# Fused multi-modal similarity network as prior in guiding brain imaging genetic association

Bing He[1], Linhui Xie[2], Pradeep Varathan[3], Kwangsik Nho[3], Shannon L. Risacher[3], Andrew J. Saykin[3], Jingwen Yan[1]* and the Alzheimer's Disease Neuroimaging Initiative[†]

[1]Luddy School of Informatics, Computing, and Engineering, Indiana University Purdue University Indianapolis (IUPUI), Indianapolis, IN, United States, [2]School of Engineering and Technology, Indiana University Purdue University Indianapolis (IUPUI), Indianapolis, IN, United States, [3]Indiana Alzheimer's Disease Research Center, Indianapolis, IN, United States

**Introduction:** Brain imaging genetics aims to explore the genetic architecture underlying brain structure and functions. Recent studies showed that the incorporation of prior knowledge, such as subject diagnosis information and brain regional correlation, can help identify significantly stronger imaging genetic associations. However, sometimes such information may be incomplete or even unavailable.

**Methods:** In this study, we explore a new data-driven prior knowledge that captures the subject-level similarity by fusing multi-modal similarity networks. It was incorporated into the sparse canonical correlation analysis (SCCA) model, which is aimed to identify a small set of brain imaging and genetic markers that explain the similarity matrix supported by both modalities. It was applied to amyloid and tau imaging data of the ADNI cohort, respectively.

**Results:** Fused similarity matrix across imaging and genetic data was found to improve the association performance better or similarly well as diagnosis information, and therefore would be a potential substitute prior when the diagnosis information is not available (i.e., studies focused on healthy controls).

**Discussion:** Our result confirmed the value of all types of prior knowledge in improving association identification. In addition, the fused network representing the subject relationship supported by multi-modal data showed consistently the best or equally best performance compared to the diagnosis network and the co-expression network.

## 1. Introduction

Brain imaging genetics studies the influence of genetic variation on brain structure and function. Its major task is to examine the association between genetic markers such as single nucleotide polymorphisms (SNPs) and quantitative traits (QTs) extracted from multi-modal neuroimaging data (e.g., MRI and PET scans). Although both gene and imaging phenotype are two well-known factors contributing to brain function, exploring their underlying connections would lead to a better mechanistic understanding of normal or disordered brain functions.

Early studies in brain imaging genetics associations typically adopt a univariate approach (Shen et al., 2010), where each pair of SNP and brain phenotype were examined individually for the association. Based on the assumption that a real imaging genetic association typically involves a small subset of SNPs and QTs, bi-multivariate association models, such as sparse canonical correlation analysis (SCCA), have been increasingly used later to identify the best linear transformation for imaging and genetics features so that the correlation between

imaging and genetic components can be maximized (Chi et al., 2013; Lin et al., 2014). Recently, to further improve the performance, various prior knowledge, such as diagnosis group, linkage disequilibrium block in SNPs, and brain co-expression networks, have been incorporated into the SCCA model. These prior knowledge mitigates the effect of limited sample size and all of them have helped yield much improved performance over the traditional SCCA model. However, these prior knowledge are not always available or sometimes not applicable. For example, the brain co-expression network used by Yan et al. (2014) requires a predefined subset of genes related to brain imaging, which would not be available for structural MRI. Some imaging genetics studies may have to deal with data without diagnosis information or with a single diagnosis group. In both cases, diagnosis information cannot be used as prior. Data-driven subject similarity network has been previously explored but was derived from a single modality with limited guidance (Du et al., 2016).

To address this problem, we propose a multi-modal subject similarity network as a new prior knowledge using the similarity network fusion (SNF) approach. In particular, we aim to build a subject similarity network that is supported by both brain imaging phenotype and genetic variants. Then, we will employ a discriminative SCCA model (Yan et al., 2017) to identify a subset of SNPs and brain imaging ROIs that are not only highly correlated but also can best explain the shared similarity network. When applied to the real brain imaging (including amyloid and tau PET) and genetic data in the ADNI cohort, we found that SCCA guided by the fused similarity network showed similar performance as that guided by diagnosis information and both outperformed those guided by other prior knowledge. Taken together, our results suggested the value of a fused similarity network as a great alternative prior in case of the absence of diagnostic network, particularly when the study focuses only on one group (like the aging process of healthy older adults).

## 2. Data

Amyloid and tau PET imaging data, together with the imputed genotype data, were downloaded from the Alzheimer's disease Neuroimaging Initiative (ADNI) (http://adni.loni.usc.edu/) database. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org. In this study, we have 800 subjects with both genotype data and amyloid imaging, including 158 cognitive normals (CN), 90 with significant memory concern (SMC), 279 early mild cognitive impairment (EMCI), 143 late MCI (LMCI), and 130 AD patients. For Tau, we have 291 subjects with both genotype data and tau imaging data, including 75 CN, 135 SMC, 30 EMCI, 32 LMCI, and 19 AD patients. The detailed demographic information of gender, age, and education years are shown in Table 1.

