Songyun Xie[c], Lei Du[b,*]

[a] Institute of Medical Research, Northwestern Polytechnical University, Xi'an, Shannxi 710072, China
[b] School of Automation, Northwestern Polytechnical University, Xi'an, Shannxi 710072, China
[c] School of Electronics and Information, Northwestern Polytechnical University, Xi'an, Shannxi 710072, China

ARTICLE INFO

Article history:
Received 9 June 2022
Revised 24 February 2023
Accepted 24 February 2023

Keywords:
Brain imaging genetics
Multi-task sparse canonical correlation analysis (MTSCCA)
Parameter decomposition
Feature selection

ABSTRACT

Background and Objectives: In brain imaging genetics, multi-task sparse canonical correlation analysis (MTSCCA) is effective to study the bi-multivariate associations between genetic variations such as single nucleotide polymorphisms (SNPs) and multi-modal imaging quantitative traits (QTs). However, most existing MTSCCA methods are neither supervised nor capable of distinguishing the shared patterns of multi-modal imaging QTs from the specific patterns.

Methods: A new diagnosis-guided MTSCCA (DDG-MTSCCA) with parameter decomposition and graph-guided pairwise group lasso penalty was proposed. Specifically, the multi-tasking modeling paradigm enables us to comprehensively identify risk genetic loci by jointly incorporating multi-modal imaging QTs. The regression sub-task was raised to guide the selection of diagnosis-related imaging QTs. To reveal the diverse genetic mechanisms, the parameter decomposition and different constraints were utilized to facilitate the identification of modality-consistent and -specific genotypic variations. Besides, a network constraint was added to find out meaningful brain networks. The proposed method was applied to synthetic data and two real neuroimaging data sets respectively from Alzheimer's disease neuroimaging initiative (ADNI) and Parkinson's progression marker initiative (PPMI) databases.

Results: Compared with the competitive methods, the proposed method exhibited higher or comparable canonical correlation coefficients (CCCs) and better feature selection results. In particular, in the simulation study, DDG-MTSCCA showed the best anti-noise ability and achieved the highest average hit rate, about 25% higher than MTSCCA. On the real data of Alzheimer's disease (AD) and Parkinson's disease (PD), our method obtained the highest average testing CCCs, about 40% ~ 50% higher than MTSCCA. Especially, our method could select more comprehensive feature subsets, and the top five SNPs and imaging QTs were all disease-related. The ablation experimental results also demonstrated the significance of each component in the model, i.e., the diagnosis guidance, parameter decomposition, and network constraint.

Conclusions: These results on simulated data, ADNI and PPMI cohorts suggested the effectiveness and generalizability of our method in identifying meaningful disease-related markers. DDG-MTSCCA could be a powerful tool in brain imaging genetics, worthy of in-depth study.

1. Introduction

As an emerging topic in the last decade, brain imaging genetics aims to uncover the associations between genetic factors and the structure or function of the brain, especially for neurological diseases [1–3]. So far, different types of imaging quantitative traits (QTs, a.k.a. endophenotypes) have been used, and many previously missed genetic alterations such as single nucleotide polymorphisms (SNPs) have been identified for brain diseases [4–7]. Since multi-modal imaging QTs can characterize an individual with different representations, the joint utilization of multi-modal information may provide more leverage in discovering risk loci and abnormal QTs in disordered brains compared to single-modality data [6,8]. Thus, how to effectively explore the bi-multivariate associations between multi-modal imaging QTs and SNPs has become a critical issue to be solved, which can further promote the understanding of brain pathology.

Previous studies have suggested that the bi-multivariate sparse canonical correlation analysis (SCCA) is a useful technique to identify the SNP-QT associations [9–12]. However, the conventional

* Corresponding author.
E-mail address: dulei@nwpu.edu.cn (L. Du).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.cmpb.2023.107450
0169-2607/

SCCA methods are generally applied via learning the bi-directional relationship between SNP data and only one modality of imaging QTs, i.e., two-view SCCA, restricting its power in multi-modal information exploration. Afterward the multi-view SCCA (mSCCA) methods were developed, attempting to take the advantage of multi-modal imaging data [13–16]. For example, Witten et al. presented a proposal for multiple groups of canonical correlation analyses (CCAs) to extend the sparse CCA methodology to the case of more than two data sets [13]. Hao et al. proposed the three-way SCCA method of T-SCCA, a stack of three SCCAs, aiming to learn the intrinsic associations among genetic markers, QTs and clinical scores [15]. And Fang et al. designed the joint SCCA (JSCCA), which used a generalized fused lasso penalty to jointly estimate multiple associations among multi-class subjects, including shared and class-specific patterns [14]. Whereas, these mSCCA methods are actually simple extensions to conventional two-view SCCA, which are unsupervised and, most importantly, use the multi-modal imaging data straightforwardly. More recently, the idea of multi-task learning has been introduced to multi-modal imaging genetics analysis, i.e., multi-task SCCA (MTSCCA) [6]. By constructing multiple SCCA sub-tasks, i.e., each modality corresponds to one task, MTSCCA made use of complementary information carried by different imaging data, and could achieve improved association identification and feature selection results. On this basis, Wei et al. proposed a sparse bivariate learning model with a linear regression model [17]. Wang et al. proposed a multi-task sparse canonical correlation analysis regression model that integrated multi-modal biological clinical indicators [18]. Chen et al. proposed a nonlinear multi-task SCCA, applied to incomplete multi-modal imaging and genetic data [19].

However, it is still insufficient before the application of MTSCCA in real data. First, similar to mSCCA, MTSCCA is still unsupervised without using the diagnosis information. This may lead to disease-irrelevant SNP-QT associations, thereby misleading the interpretability of the method. For example, Wei and Wang *et al*'s studies proved the importance of diagnosis in feature selection [17,18]. Second, the disentanglement of the shared and specific information is inadequate due to the lack of an uncoupling mechanism [20]. These two kinds of information are important for further identification of meaningful features and tracing back to the complex genetic mechanism of disease, as some brain regions may exhibit modality-consistent characteristics and other regions may own modality-specific characteristics. Moreover, according to neuroimaging findings, the brain is organized in the form of networks instead of isolated regions [21,22]. This means that strong connections may exist between some QT pairs (one brain region may influence another brain region), while weak or no connections exist between other pairs. Unfortunately, current MTSCCA methods ignore this which further limits their capability.

Given the above considerations, we proposed an improved multi-task SCCA method for multi-modal imaging genetic association analysis. First, the basic multi-task SCCA model was constructed to explore the associations between SNPs and multi-modal QTs. To achieve more precise disease-related SNP-QT associations, one regression task corresponding to diagnosis status was raised to guide the selection of diagnosis-related imaging QTs. Moreover, the canonical weight associated with genetic data was decomposed into the task-consistent and task-specific parts, and different penalties were imposed to pursue the modality-consistent and -specific subsets of SNPs [20]. Notably, an orthogonal constraint was employed to better disentangle the shared and specific components. As for the brain network constraints, a graph-guided pairwise group lasso penalty (named GGL-penalty [9]) was employed for network identification. Finally, the proposed method was named DDG-MTSCCA as it jointly applies parameter decomposition, diagnosis information, and GGL-penalty. An efficient optimization algorithm was provided to solve the overall problem. To evaluate the performance of DDG-MTSCCA, we respectively used simulated data, and two independent real neuroimaging genetic data sets from the Alzheimer's disease neuroimaging initiative (ADNI) and Parkinson progression marker initiative (PPMI) databases. Our experimental results showed that DDG-MTSCCA performed better or comparably compared to benchmarks in canonical correlation coefficients (CCCs) and feature subsets selection of meaningful SNPs and imaging QTs. More importantly, DDG-MTSCCA could identify modality-consistent SNPs as well as modality-specific SNPs showing relevance to disorders. All of the results demonstrated the potential of DDG-MTSCCA in multi-modal brain imaging genetics research, as a promising tool to further help understand the mechanism of brain diseases.

2. Method

Herein, we denote scalars as italic letters, column vectors as boldface lowercase letters, and matrices as boldface uppercase ones. For a given matrix $\mathbf{M} = (m_{ij})$, m^i and m_j indicate the i th row and j-th column of the matrix, respectively. $\|\mathbf{m}\|_2$ denotes the Euclidean norm of the vector $\mathbf{m}$, $\|\mathrm{M}\|_F = \sqrt{\sum_i \sum_j m_{ij}^2}$ denotes the Frobenius norm, and $\|\mathbf{M}\|_{1,1} = \Sigma_i \Sigma_j |m_{ij}|$ denotes the $\ell_{1,1}$-norm of M. Specifically, we use the matrix $\mathrm{X} \in \mathbb{R}^{n \times p}$ to represent the genetic data with n subjects and p SNPs, and the matrix $\mathbf{Y}_c \in \mathbb{R}^{n \times q} (c = 1, \cdots, C)$ to represent the phenotype data with n subjects and q imaging QTs for the c-th imaging modality, where C is the number of imaging modalities (tasks).

2.1. Overview of the DDG-MTSCCA workflow

Fig. 1 presents the flowchart of DDG-MTSCCA, consisting of three components. First, we preprocessed and organized the multi-modal brain imaging, SNP and diagnostic data, and then fed them to the model. Second, the DDG-MTSCCA model was designed to explore the associations between SNPs and multi-modal imaging QTs, which included three key parts. The canonical correlation analysis (CCA) part with parameter decomposition was responsible for constructing the association between SNPs and QTs, while decoupling modality-consistent and -specific SNPs. The linear regression part was in charge of determining the diagnosis-related QTs. The sparse constraints assisted to obtain meaningful sparse feature subsets, including modality-consistent and -specific sparsity of decomposed parameters and GGL-penalty to select QT pairs with strong connections. Finally, the proposed method was applied to two kinds of neurodegenerative diseases: Alzheimer's disease (AD) and Parkinson's disease (PD). With the above processing steps, some specific biomarkers of AD and PD, including brain imaging QT and SNP features, can be obtained.

2.2. MTSCCA

The MTSCCA model identifies the bi-multivariate association between SNPs and multi-modal imaging QTs which is defined as follows:

$$\begin{aligned} &\min_{\mathbf{u}_c, \mathbf{v}_c} \textstyle\sum_{c=1}^{C} \|\mathbf{X}\mathbf{u}_c - \mathbf{Y}_c\mathbf{v}_c\|_2^2 \\ &s.t. \|\mathbf{X}\mathbf{u}_c\|_2^2 = 1, \|\mathbf{Y}_c\mathbf{v}_c\|_2^2 = 1, \Omega(\mathbf{U}) \le b_1, \Omega(\mathbf{V}) \le b_2, \forall c. \end{aligned} \tag{1}$$

In this model, $\mathbf{U} \in \mathbb{R}^{p \times C}$ denotes the canonical weight matrix associated with genetic data $\mathbf{X}$ and each $\mathbf{u}_c$corresponds to each sub-task $\mathbf{Y}_c$. $\mathbf{V} \in \mathbb{R}^{q \times C}$ denotes that associated with imaging QTs $\mathbf{Y}_j$and each $\mathbf{v}_c$ corresponds to each sub-task $\mathbf{Y}_c$. $\Omega(\mathbf{U})$ and $\Omega(\mathbf{V})$ are penalty functions that control the sparsity and can prevent overfitting as well.

As indicated above, though MTSCCA has shown good effectiveness in multi-modal brain imaging genetics, it is still insufficient.

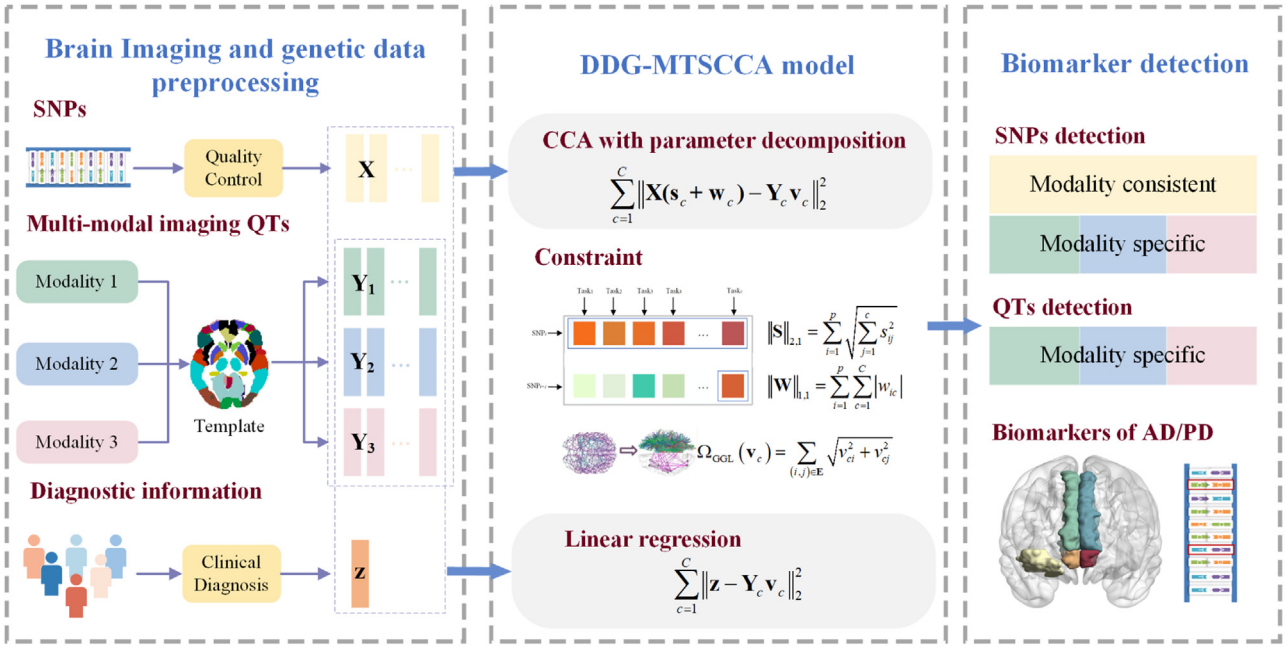

Fig. 1. The flowchart of the proposed DDG-MTSCCA method.

First, it is unsupervised which may achieve disease-irrelevant SNP-QT associations. Second, its halfway disentanglement of the shared and specific information may still lead to tangled results which is hard to interpret. Besides, it ignores the intrinsic network structure of the brain. Therefore, the present study improved this MTSCCA model based on the above considerations.

2.3. The proposed DDG-MTSCCA model

2.3.1. Diagnosis-guided MTSCCA

In order to make MTSCCA supervised, one regression component was added to Eq. (1) to make use of the diagnostic information. This diagnosis-guided MTSCCA model is defined as

$$\begin{aligned}&\min_{\mathbf{u}_c,\mathbf{v}_c} \textstyle\sum_{c=1}^{C} \|\mathbf{X}\mathbf{u}_c - \mathbf{Y}_c\mathbf{v}_c\|_2^2 + \sum_{c=1}^{C} \|\mathbf{z} - \mathbf{Y}_c\mathbf{v}_c\|_2^2 \\ &s.t. \|\mathbf{X}\mathbf{u}_c\|_2^2 = 1, \|\mathbf{Y}_c\mathbf{v}_c\|_2^2 = 1, \Omega(\mathbf{U}) \le b_1, \Omega(\mathbf{V}) \le b_2, \forall c,\end{aligned} \tag{2}$$

where $\mathbf{z} \in \mathbb{R}^{n\times 1}$ is a column vector corresponding to the diagnosis status of n subjects.

This model consists of two parts: the regression part and the MTSCCA part. Specifically, the regression part is responsible for identifying diagnosis-related imaging QTs, and the MTSCCA part is used to select relevant SNPs being related to imaging QTs obtained by the regression. Thus, the ultimately selected features are more reasonable than those unsupervised methods.

2.3.2. DDG-MTSCCA

The multi-modal imaging data usually characterizes an individual with different representations, and thus they could somehow share information as well as carry specific information associated with each modality. Both types of information could be useful for understanding the pathology of brain diseases. Moreover, this kind of heterogeneous but coupled information may play a crucial role in identifying risk loci. Thus, a diverse sparsity at both the individual level and group level is preferred for feature selection. Parameter decomposition is a good way to achieve this, i.e., decomposing the canonical weight of SNP into task-consistent and task-specific ones to identify the corresponding two kinds of features. Whereas, the tangle problem could still exist [23]. To overcome this issue, we further use an additional orthogonal constraint, which can well disentangle the shared and specific information. In addition, according to current neuroimaging studies, the brain operates in the form of a network. In other words, the brain regions prefer to work jointly rather than individually and separately. Then, it is meaningful to consider the network constraint in the model. To identify meaningful brain sub-networks, the graph-guided GGL-penalty was further added in our model for the brain imaging data [15].

Formally, the proposed DDG-MTSCCA model is given by:

$$\begin{aligned}&\min_{\mathbf{S},\mathbf{W},\mathbf{v}_c} \textstyle\sum_{c=1}^{C} \|\mathbf{X}(\mathbf{s}_c + \mathbf{w}_c) - \mathbf{Y}_c\mathbf{v}_c\|_2^2 + \sum_{c=1}^{C} \|\mathbf{z} - \mathbf{Y}_c\mathbf{v}_c\|_2^2 \\ &\quad + \beta_s\|\mathbf{S}\|_{2,1} + \lambda_w\|\mathbf{W}\|_{1,1} + \gamma \textstyle\sum_{c=1}^{C} \mathbf{s}_c^\top \mathbf{w}_c \\ &\quad + \lambda_{v_1} \textstyle\sum_{c=1}^{C} \|\mathbf{v}_c\|_1 + \lambda_{v_2} \sum_{c=1}^{C} \Omega_{\mathrm{GGL}}(\mathbf{v}_c) \\ &s.t. \quad \|\mathbf{X}(\mathbf{s}_c + \mathbf{w}_c)\|_2^2 = 1, \|\mathbf{Y}_c\mathbf{v}_c\|_2^2 = 1, \forall c,\end{aligned} \tag{3}$$

where β_s, λ_w, γ, λ_{v_1}, λ_{v_2} are regularization parameters to help identify those relevant features. Compared with Eq. (2), the DDG-MTSCCA decomposes the canonical weight $\mathbf{U}$ associated with genetic data into the task-consistent component $\mathbf{S} \in \mathbb{R}^{p\times C}$ and the task-specific component $\mathbf{W} \in \mathbb{R}^{p\times C}$ with $\mathbf{s}_c$ and $\mathbf{w}_c$ corresponding to each sub-task $\mathbf{Y}_c$, namely $\mathbf{U} = \mathbf{S} + \mathbf{W}$.

After parameter decomposition, distinct penalties are imposed on $\mathbf{S}$ and $\mathbf{W}$ respectively. Specifically, considering an individual SNP could be relevant to all modalities, the$\ell_{2,1}$-norm is imposed onto the shared component $\mathbf{S}$, i.e.

$$\|\mathbf{S}\|_{2,1} = \sum_{i=1}^{p} \|\mathbf{s}^i\|_2 = \sum_{i=1}^{p} \sqrt{\sum_{c=1}^{C} (s_{ic})^2} \tag{4}$$

to pursuit the task-consistent selection for SNPs. Meanwhile, some SNPs could be associated with only one specific task (modality). Accordingly, $\ell_{1,1}$-norm is imposed onto the task-specific component $\mathbf{W}$.

Besides, we impose an orthogonal constraint on $\mathbf{S}$ and $\mathbf{W}$ ($\mathbf{SW} = \mathbf{I}$) to guarantee the independence between $\mathbf{S}$ and $\mathbf{W}$, and to better decouple the task-consistent and -specific information simultaneously.

The GGL-penalty is used in our algorithm to embody the network connection information. It mainly takes advantage of both group lasso and graph-guided fused lasso, which is robust to the correlation directionality and requires no prior knowledge. The

GGL-penalty is defined as

$$\Omega_{\mathrm{GGL}}(\mathbf{v}_c) = \sum_{(i,j)\in\mathbf{E}_c} \sqrt{v_{ci}^2 + v_{cj}^2}, \tag{5}$$

where $\mathbf{E}_c$ is the edge set of the graph (network) of the c-th modality (task). Meanwhile, the ℓ_1-norm penalty is imposed on $\mathbf{v}_c$ in order to obtain individual level sparsity.

2.4. The optimization algorithm

As $\mathbf{S}$, $\mathbf{W}$, and $\mathbf{v}_c$ are jointly non-convex, it is difficult to solve Eq. (3) straightly. Fortunately, it is a bi-convex problem. Thus, we use the alternative convex search (ACS) strategy to solve Eq. (3). The detailed steps are described as follows.