## 2.1. Imaging data preprocessing

Both amyloid and tau imaging data have been downloaded from the ADNI website as preprocessed. Briefly, amyloid PET used florbetapir (18F) as a tracer to measure amyloid-$\beta$ (A$\beta$) plaques (Okamura and Yanai, 2010). For each subject, brain regions of interest (ROIs) were defined from structural MRI through segmentation and parcellation using Freesurfer (Fischl, 2012). Then, each florbetapir scan was coregistered to the corresponding MRI and calculated the mean florbetapir uptake within the predefined ROIs. All the regional amyloid deposition was re-normalized using the whole cerebellum as a reference region. Tau PET used flortaucipir as a tracer to detect the aggregated tau (Fleisher et al., 2020), and the regional tau aggregation was obtained similarly as amyloid. All the regional tau tangle accumulation was re-normalized using inferior cerebellar as reference region. Finally, we have amyloid measurement in 68 cortical ROIs and tau measurement in 73 ROIs. More detailed image processing information can be found in Landau et al. (2013) and Landau et al. (2016). To remove potential bias, both amyloid and tau measures were pre-adjusted using baseline age, gender, and the weight derived from healthy controls. Finally, they were normalized to zero mean and unit variance for subsequent analysis.

## 2.2. Genotype data processing

Genotype data of both ADNI-1 and ADNI-2/GO phases were also obtained from the ADNI cohort (adni.loni.usc.edu). We focused our analysis on top SNPs from the International Genomics of Alzheimer's Project (IGAP), a large-scale genome-wide association study of AD (Schellenberg and IGAP, 2012). It tested the association of 7,055,881 single nucleotide polymorphisms (SNPs) of 17,008 Alzheimer's disease cases and 37,154 controls. SNPs with $p \leq 5 \times 10^{-6}$ in their meta analysis were used as our candidates and their genotypes were extracted based on the quality controlled and imputed genetic data in the ADNI using PLINK (Purcell et al., 2007). Finally, we have 1,080 SNPs for the subsequent imaging genetics association.

## 3. Methods

To evaluate the proposed prior knowledge, we apply it to amyloid and tau imaging and genetic data in the study of Alzheimer's disease (AD). Deposition of amyloid-$\beta$ and abnormal accumulations of tau protein are two major hallmarks in AD pathogenesis. Prior knowledge tested in this analysis for comparison include (1) subject diagnosis information, (2) brain co-expression network using amyloid- and tau-related genes, respectively (Zeng et al., 2012), and (3) fused similarity network built on imaging and genetics data.

### 3.1. Fused similarity network

In this study, we proposed to use a fused similarity network as a new prior knowledge, as inspired by Wang et al. (2014)

TABLE 1 Demographic information of ADNI image data.

| | Subjects | NC | SMC | EMCI | LMCI | AD |
|---|---|---|---|---|---|---|
| Amyloid | Number | 158 | 90 | 279 | 143 | 130 |
| | Gender (M/F) | 79/79 | 36/54 | 158/121 | 79/64 | 78/52 |
| | Age (mean ± SD) | 73.25 ± 6.05 | 71.62 ± 5.45 | 71.05 ± 7.27 | 71.41 ± 7.49 | 73.95 ± 8.03 |
| | Educ (mean ± SD) | 16.62 ± 2.50 | 16.79 ± 2.62 | 16.09 ± 2.66 | 16.71 ± 2.51 | 15.72 ± 2.69 |
| TAU | Number | 75 | 135 | 30 | 32 | 19 |
| | Gender (M/F) | 27/48 | 55/80 | 18/12 | 20/12 | 12/7 |
| | Age (mean ± SD) | 69.25 ± 5.40 | 70.96 ± 6.15 | 70.2 ± 7.23 | 71.97 ± 8.64 | 73.42 ± 10.80 |
| | Edu (mean ± SD) | 17.12 ± 2.14 | 16.84 ± 2.21 | 16.03 ± 2.68 | 15.94 ± 2.23 | 16.16 ± 2.75 |

and hypothesize that it will help improve the performance of imaging genetics association. First, we have original SNP data and imaging data showed in Figure 1A, we build a sample-sample similarity matrix for imaging and genetic data, respectively (Figure 1B) and their subject similarity network look like in Figure 1C. This similarity matrix can be seen as a similarity network G = (V,E,W), where nodes V represent subjects $\{x_1, x_2, ..., x_n\}$, the weighted edges $E$ represent similarities of a subject to others and W is a $n \times n$ similarity weighted matrix representing the similarity between subjects $x_i$ and $x_j$. Suppose $\rho(x_i, x_j)$ is euclidean distance between subjects $x_i$ and $x_j$. Then a scaled exponential similarity kernel was used to determine the weight of the edge:

$$\mathbf{W}(i,j) = \exp\left(-\frac{\rho^2(x_i, x_j)}{\mu \varepsilon_{i,j}}\right) \quad (1)$$

where $\mu$ is a hyper parameter that can be empirically set. It was recommended from [0.3, 0.8], and we set it as 0.5 by default (Wang et al., 2014). and $\varepsilon_{i,j}$ is used to eliminate the scaling problem. Here we define:

$$\varepsilon_{i,j} = \frac{\text{mean}(\rho(x_i, N_i)) + \text{mean}(\rho(x_j, N_j)) + \rho(x_i, x_j)}{3} \quad (2)$$

where $N_i$ denote a set of $x_i$'s neighbors including $x_i$ in $G$, and $\rho(x_i, N_i)$ is the average value of the distance between $x_i$ and each of its neighbors. Each row of $\mathbf{W}$ was then normalized as below:

$$\mathbf{P}(i,j) = \begin{cases} \frac{\mathbf{W}(i,j)}{2\Sigma_{k \neq i}\mathbf{W}(i,k)}, & j \neq i \\ 1/2, & j = i \end{cases} \quad (3)$$

Given a graph $G$, we use $K$ nearest neighbors (KNN) to measure local affinity as:

$$\mathbf{S}(i,j) = \begin{cases} \frac{\mathbf{W}(i,j)}{\Sigma_{k \in N_i}\mathbf{W}(i,k)}, & j \in N_i \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$\mathbf{P}$ offers the similarity information of each subject to all others and $\mathbf{S}$ encodes the similarity to the $\mathbf{K}$ most similar neighbors for each subject. In this article, we have two types of data, genomics data and imaging data. We first calculated the status

matrices $P^{(1)}$ and $P^{(2)}$ following equation (3), and then the kernel matrices $S^{(1)}$ and $S^{(2)}$ following in equation (4). For both amyloid and tau data, we tested the model performance with varying $K$ values from 5 to 50. Association performance was found highly stable across varying $K$ values, therefore we set $K = 20$ as default.

Next, we performed the network fusion of two kernel matrices using a message-passing theory (Pearl, 1988) non-linear method. This is an iterative process where both networks keep getting updated until they converge (i.e., not change much). The final network, known as the fused network, is expected to represent the subject relationships supported by both brain image data and genotype data. Let $P_{t=0}^{(1)} = P^{(1)}$ and $P_{t=0}^{(2)} = P^{(2)}$ be the initial two status matrices when $t = 0$. The fusion process will iteratively update two similarity matrices corresponding to two data types as follows:

$$\mathbf{P}_{t+1}^{(1)} = \mathbf{S}^{(1)} \times \mathbf{P}_t^{(2)} \times \left(\mathbf{S}^{(1)}\right)^T \quad (5)$$

$$\mathbf{P}_{t+1}^{(2)} = \mathbf{S}^{(2)} \times \mathbf{P}_t^{(1)} \times \left(\mathbf{S}^{(2)}\right)^T \quad (6)$$

where $\mathbf{P}_{t+1}^{(1)}$ and $\mathbf{P}_{t+1}^{(2)}$ are the status matrix of these two data types after t iterations. After each iteration, we performed normalization on $\mathbf{P}_{t+1}^{(1)}$ and $\mathbf{P}_{t+1}^{(2)}$ as in equation (3). This step ensures that subject self-similarity is always higher than the similarity to other neighbors. Here, the alternating multiplication of the squared KNN similarity of the two modalities essentially combines the local information of the two modalities in a way that reinforces their shared information. By multiplying the initial similarity matrix of one modality with the squared KNN similarity matrix of the other modality, the shared information between the two modalities is amplified and the unique information in each modality is retained. This process is then repeated in an alternating manner to ensure that both modalities contribute equally to the final similarity matrix, thereby achieving a balanced fusion of the two modalities (The fused network as showed in Figure 1D). This approach is expected to result in a more informative similarity matrix that captures the shared and unique features of both modalities, which in turn can improve the performance of downstream analysis such as association identification between brain imaging and genetic features.