2.4.1. Updating S and W

Firstly, we update $\mathbf{s}_c$ with $\mathbf{W}$, $\mathbf{v}_c(c = 1, \cdots C)$ and $\mathbf{s}_k(k \neq c)$ fixed. Considering $\mathbf{W}$, $\mathbf{v}_c$ and $\mathbf{s}_k$ as constants, we can rewrite the objective with respect to $\mathbf{s}_c$as

$$\min_{\mathbf{S}} \sum_{c=1}^{C} \|\mathbf{X}\mathbf{s}_c - \mathbf{Y}_c\mathbf{v}_c\|_2^2 + \beta_s\|\mathbf{S}\|_{2,1} + \gamma \sum_{c=1}^{C} \mathbf{s}_c^\top \mathbf{w}_c \tag{6}$$

Then we derive the Eq. (6) with respect to $\mathbf{s}_c$and set it to zero,

$$2\mathbf{X}^\top\mathbf{X}\mathbf{s}_c - 2\mathbf{X}^\top\mathbf{Y}_c\mathbf{v}_c + 2\beta_s\widetilde{\mathbf{D}}\mathbf{s}_c + \gamma\mathbf{w}_c = 0, \tag{7}$$

where $\widetilde{\mathbf{D}}$ is a diagonal matrix with i th diagonal element being $\frac{1}{2\|\mathbf{s}^i\|}(i = 1, \cdots, p)$. Lastly, $\hat{\mathbf{s}}_c$ can be obtained as follows

$$\hat{\mathbf{s}}_c = \left(\mathbf{X}^\top\mathbf{X} + \beta_s\widetilde{\mathbf{D}}\right)^{-1}\left(\mathbf{X}^\top\mathbf{Y}_c\mathbf{v}_c - \frac{1}{2}\gamma\mathbf{w}_c\right). \tag{8}$$

Similarly, with $\mathbf{S}$, $\mathbf{v}_c(c = 1, \cdots C)$ and $\mathbf{w}_k(k \neq c)$ fixed, we easily have

$$\hat{\mathbf{w}}_c = \left(\mathbf{X}^\top\mathbf{X} + \lambda_w\widetilde{\mathbf{D}}_c\right)^{-1}\left(\mathbf{X}^\top\mathbf{Y}_c\mathbf{v}_c - \frac{1}{2}\gamma\mathbf{s}_c\right) \tag{9}$$

by taking the derivative of Eq. (3) with respect to each $\mathbf{w}_c$ separately, and letting them be zero. $\widetilde{\mathbf{D}}_c$ here is a diagonal matrix, and its i th element being $\frac{1}{2|w_{ic}|}(i = 1, \cdots, p)$.

In order to satisfy the equality constraints of Eq. (3) [23], we scale $\hat{\mathbf{s}}_c$ and $\hat{\mathbf{w}}_c$ as follows.

$$\mathbf{s}_c = \frac{\hat{\mathbf{s}}_c}{\|\mathbf{X}(\hat{\mathbf{s}}_c + \hat{\mathbf{w}}_c)\|}, \ \mathrm{w}_c = \frac{\hat{\mathbf{w}}_c}{\|\mathbf{X}(\hat{\mathbf{s}}_c + \hat{\mathbf{w}}_c)\|} \tag{10}$$

2.4.2. Updating V_c

The $\mathbf{v}_c(c = 1, \cdots C)$ should be solved separately since each $\mathbf{v}_c$ is associated with each $\mathbf{Y}_c$ respectively. Following the same procedures of solving $\hat{\mathbf{s}}_c$ and $\hat{\mathbf{w}}_c$, we have

$$\hat{\mathbf{v}}_c = \left(\lambda_{v_1}\mathbf{D}_c + \frac{1}{2}\lambda_{v_2}\widetilde{\mathbf{D}}_c + 2\mathbf{Y}_c^\top\mathbf{Y}_c\right)^{-1}\mathbf{Y}_c^\top(\mathbf{X}\mathbf{u}_c + \mathbf{z}), \tag{11}$$

where $2\mathbf{D}_c\mathbf{v}_c$ is the subgradient of the ℓ_1-norm penalty of$\mathrm{v}_c$and $\widetilde{\mathbf{D}}_c\mathbf{v}_c$ is that of GGL-penalty. Specifically, $\mathbf{D}_c$ is a diagonal matrix with its i th element being $\frac{1}{2|\mathrm{v}_{ic}|}(i = 1, \cdots, p)$; $\widetilde{\mathbf{D}}_c$ is also a diagonal matrix with the k-th entry being $\sum_{j,j\neq k}\frac{1}{\sqrt{v_{kc}^2+v_{jc}^2}}(k \in [1, q_c], c \in [1, C])$. To satisfy the equality constraints in Eq. (3), we scale $\hat{\mathbf{v}}_c$by

$$\mathbf{v}_c = \frac{\hat{\mathbf{v}}_c}{\|\mathbf{Y}\hat{\mathbf{v}}_c\|}. \tag{12}$$

According to the ACS method, our model is optimized by updating $\mathbf{S}$, $\mathbf{W}$, and $\mathbf{v}_c$ alternatively. Algorithm 1 summarizes the optimization pseudocode of DDG-MTSCCA. These variables are updated in turn in each iteration until the algorithm converges or reaches a predefined stopping condition.

Algorithm 1
Algorithm to solve Eq. (3).

Require:
$\mathbf{X} \in \mathbb{R}^{n\times p}$, $\mathbf{Y}_c \in \mathbb{R}^{n\times q_c}$, $\mathbf{E}_c \in \mathbb{R}^{q_c\times q_c}$, $\beta_s, \lambda_w, \lambda_{v_1}, \lambda_{v_2}, \ \gamma, \ c \in [1, C]$
Ensure:
Canonical weights $\mathbf{S}$, $\mathbf{W}$, $\mathbf{v}_c$
Initialize $\mathbf{S} \in \mathbb{R}^{p\times C}$, $\mathbf{W} \in \mathbb{R}^{p\times C}$ and $\mathbf{v}_c \in \mathbb{R}^{q_c\times C}$
while not convergence do
  Update $\widetilde{\mathbf{D}}$ and solve $\hat{\mathbf{s}}_c$ according to Eq. (8)
  Update $\widetilde{\mathbf{D}}_c$ and solve $\hat{\mathbf{w}}_c$ according to Eq. (9)
  Solve $\mathbf{S}$ and $\mathbf{W}$ according to Eq. (10)
  Update $\mathbf{D}_c$ and $\widetilde{\mathbf{D}}_c$
  Solve v_c according to Eq. (11), and normalize v_c to $\|\mathbf{Y}_c\mathbf{v}_c\|_2^2 = 1$ according to Eq. (12)
end while

3. Experiments and results

3.1. Experimental setup

We compared the proposed DDG-MTSCCA method to the three most related methods including the conventional two-view SCCA [11], multi-view SCCA (mSCCA) [13], and MTSCCA to evaluate their performance in this study. For DDG-MTSCCA, it employs a multi-task learning paradigm, and generates canonical weight matrices $\mathbf{S}$ and $\mathbf{W}$ for SNP data and one canonical weight vector $\mathbf{v}_c$ for each specific modality of imaging QTs respectively. However, the two-view SCCA is a single-task model. When processing multi-model imaging data, the SCCA splits it into multiple single-task two-view SCCA models and calculates a canonical weight vector $\mathbf{u}_c$ and $\mathbf{v}_c$ for each task independently. Since mSCCA only learns one canonical weight vector u for SNP data, we stack the weight $\mathbf{u}$ several times to yield canonical weight matrix $\mathbf{U}$. Similarly, MTSCCA learns multiple SCCA tasks together, yielding canonical weight $\mathbf{U}$ for SNPs and canonical weight vector $\mathbf{v}_c$ for the c-th modality of imaging QTs.

Using a proportion of 6:2:2, we divide the data set into three parts, namely training set, validation set and testing set. The parameters in the model are tuned based on the training and validation set. Particularly, the set of parameters with the highest correlation coefficient is retained as the optimal parameter combination. Using the optimal parameters, the model is retrained on the combined training and validation set, and the final results are calculated on the testing set. Five parameters, i.e., $\beta_s, \lambda_w, \lambda_{v_1}, \lambda_{v_2}, \ \gamma$, are tuned for the DDG-MTSCCA model. Specifically, the first three parameters govern the sparsity level of the feature subsets, the fourth parameter controls the selected brain networks, and the last parameter restrains the independence of the shared and specific information. We apply the grid search strategy to tune those parameters from a moderate range $10^i(i = -5, -4, \ldots, 0, \ldots, 4, 5)$. To ensure efficiency, we set two stopping conditions, i.e.,$\max_c|(\mathbf{s}_c + \mathbf{w}_c)^{t+1} - (\mathbf{s}_c + \mathbf{w}_c)^t| \leq \varepsilon$ and $\max_c|\mathbf{v}_c^{t+1} - \mathbf{v}_c^t| \leq \varepsilon$, where t is the number of iterations, the estimation tolerance error $\varepsilon = 10^{-5}$and the maximum number of iterations was set to 100. Both stopping conditions could obtain good results experimentally. We repeat each experiment using the same experimental setup 100 times to ensure stable results. Finally, we show the average results by removing those results of failure trials (about 20 out of 100), since those failure ones are probably due to the inappropriate data partition.

3.2. Results on synthetic data

3.2.1. Data sources

In this section, we performed four experiments on the synthetic data. We generated four data sets with different numbers of samples, features, and noise levels. The first three data sets shared the

Table 1
Training and Testing CCCs (mean ± std) Estimated from Synthetic Data Sets.

Data set	Method	Training CCCs			Testing CCCs		
		Task 1	Task 2	Task 3	Task 1	Task 2	Task 3
Data 1	*SCCA*	0.94±0.01	0.90±0.02	0.93±0.01	0.35±0.09	0.41±0.20	0.59±0.19
	mSCCA	0.98±0.00	0.98±0.00	0.98±0.00	0.30±0.14	**0.53±0.20**	0.36±0.18
	MTSCCA	0.97±0.00	0.96±0.00	0.98±0.01	0.24±0.20	0.43±0.20	0.53±0.27
	DDG-MTSCCA	**0.99±0.00**	**0.99±0.00**	**0.99±0.00**	**0.38±0.15**	0.49±0.15	**0.54±0.20**
Data 2	*SCCA*	0.88±0.01	0.89±0.01	0.91±0.01	0.50±0.17	0.74±0.10	0.77±0.08
	mSCCA	0.96±0.00	0.96±0.01	0.96±0.01	0.46±0.18	0.60±0.14	0.62±0.14
	MTSCCA	0.97±0.00	0.98±0.00	0.98±0.00	0.45±0.18	0.65±0.11	0.74±0.09
	DDG-MTSCCA	**0.99±0.00**	**0.99±0.00**	**0.99±0.00**	**0.66±0.12**	**0.74±0.09**	**0.78±0.08**
Data 3	*SCCA*	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00
	mSCCA	0.99±0.00	0.99±0.00	0.99±0.00	0.96±0.01	0.97±0.01	0.98±0.01
	MTSCCA	0.99±0.00	0.99±0.00	0.99±0.00	0.96±0.02	0.98±0.01	0.98±0.01
	DDG-MTSCCA	**0.99±0.00**	**0.99±0.00**	**0.99±0.00**	**0.99±0.01**	**0.99±0.00**	**0.99±0.00**
Data 4	*SCCA*	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.01	0.99±0.00	0.99±0.00
	mSCCA	0.99±0.00	0.99±0.00	0.99±0.00	0.97±0.00	0.98±0.00	0.98±0.00
	MTSCCA	0.99±0.00	0.99±0.00	0.99±0.00	0.98±0.00	0.98±0.00	0.98±0.00
	DDG-MTSCCA	**0.99±0.00**	**0.99±0.00**	**0.99±0.00**	**0.99±0.01**	**0.99±0.00**	**0.99±0.00**

same ground truth but with different noise strengths, which could demonstrate the performance of one method under different noise intensities. The fourth data set was generated to access the performance under a high-dimensional situation. The details of each data set synthesis are described as follows.

Data 1: We set the number of subjects n to 60, the ground truth of SNP data to $\mathbf{u} = (\underbrace{0,\cdots,0}_{50},\underbrace{1,\cdots,1}_{40},\underbrace{0,\cdots,0}_{60})^\top$, and the ground truth of three imaging modalities (tasks) to $\mathbf{v}_1 = (\underbrace{0,\cdots,0}_{45},\underbrace{1,\cdots,1}_{30},\underbrace{0,\cdots,0}_{45})^\top$, $\mathbf{v}_2 = (\underbrace{0,\cdots,0}_{20},\underbrace{2,\cdots,2}_{20},\underbrace{0,\cdots,0}_{20},\underbrace{1,\cdots,1}_{20},\underbrace{0,\cdots,0}_{20})^\top$, $\mathbf{v}_3 = (\underbrace{1,\cdots,1}_{20},\underbrace{0,\cdots,0}_{30},\underbrace{2,\cdots,2}_{40},\underbrace{0,\cdots,0}_{30})^\top$ respectively. According to the population classification information, a latent vector $\mathbf{z} \in \mathbb{R}^{n\times 1}$ of length n with unit norm was generated. Then the data matrix X was generated by $x_{l,i} \sim N(z_l u_i, \sigma_x)$, and each $\mathbf{Y}_c$ by $(y_{l,j})_c \sim N(z_l v_j, \sigma_{y_c})$, where $\sigma = \sigma_x = \sigma_{y_1} = \sigma_{y_2} = \sigma_{y_3} = 5$ denotes the noise strength.

Data 2 ∼ Data 3: These two data sets were produced by utilizing the same settings as Data 1, but with different noise levels, that was$\sigma = \sigma_x = \sigma_{y_1} = \sigma_{y_2} = \sigma_{y_3} = 2$for Data 2 and $\sigma = \sigma_x = \sigma_{y_1} = \sigma_{y_2} = \sigma_{y_3} = 0.1$ for Data 3. Accordingly, the true correlation coefficients of these three data sets increased gradually.

Data 4: $\mathbf{n} = 500$, $\sigma = \sigma_x = \sigma_{y_1} = \sigma_{y_2} = \sigma_{y_3} = 0.1$, $\mathbf{u} = (\underbrace{0,\cdots,0}_{400},\underbrace{1,\cdots,1}_{200},\underbrace{0,\cdots,0}_{300},\underbrace{2,\cdots,2}_{100},\underbrace{0,\cdots,0}_{1000})^\top$, $\mathbf{v}_1 = (\underbrace{0,\cdots,0}_{300},\underbrace{1.5,\cdots,1.5}_{100},\underbrace{0,\cdots,0}_{200})^\top$, $\mathbf{v}_2 = (\underbrace{0,\cdots,0}_{250},\underbrace{1.5,\cdots,1.5}_{150},\underbrace{0,\cdots,0}_{150})^\top$, $\mathbf{v}_3 = (\underbrace{0,\cdots,0}_{250},\underbrace{1.5,\cdots,1.5}_{150},\underbrace{0,\cdots,0}_{150})^\top$. Similarly, the data matrix $\mathbf{X}$ was generated by $x_{l,i} \sim N(z_l u_i, \sigma_x)$and $\mathbf{Y}_c$ was generated by $(y_{l,j})_c \sim N(z_l v_j, \sigma_{y_c})$, with the latent vector z of length n.

3.2.2. Experiment results on synthetic data

We ran all methods on four synthetic data, and showed the mean and standard deviations (STDs) of training and testing canonical correlation coefficients (CCCs) for each task (modality) in Table 1. The CCC is widely used to evaluate the performance of CCA methods.

For Data 1, all methods exhibited significantly lower testing CCCs than their training CCCs, suggesting overfitting due to the high percentage of noise exerted in this data set. From the first data set to the third one, the testing CCCs of these three data sets increased gradually due to the decrease of noise strength, and the testing CCCs reached the same level as training CCCs in Data 3. The DDG-MTSCCA obtained the highest testing CCCs in Task 1 (0.38 ± 0.15) and Task 2 (0.49 ± 0.15) while SCCA achieved the highest testing CCCs in Task 3 for Data 1. DDG-MTSCCA both obtained the highest training and testing CCCs among all four methods in Data 2 and Data 3. These results suggest that our method has a greater advantage than the benchmarks under the low Signal to Noise Ratio (SNR) situation, owing to the modeling paradigm and utilization of diagnosis information. In the high-dimensional data set of Data 4, DDG-MTSCCA also obtained higher testing CCCs than mSCCA and MTSCCA for each task, comparable with SCCA.

For brain imaging genetics studies, selecting correct feature subsets is of great interest and importance. The heat maps in Fig. 2 showed the feature selection results of each method on the four synthetic data sets, and the ground truths were also presented in the first row for reference in each subfigure. From Fig. 2, it can be observed that the identified genetic and imaging features improve from Data 1 to Data 3 for all the methods, when considering their consistence with the ground truths. Compared with mSCCA and MTSCCA, DDG-MTSCCA holds better canonical profiles being consistent with the ground truths. Especially in Data 2 and Data 3, the canonical weight $\mathbf{U}$ of our method selects more comprehensive feature subsets than mSCCA and MTSCCA, indicating the advantages of parameter decomposition and multi-task modeling. Besides, in Data 3, the canonical weight $\mathbf{V}$ of DDG-MTSCCA could select the weak signal (in Task 2 and Task 3), while MTSCCA cannot, suggesting introducing diagnosis information could assist label-related feature selection. In the high-dimensional data set (Data 4), DDG-MTSCCA could also identify correct signal positions.

Moreover, in order to evaluate feature subsets intuitively, we calculated the hit rates of each canonical weight as shown in Fig. 3. Specifically, the hit rate is the proportion of real features in the first k features with larger weights, where k is the number of non-zero features of the ground truth. The higher hit rate means better identification performance. DDG-MTSCCA achieved the highest hit rates of canonical weights $\mathbf{u}$, $\mathbf{v}_1$, $\mathbf{v}_3$ on all four data sets and $\mathbf{v}_2$ on Data 3 and Data 4, expect less hit rates than SCCA for Task 2 on

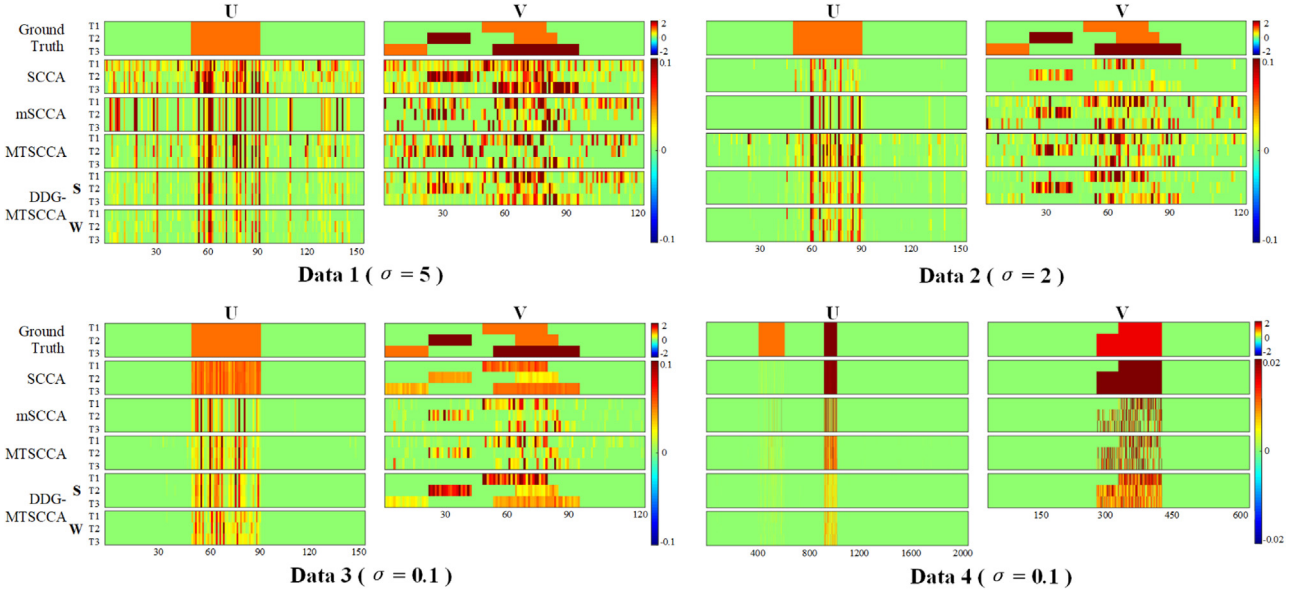

Fig. 2. Comparison of canonical weights in terms of each task for synthetic data sets. For each data set, the canonical weight **U** is shown on the left, and **V** is shown on the right. The top row shows the ground truth of **U** and **V**, and the remaining rows correspond to the methods: (1) SCCA; (2) mSCCA; (3) MTSCCA; (4) DDG-MTSCCA. Our method has two weights for **X** owing to the parameter decomposition, i.e., **S** and **W**. Within each panel, there are three rows corresponding to three SCCA tasks (denoted as T1∼T3).

Data 1 and Data 2. Moreover, the average hit rate of DDG-MTSCCA for the four data sets was the highest, about 25% higher than that of MTSCCA. The above results further highlight the anti-noise ability and recognition sensitivity of our method.

In summary, using data sets with different noise levels and different characteristics, this simulation study suggested the effectiveness of DDG-MTSCCA in the bi-multivariate association identification of multi-modality data. The parameter decomposition and diagnosis-guided strategy helped the identification of more comprehensive features, and what's more, our method could achieve task-consistent and -specific features while the other three methods could not.

3.3. Results on real neuroimaging genetic data

To further validate the effectiveness and reliability of our method, we performed and compared all four methods on real neuroimaging genetic data sets. Two independent real data sets were used, i.e., ADNI and PPMI, with the aim to discover potential brain regions and genetic variations associated with AD or PD.

3.3.1. Results on ADNI research database

1) Data Sources

We first used genotyping and multi-modal brain imaging data from the ADNI database (adni.loni.ucla.edu). This project aims to combine neuroimaging, clinical and neuropsychological assessment and other biological markers to investigate the progression of mild cognitive impairment (MCI) and early AD. The up-to-date information can be checked out at www.adni-info.org.

A total of 755 non-Hispanic Caucasian subjects participated in this experiment, including 182 healthy control (HC) subjects, 75 significant memory concern (SMC) subjects, 217 early mild cognitive impairment (EMCI) subjects, 184 late mild cognitive impairment (LMCI) subjects and 97 CE subjects. The explicit characteristics of the participants are listed in Table 2. We collected the neuroimaging data of three modalities of each subject, including 18F florbetapir positron-emission tomography (AV45 PET) scans, fluorodeoxyglucose positron-emission tomography (FDG PET) scans, and structural magnetic resonance imaging (sMRI) scans, and aligned the multi-modal imaging data to each other. For sMRI data, Statistical Parametric Mapping (SPM) software was used to perform voxel-based morphometry (VBM) processing on it. Specially, all the sMRI images were firstly registered to a T1-weighted template image, and then segmented into three parts: the gray matter (GM), the white matter (WM), and the cerebrospinal fluid (CSF) maps. These maps were normalized to the standard space of Montreal Neurological Institute (MNI) with the voxel size of $2 \times 2 \times 2$ mm^3, and smoothed with an 8 mm full-width-half-maximum (FWHM) kernel. For PET images, we co-registered the AV45 PET and FDG PET scans to the same MNI space. Lastly, based on the MarsBaR automated anatomical labeling (AAL) atlas, the sMRI and PET images were segmented into 116 regions of interest (ROIs). At the ROI level, we extracted the mean GM densities of sMRI scans, beta-amyloid depositions of AV45 PET scans, and glucose utilization of FDG PET scans for each ROI as three kinds of imaging QTs (represented as VBM, AV45 and FDG respectively). Then, for each imaging modality, a total of 116 QTs were acquired, representing 116 AAL brain regions.

In addition, we downloaded the genotyping data corresponding to each subject from the ADNI website. They were genotyped using the Human 610-Quad or Omni Express Array (Illumina, Inc., San Diego, CA, USA), and preprocessed by the standard quality control (QC) and imputation steps. According to the ANNOVAR annotation, we collected 4000 SNPs from neighbors of the AD risk gene *APOE* in chromosome 19 for experiment [24,25].

2) Experimental Results on ADNI

In this subsection, we applied the above four algorithms to the ADNI data to investigate the bi-multivariate associations between genetic data and three sets of imaging QTs. For simplicity, there were three tasks respectively denoted as SNP-AV45, SNP-FDG and SNP-VBM. Table 3 exhibited the training and testing CCCs for each task, and the averaged results for each method.

As shown in Table 3, the performance of DDG-MTSCCA surpassed that of all the other three methods across multiple tasks except the training results of SNP-VBM. The two-view SCCA obtained the highest training CCCs (0.30 ± 0.04) in the SNP-VBM

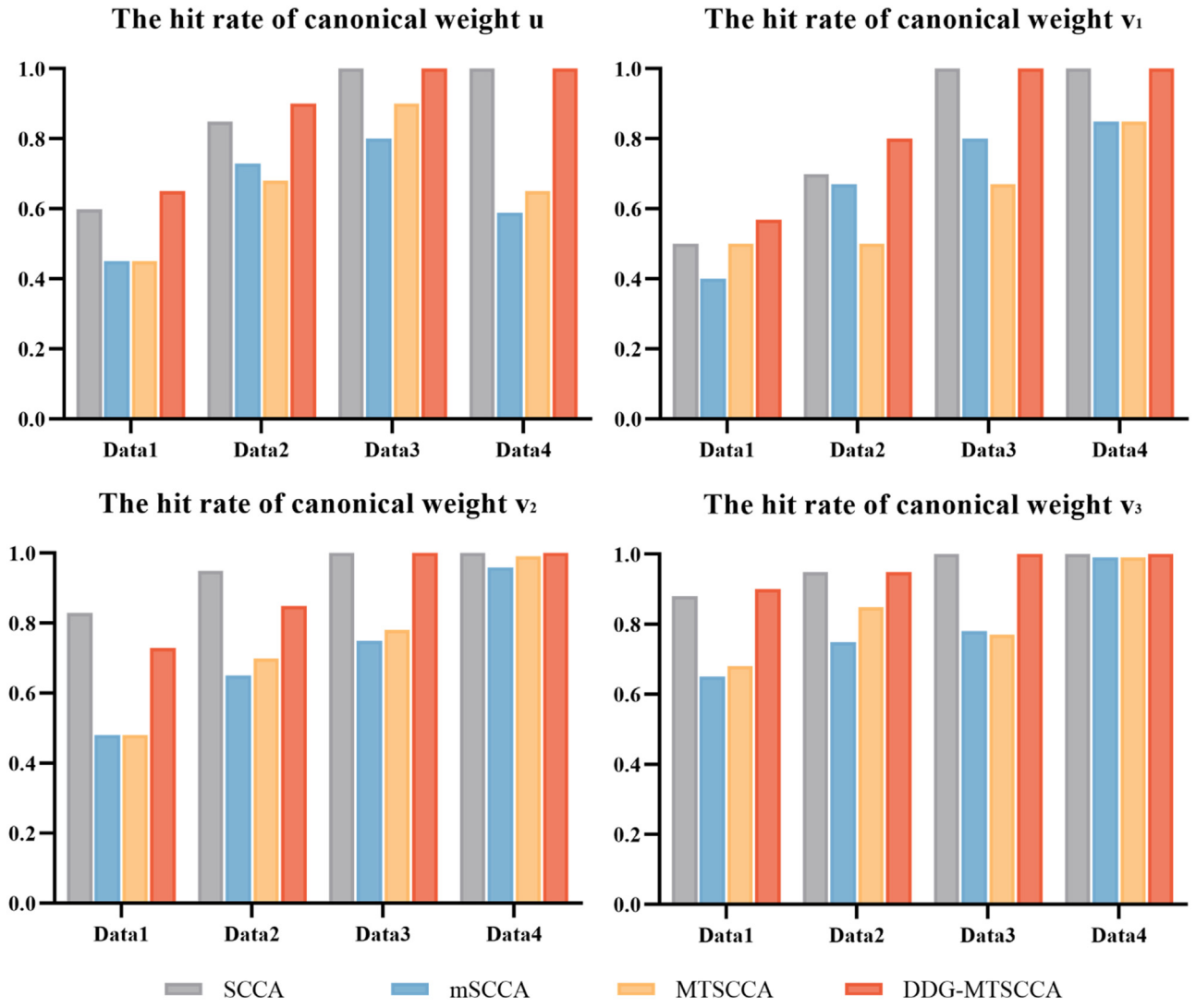

Fig. 3. Comparison of the hit rates of each canonical weight ($\mathbf{u}$, $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$) obtained from average results on synthetic data.

Table 2
ADNI participant characteristics.

	HC	SMC	EMCI	LMCI	AD
Num	182	75	217	184	97
Gender (M/F)	89/93	29/46	113/104	96/88	54/43
Handedness (R/L)	163/19	65/10	194/23	165/19	89/8
Age (mean±std)	73.93±5.51	71.77±5.76	70.59±7.16	71.89±7.92	73.99±8.44
Education (mean±std)	16.43±2.68	16.87±2.71	15.94±2.64	16.14±2.92	15.60±2.61

Table 3
CCCs (mean ± std) estimated between SNPs and imaging QTs of three modalities from ADNI data set.

		SNP-AV45	SNP-FDG	SNP-VBM	Average
Training	*SCCA*	0.42±0.01	0.29±0.01	**0.30±0.04**	**0.34±0.02**
	mSCCA	0.33±0.06	0.29±0.03	0.26±0.01	0.29±0.03
	MTSCCA	0.41±0.04	0.23±0.03	0.25±0.03	0.30±0.03
	DDG-MTSCCA	**0.45±0.01**	**0.31±0.01**	0.21±0.02	0.32±0.01
Testing	*SCCA*	0.42±0.06	0.29±0.06	0.09±0.06	0.27±0.06
	mSCCA	0.21±0.11	0.18±0.08	0.13±0.09	0.17±0.09
	MTSCCA	0.39±0.07	0.19±0.08	0.09±0.08	0.22±0.08
	DDG-MTSCCA	**0.47±0.06**	**0.33±0.06**	**0.20±0.07**	**0.33±0.06**

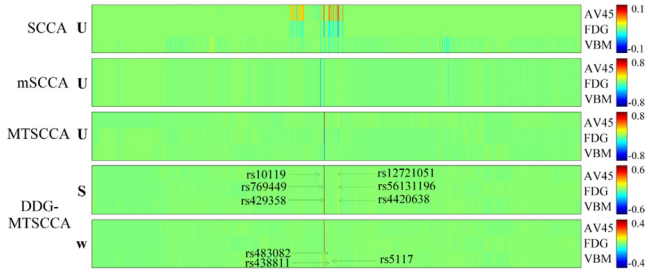

Fig. 4. The visualization of canonical weights of SNPs in terms of each task (i.e. AV45, FDG, VBM) for each method. From top to bottom: SCCA (**U**), mSCCA (**U**), MTSCCA (**U**), DDG-MTSCCA (**S**) and DDG-MTSCCA (**W**).

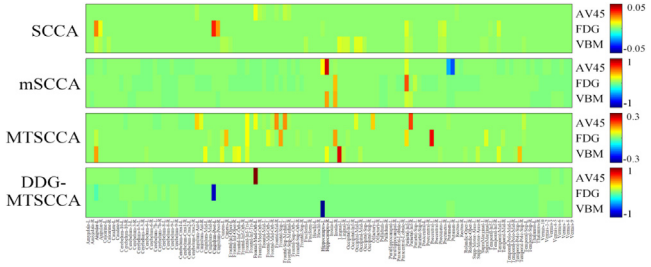

Fig. 5. The visualization of identified imaging QTs in each imaging modality (i.e. AV45, FDG and VBM) for each method. From top to bottom: SCCA, mSCCA, MTSCCA and the proposed DDG-MTSCCA.

Table 4
Top five imaging modality-specific SNPs selected by canonical weights of each imaging modality of DDG-MTSCCA except modality-consistent SNPs.

AV45	FDG	VBM
rs10414043	rs7256200	rs483082
rs7256200	rs10414043	rs3786497
rs769449	rs438811	rs438811
rs73052335	rs111789331	rs1727743
rs438811	rs73052335	rs5117

task, however, the corresponding testing CCCs (0.09 ± 0.06) was extremely low, suggesting that SCCA might fall into overfitting and further implying its limited performance in real neural data sets. Comparatively, the DDG-MTSCCA exhibited balanced training and testing CCCs for SNP-VBM, despite the relatively low values. Though mSCCA could simultaneously utilize three kinds of imaging QTs, its performance is degraded with lower CCCs than SCCA except the testing result of SNP-VBM task. This phenomenon could be attributed to its over-strict modeling strategy. That is, mSCCA requires the set of genetic data to be associated with three sets of imaging QTs at the same time [6]. As to MTSCCA, it achieved similar training and testing results on SNP-AV45 and SNP-VBM to two-view SCCA, while the training and testing CCCs of the SNP-FDG task were relatively small. In addition, DDG-MTSCCA achieved the highest average testing CCCs, about 50% higher than MTSCCA. To sum up, compared to the other three comparison methods, DDG-MTSCCA is outstanding across all three tasks.

Apart from the CCCs, identifying risk loci and imaging markers for AD will assist scientists or clinicians in exploring and developing more targeted treatment plans, which is a primary concern for imaging genetics. The heat maps in Fig. 4 show SNPs identified based on the amplitude of canonical weights for each method.

Obviously, compared to benchmarks, DDG-MTSCCA achieved much cleaner weight patterns, indicating its ability in identifying significant SNPs from massive markers. As expected, the notable AD risk marker rs429358 (*APOE*) could be identified by mSCCA, MTSCCA and our method, suggesting its essential correlation with AD. The SCCA performed unacceptable in this comparison as it did not find out this SNP. Due to the defects of the modeling strategies, only partial AD-related modality-consistent SNPs were selected by mSCCA and MTSCCA. Comparatively, benefitting from the parameter decomposition strategy, the DDG-MTSCCA method obtained both modality-consistent and -specific SNPs, and identified the most comprehensive AD-related markers. Specifically, the modality-consistent SNPs included rs10119 (*TOMM40*), rs769449 (*APOE*), rs12721051 (*APOC1*), rs56131196 (*APOC1*) and rs4420638 (*APOC1*). Table 4 lists the top five modality-specific SNPs of each imaging modality of our method. These selected SNPs were associated with the corresponding imaging modalities, e.g., rs483082 (*APOC1*) is highly correlated with VBM, enhancing the role of sMRI in identifying the loci of relevance. Importantly, the above identified modality-consistent and -specific SNPs have been reported to be AD-related previously [26–28].

What's more, the selected imaging QTs for the three modalities (AV45, FDG and VBM) of each method are also exhibited in Fig. 5. To make it clear, the distributions of the top five selected QTs of each modality selected by DDG-MTSCCA in the brain are shown in Fig. 6. The AV45 mode selected five imaging QTs located in the frontal lobe, indicating the amyloid deposition of the frontal lobe could well reflect the state of disease, which is consistent with the imaging studies [29]. The FDG identified the left post cingulum cortex, which is a sign of MCI [30]. And the significant reduction of glucose metabolism in the angular gyrus has been found to be associated with cognitive dysfunction [31]. Meanwhile, based on VBM, we identified atrophy of the bilateral hippocampus and also adjacent amygdala, which has been reported to be significant markers of AD and MCI [32]. The SCCA performed poorly since it did not find out the distinct mark related to AD such as the hippocampus. The mSCCA only identified a few AD-related QTs such as the bilateral hippocampus, but not nearly enough. The MTSCCA performed better than SCCA and mSCCA, which could identify the biomarkers we reported, but the results were disorganized and lacked a clear pattern of feature selection. The above results demonstrated that our method could effectively and clearly identify meaningful imaging QTs with the aid of parameter decomposition, network constraint and diagnosis-guided regression.

3.3.2. Results on PPMI research database

1) Data Sources

We use neuroimaging data and SNP genotyping data of subjects from the PPMI database (data downloaded via the PPMI website, www.ppmi-info.org) for further validation, which focuses on the efforts in identifying new potential biomarkers for PD progression and onset. Moreover, PPMI aims to enhance the development of new therapies and treatments for PD through longitudinal studies considering different types of data.

Neuroimaging Data: Throughout this study, we selected diffusion tensor imaging (DTI) and sMRI data since they have been shown to be proficient potential biomarkers for PD onset and progression [33,34]. Subjects were excluded if their neuroimaging data, subject information or genotyping data were missing. Finally, 100 non-Hispanic Caucasian participants were included in the experiment, including 35 HC and 65 PD participants. And the details of the participants characteristics are shown in Table 5.

The DTI and sMRI scans were processed by FMRIB's Software Library (FSL) software (http://www.fmrib.ox.ac.uk/fsl). For DTI data, we first corrected for motion artifacts and eddy current distortions by normalizing each volume to non-diffusion-weight volume (b0) utilizing FMRIB's Linear Image Registration Tool (FLIRT). Besides, by using the brain-extraction tool (BET), the brain masks of the b0 image were generated. Finally, we calculated the diffusion tensor with the FSL DTIFIT program for the whole brain and derived fractional

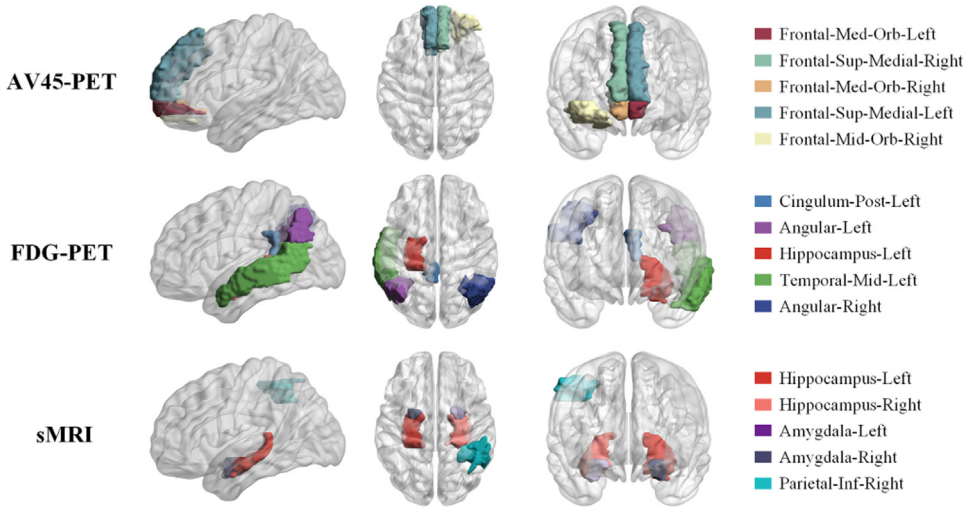

Fig. 6. Top five important brain regions (listed in descending order of the canonical weight on the right side of the picture) and overall distribution from the AV45-PET (first row), FDG-PET (second row), and sMRI imaging data of AD (last row).

Table 5
PPMI participant characteristics.

	HC	PD
Num	35	65
Gender (M/F)	27/8	43/22
Age (mean±std)	62.10±10.16	60.83±10.00
Education (mean±std)	16.80±2.45	16.68±1.56

anisotropy (FA) and mean diffusivity (MD) parameter maps, then nonlinearly registered each subject's parameter images to the FMRIB58_FA template in MNI space by using Advanced Normalization Tools (ANTs). Furthermore, we extracted the ROI-level DTI metrics based on 48 ROIs of JHU-ICBM-labels. Similarly, the sMRI scans were processed with VBM by FSL. After brain extraction and segmentation, all the native GM images were non-linearly registered into MNI space and smoothed using an FWHM of 6 mm. Based on the AAL atlas, the mean GM densities of 116 ROIs were also derived. Then, we obtained 116 VBM QTs, 48 FA QTs and 48 MD QTs for imaging data.

Genotyping Data: The genotyping data of the same population downloaded from the PPMI website were preprocessed using the standard quality control (QC), which includes the call rate check per subject and per SNP marker, gender check, the Hardy-Weinberg equilibrium test, and marker removal by the minor allele frequency. According to ANNOVAR annotation, 2000 SNPs collected from neighbors of the PD risk gene *TMEM175* were used in this paper [35,36].

2) Experimental Results on PPMI

The bi-multivariate association analysis between one set of SNP data and three sets of imaging QTs, i.e., SNP-FA, SNP-MD, and SNP-VBM, was performed using the four methods. Similarly, Table 6 shows the training and testing CCC results of each method. It can be observed that SCCA obtained the highest CCCs on the training set, but extremely low CCCs on the testing set for all tasks, indicating the existing overfitting problem. By comparison, the other methods exhibited relatively low but more comparable training and testing CCCs. mSCCA achieved the lowest CCCs in both training and testing sets of each task due to its suboptimal modeling strategy. For MTSCCA, it obtained a bit higher training CCCs than DDG-MTSCCA for each task. However, the latter yielded better testing results, the average of which is about 40% higher than that of MTSCCA. Besides, DDG-MTSCCA achieved the best testing CCCs

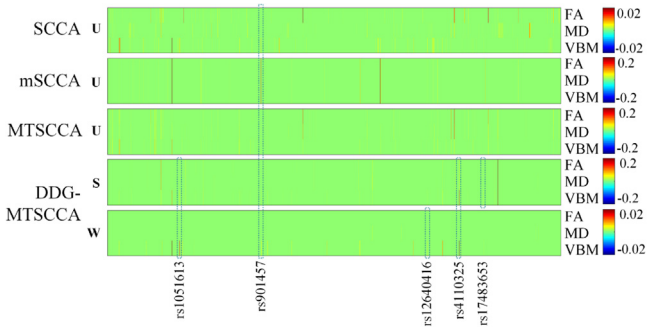

Fig. 7. The visualization of canonical weights of SNPs in terms of each task (i.e. FA, MD, VBM) for each method. From top to bottom: SCCA (**U**), mSCCA (**U**), MTSCCA (**U**), DDG-MTSCCA (**S**) and DDG-MTSCCA (**W**).

which were much close to the corresponding training CCCs, suggesting its relatively good performance. As a result, DDG-MTSCCA outperformed other benchmark methods on the PPMI data set when evaluating the performance of CCC.

The identified SNPs and imaging QTs were also investigated by the absolute values of canonical weights. The feature selection results of SNPs are shown in Fig. 7. In order to verify whether the identified SNPs are effective, SNPs with non-zero canonical weight were compared to the PD-related SNPs in the PDGene database [36,37]. Both the DDG-MTSCCA's modality-consistent component and the benchmark methods identified rs901457, which is an intergenic locus associated with PD. Additionally, DDG-MTSCCA independently identified PD-related loci, e.g., rs1051613 (*TMEM175*), rs12640416 (*GPRIN3*), rs4110325 (intergenic), and rs17483653 (*RNU1–138P*). Table 7 lists the top five modality-specific SNPs of each imaging modality of our method. Some SNPs have not been currently reported, but they might provide a novel clue, and further investigation should be warranted for this. Obviously, our method could identify more PD-related loci and produce much cleaner weight patterns than the benchmarks, illustrating the advantage of using the diagnosis status and parameter decomposition in comprehensive feature selection ability.

The heat maps of imaging QTs for each method are shown in Fig. 8. Compared with the baseline methods, we obtained a cleaner feature selection model, and those QTs with non-zero coefficients have been shown to be associated with the progression of PD. Similarly, the top five QTs of each modality detected by the DDG-

Table 6
CCCs (mean ± std) estimated between SNPs and imaging QTs of three modalities from PPMI data set.

		SNP-FA	SNP-MD	SNP-VBM	Average
Training	*SCCA*	**0.79±0.07**	**0.67±0.11**	**0.71±0.08**	**0.72±0.09**
	mSCCA	0.21±0.10	0.23±0.09	0.35±0.07	0.26±0.09
	MTSCCA	0.52±0.06	0.48±0.06	0.47±0.06	0.49±0.06
	DDG-MTSCCA	0.45±0.06	0.40±0.06	0.35±0.10	0.40±0.07
Testing	*SCCA*	0.20±0.13	0.14±0.11	0.18±0.13	0.17±0.12
	mSCCA	0.18±0.12	0.17±0.14	0.20±0.14	0.18±0.13
	MTSCCA	0.23±0.14	0.17±0.14	0.21±0.16	0.20±0.15
	DDG-MTSCCA	**0.34±0.15**	**0.27±0.13**	**0.22±0.12**	**0.28±0.13**

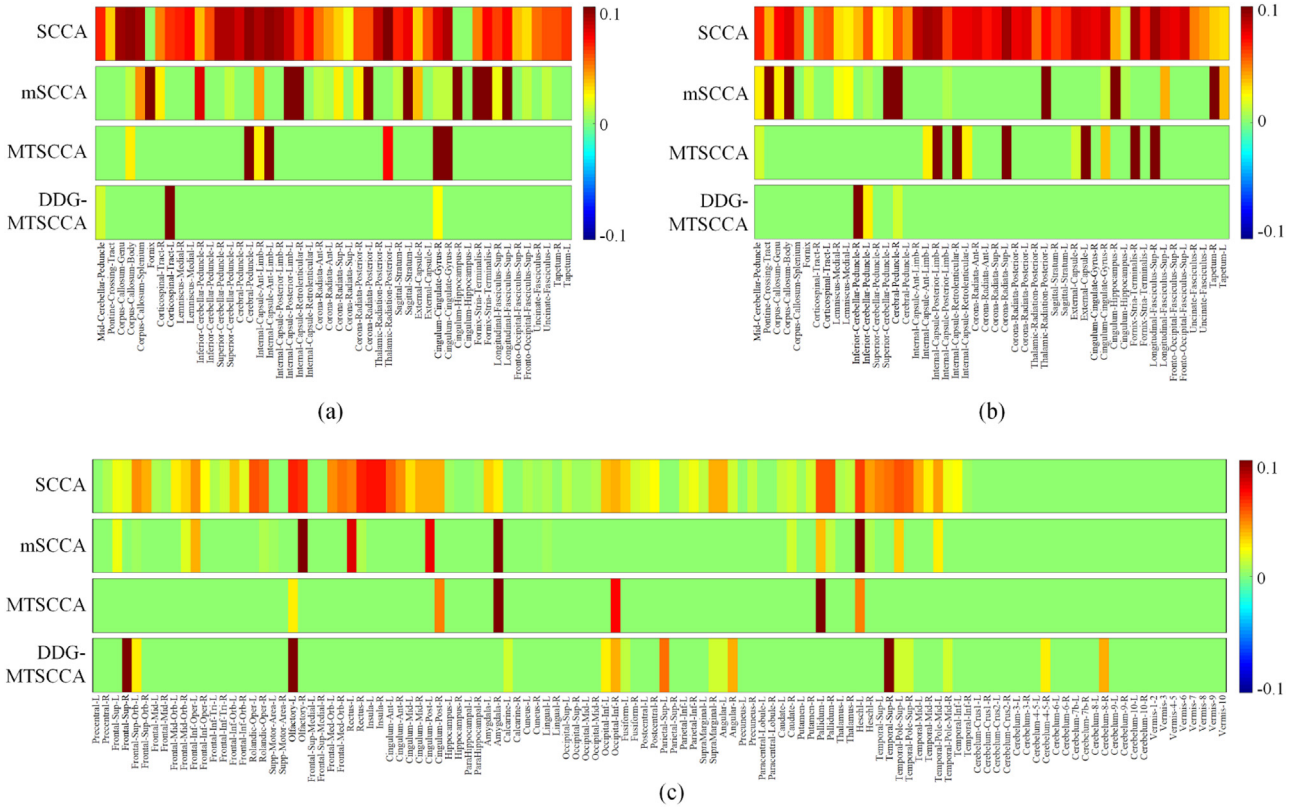

Fig. 8. The visualization of identified imaging QTs in each imaging modality, i.e. (a) FA, (b) MD and (c) VBM for each method. From top to bottom: SCCA, mSCCA, MTSCCA, and DDG-MTSCCA.

Table 7
Top five imaging modality-specific SNPs selected by Canonical weights of each imaging modality of DDG-MTSCCA.

FA	MD	VBM
rs4677722	rs10030417	rs7639719
rs9999619	rs324690	rs4110325
rs308435	rs324727	rs1051613
rs11945928	rs6818271	rs10026084
rs34048166	rs11098654	rs7439876

MTSCCA method are mapped to the brain, and their distribution is shown in Fig. 9. The alteration of left corticospinal tract has been reported in both FA and MD, a region of damage that has an important impact on motor disorders in PD patients [38]. The structural changes of the corpus callosum (MD) and cerebellum (FA and MD) seem to be mainly related to movement dysfunction and impulse control disorders. In addition, the cognitive status of PD patients is mainly related to the right cingulate gyrus (FA) injury [39]. The right superior temporal gyrus (VBM), the right superior frontal (VBM),and the left olfactory cortex (VBM) are hypothesized to play an important role in cognition, emotion, olfaction, and autonomic functions [40]. These results also demonstrated the meaning of our proposed method which introduced parameter decomposition, network constraint and diagnosis information.

3.4. Ablation experiments

Compared to MTSCCA, three components including the diagnosis-guide regression component, the parameter decomposition component, and the GGL-penalty network constraint were added to derive the DDG-MTSCCA. To further illustrate the effect of each additional item, we performed the ablation experiments. Specifically, we alternatively remove each component to generate three new models, namely noDG (no diagnosis information), noPD (no parameter decomposition) and noGGL (no GGL-penalty). The ADNI data set was used for comparison with the same experimental setup for each method. The training and testing results of CCCs of 80 trails per model are shown in Table 8.

It can be observed that DDG-MTSCCA achieves the highest training CCCs on both SNP-AV45 and SNP-FDG tasks, and the best testing CCCs only on the task SNP-AV45. On the remaining tasks,

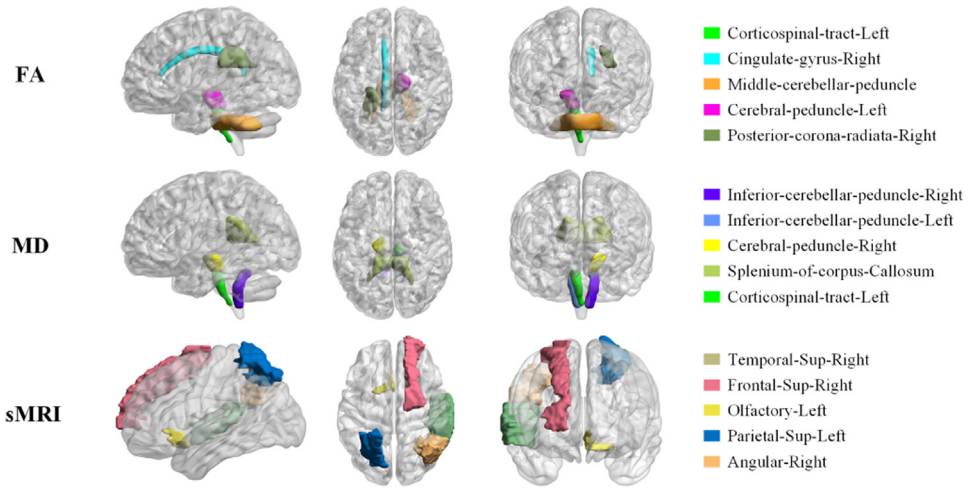

Fig. 9. Top five important brain regions (listed in descending order of the canonical weight on the right side of the picture) and overall distribution from the FA (first row), MD (second row), and sMRI imaging data of PD (last row).

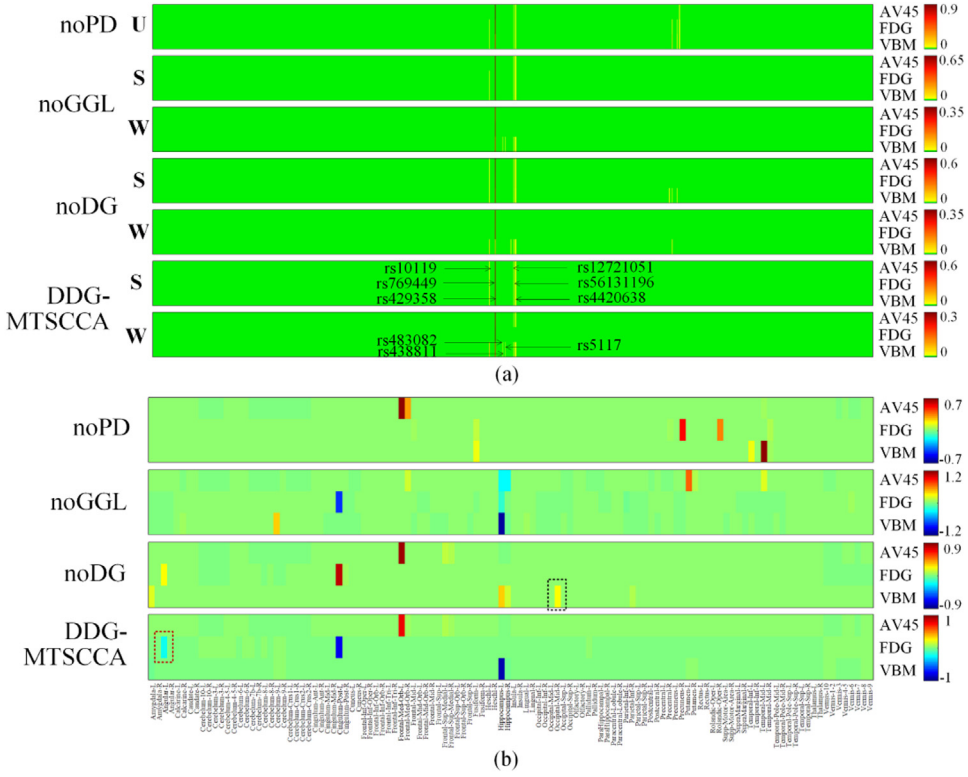

Fig. 10. Visualization of the canonical weights of SNPs (a) and imaging QTs (b) for the ablation experiments in terms of each task (i.e., AV45, FDG, VBM). From top to bottom: noPD, noGGL, noDG and the proposed DDG-MTSCCA.

Table 8
CCCs (mean ± std) of Ablation experiments.

		SNP-AV45	SNP-FDG	SNP-VBM	Average
Training	*noPD*	0.46±0.01	0.28±0.03	**0.28±0.02**	**0.34±0.02**
	noGGL	0.43±0.02	0.32±0.02	0.23±0.02	0.33±0.02
	noDG	0.46±0.01	0.32±0.01	0.24±0.02	0.34±0.02
	DDG-MTSCCA	**0.46±0.01**	**0.32±0.02**	0.21±0.02	0.33±0.02
Testing	*noPD*	0.47±0.05	0.24±0.07	0.12±0.07	0.28±0.06
	noGGL	0.44±0.05	**0.32±0.06**	**0.22±0.07**	**0.33±0.06**
	noDG	0.47±0.05	0.31±0.06	0.12±0.07	0.30±0.06
	DDG-MTSCCA	**0.47±0.05**	0.31±0.06	0.19±0.07	0.32±0.06

DDG-MTSCCA did not obtain the best score, implying that the prediction performance of our method has a bit decrease due to additional constraints. Nonetheless, our method still achieved relatively high scores across all tasks, and the average results among the three tasks were also similar to the top values.

To further explore the impact of each component, the feature selection results were further compared in Fig. 10. It is clear that the noPD model cannot distinguish between modality-consistent and modality-specific SNPs since the lack of parameter decomposition of the canonical weight **U**, thus only identifies one single feature selection (the first row in Fig. 10a). In our model, however, the significance of this division (the sixth and seventh rows in Fig. 10a) has been demonstrated. Secondly, our method identified the left angular (in the red box of Fig. 10b) region which has been reported to be associated with AD [41], whereas the noGGL model ignores this AD-related area. This may be because GGL-penalty retained a strong connection between this region and the left post cingulum, thus the corresponding weights tended to be consistent. What's more, we find that, compared with the noDG method, the QT pattern identified by DDG-MTSCCCA is much cleaner. It suggests that the noDG method might identify some AD-irrelevant QTs

Such as the left medial occipital (in the black box of Fig. 10b), demonstrating the necessity of introducing the diagnosis information into the model. In summary, all three components, including parameter decomposition, network connection and diagnosis status, play a crucial and indispensable role in improving the performance of the synthetic DDG-MTSCCA model.

4. Discussion

We have proposed the DDG-MTSCCA with joint consideration of diagnosis status, parameter decomposition and network connection constraints to identify the multi-SNPs-multi-QTs relationship for neurodegenerative disorders. The above results demonstrated that DDG-MTSCCA was better than comparison methods for the identification of risk genetic factors and abnormal brain regions. But there are still some limitations. First, more participants may be needed to reduce the overfitting risk. In practice, due to strict data requirements such as the data acquisition cost, data quality and patient burden, we could not obtain sufficient samples and imaging modalities. In view of this, we need to design new model to handle multi-modal data with missing modalities for small sample size. Second, though several meaningful SNPs and QTs could be identified, it is difficult to clearly indicate the number of significant features. Third, we only tested DDG-MTSCCA on a small set of SNPs due to the computational complexity of the GGL penalty. In the future research, more types of neuroimaging data (such as fMRI, other PET), omics data (such as gene expression, transcriptome, and proteome data) and clinical data may be included, in hope of identifying more comprehensive risk biomarkers. Besides, the performance of the whole-genome could be more interesting, which is also another future direction.

5. Conclusion

To help identify risk genetic factors and imaging QTs for neurodegenerative disorders, we proposed the DDG-MTSCCA method with joint consideration of diagnosis status, parameter decomposition and network connection constraints The feasibility of the algorithm was validated on both simulated and two independent real neuroimaging genetic data sets of ADNI and PPMI. Compared to SCCA, mSCCA and MTSCCA methods, DDG-MTSCCA method showed superior performance to the other three CCA models in canonical correlation and identifying disease-related SNPs and QTs.

In the proposed method, we added some key components on the basic MTSCCA model. Firstly, through the regression task, the diagnostic data were introduced to assist in identifying the disease-related imaging indicators, which not only made the biomarkers more meaningful, but also made the correlation coefficient between the detected images and the genetic markers higher. Secondly, the parameter decomposition strategy was adopted to decouple the consistent and specific information of multi-modal imaging data. With the sparsity constraint of the decomposed canonical weights and an additional orthogonal constraint to ensure their independence, more comprehensive SNPs could be identified for both AD and PD data sets, providing more reference to understanding the genetic mechanism of brain diseases. What's more, the network property of brain regions was also considered in our model. The GGL-penalty tended to identify QT pairs with strong connections, the impact of which has been validated by the ablation experiments.

In summary, DDG-MTSCCA identified stronger associations and obtained clearer feature selection patterns on simulated data sets and real data sets of AD and PD, and the recognized features are more biological significance, compared with competitive methods. These promising results revealed that the proposed model could be a powerful tool in brain imaging genetics, which is worthy of in-depth study.

Declaration of Competing Interest

The authors declared that they have no conflicts of interest to this work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61603399, 61973255), National Natural Science Foundation of Shaanxi Province (No. 2021JQ-119, 2020JM-142) and China Postdoctoral Science Foundation (2020T130537).

References

[1] L. Shen, et al., Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: a study of the ADNI cohort, Neuroimage 53 (3) (2010) 1051–1063.
[2] A.J. Saykin, et al., Genetic studies of quantitative MCI and AD phenotypes in ADNI: progress, opportunities, and plans, Alzheimer's Dement. 11 (7) (2015) 792–814.
[3] M. Kim, et al., Imaging genetics approach to Parkinson's disease and its correlation with clinical score, Sci. Rep. 7 (1) (2017) 1–10.
[4] A. Kawaguchi, F.J.B. Yamashita, Supervised multiblock sparse multivariable analysis with application to multimodal brain imaging genetics, Biostatistics 18 (4) (2017) 651–665.
[5] L. Shen, et al., Genetic analysis of quantitative phenotypes in AD and MCI: imaging, cognition and biomarkers, Brain Imaging Behav. 8 (2) (2014) 183–207.
[6] L. Du, et al., Multi-task sparse canonical correlation analysis with application to multi-modal brain imaging genetics, IEEE/ACM Trans. Comput. Biol. Bioinform. 18 (1) (2019) 227–239.
[7] M. Mroczek, et al., Imaging transcriptomics in neurodegenerative diseases, J. Neuroimaging 31 (2) (2021) 244–250.
[8] S. Bharati, et al., Dementia classification using MR imaging and clinical data with voting based machine learning models, Multimed. Tools Appl. 81 (18) (2022) 25971–25992.
[9] L. Du, et al., Identifying associations between brain imaging phenotypes and genetic factors via a novel structured SCCA approach, in: Proceedings of the International Conference on Information Processing in Medical Imaging, 2017, pp. 543–555. Springer.
[10] Q. Mai, X.J.B. Zhang, An iterative penalized least squares approach to sparse canonical correlation analysis, Biometrics 75 (3) (2019) 734–744.
[11] D.M. Witten, et al., A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics 10 (3) (2009) 515–534.
[12] J. Chen, et al., Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis, Biostatistics 14 (2) (2013) 244–258.
[13] D.M. Witten, R.J. Tibshirani, Extensions of sparse canonical correlation analysis with applications to genomic data, Stat. Appl. Genet. Mol. Biol. 8 (1) (2009).
[14] J. Fang, et al., Joint sparse canonical correlation analysis for detecting differential imaging genetics modules, Bioinformatics 32 (22) (2016) 3480–3488.

[15] X. Hao, et al., Mining outcome-relevant brain imaging genetic associations via three-way sparse canonical correlation analysis in Alzheimer's disease, Sci. Rep. 7 (1) (2017) 1–12.
[16] J.H. Sheng, et al., Strategies for multivariate analyses of imaging genetics study in Alzheimer's disease, Neurosci. Lett. 762 (2021).
[17] K. Wei, et al., An improved multi-task sparse canonical correlation analysis of imaging genetics for detecting biomarkers of Alzheimer's disease, IEEE Access 9 (2021) 30528–30538.
[18] S. Wang, et al., Identify biomarkers of Alzheimer's disease based on multi-task canonical correlation analysis and regression model, J. Mol. Neurosci. 72 (8) (2022) 1749–1763.
[19] X.M. Chen, et al., Structure-constraine d combination-base d nonlinear association analysis between incomplete multimodal imaging and genetic data for biomarker detection of neurodegenerative diseases, Med. Image Anal. 78 (2022) 102419.
[20] A. Jalali, et al., A dirty model for multiple sparse regression, IEEE Trans. Inf. Theory 59 (12) (2013) 7947–7968.
[21] E. Bullmore, O.J.N.R.N. Sporns, Complex brain networks: graph theoretical analysis of structural and functional systems, Nat. Rev. Neurosci. 10 (3) (2009) 186–198.
[22] K. Wei, et al., Integration of imaging genomics data for the study of Alzheimer's disease using joint-connectivity-based sparse nonnegative matrix factorization, J. Mol. Neurosci. 72 (2) (2022) 255–272.
[23] L. Du, et al., Associating multi-modal brain imaging phenotypes and genetic risk factors via a dirty multi-task learning method, IEEE Trans. Med. Imaging 39 (11) (2020) 3416–3428.
[24] J. Grimwood, et al., The DNA sequence and biology of human chromosome 19, Nature 428 (6982) (2004) 529–535.
[25] J.C. Venter, et al., The sequence of the human genome, Science 291 (5507) (2001) 1304–1351.
[26] L. Gao, et al., Shared genetic etiology between type 2 diabetes and Alzheimer's disease identified by bioinformatics analysis, J. Alzheimers Dis. 50 (1) (2016) 13–17.
[27] X. Zhou, et al., Non-coding variability at the APOE locus contributes to the Alzheimer's risk, Nat. Commun. 10 (1) (2019) 1–16.
[28] Q. Yan, et al., Genome-wide association study of brain amyloid deposition as measured by pittsburgh compound-B (PiB)-PET imaging, Mol. Psychiatry 26 (1) (2021) 309–321.
[29] M.J. Grothe, et al., *In vivo* staging of regional amyloid deposition, Neurology 89 (20) (2017) 2031–2038.
[30] L. Delano-Wood, et al., Posterior cingulum white matter disruption and its associations with verbal memory and stroke risk in mild cognitive impairment, J. Alzheimers Dis. 29 (3) (2012) 589–603.
[31] A. Hunt, et al., Reduced cerebral glucose metabolism in patients at risk for Alzheimer's disease, Psychiatry Res. Neuroimaging 155 (2) (2007) 147–154.
[32] J. Yang, et al., Voxelwise meta-analysis of gray matter anomalies in Alzheimer's disease and mild cognitive impairment using anatomic likelihood estimation, J. Neurol. Sci. 316 (1–2) (2012) 21–29.
[33] C. Atkinson-Clement, et al., Diffusion tensor imaging in Parkinson's disease: review and meta-analysis, Neuroimage Clin. 16 (2017) 98–110.
[34] P.J.T.r. Tuite, Magnetic resonance imaging as a potential biomarker for Parkinson's disease, Transl. Res. 175 (2016) 4–16.
[35] D. Chang, et al., A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci, Nat. Genet. (2017).
[36] M.A. Nalls, et al., Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease, Nat. Genet. 46 (9) (2014) 989–993.
[37] C.M. Lill, et al., Comprehensive research synopsis and systematic meta-analyses in Parkinson's disease genetics: the PDGene database, PLoS Genet. 8 (3) (2012) e1002548.
[38] K.I. Taylor, et al., Progressive decline in gray and white matter integrity in de novo Parkinson's disease: an analysis of longitudinal Parkinson progression markers initiative diffusion tensor imaging data, Front. Aging Neurosci. 10 (2018) 318.
[39] C. Atkinson-Clement, et al., Diffusion tensor imaging in Parkinson's disease: review and meta-analysis, Neuroimage Clin. 16 (2017) 98–110.
[40] P. Pan, et al., Voxel-wise meta-analysis of gray matter abnormalities in idiopathic Parkinson's disease, Eur. J. Neurol. 19 (2) (2012) 199–206.
[41] S.H. Lee, et al., Tract-based analysis of white matter degeneration in Alzheimer's disease, Neuroscience 301 (2015) 79–89.