**FIGURE 1**
Workflow of similarity network fusion. **(A)** Original SNP data and imaging data. **(B)** Subject similarity matrix generated using normalized mutual information (NMI). **(C)** Subject similarity network (equivalent to the subject by subject matrix). **(D)** Fused network by integrating information from both imaging and genotype data.
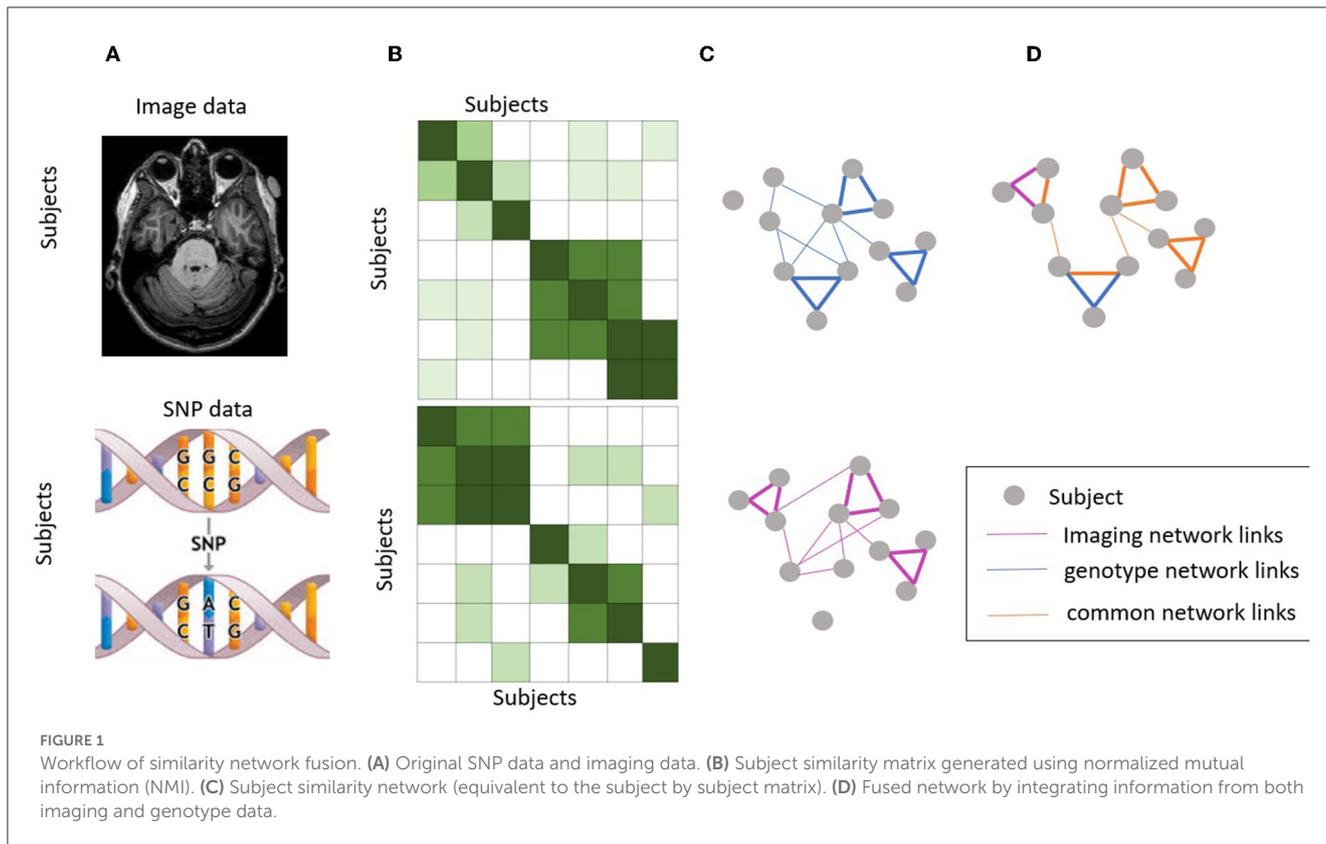
## 3.2. Prior knowledge for comparison

### 3.2.1. Diagnosis network

A similarity matrix based on diagnosis was built by assigning 1 s between samples in the same diagnosis and 0 s otherwise. In other words, we build a complete graph for all the subjects belonging to the same diagnosis group. To ensure $\Sigma_j \mathbf{P}(i,j) = 1$, it was then normalized by setting the diagonal entries as 0.5, and other elements as 0.5 divided by the group size.

$$\mathbf{P}(i,j) = \begin{cases} \frac{1}{2 \cdot Length(Group(k))}, j \neq i \\ 1/2, j = i \end{cases} \qquad (7)$$

where $Length(Group(k))$ means the size of diagnosis group, and there are totally five groups in this article.

### 3.2.2. Brain co-expression network

We use amyloid as an example to demonstrate the co-expression network construction process. We first identified 15 genes related to amyloid pathways according to previous studies (Swaminathan et al., 2012). We then extracted the expression level of these genes across 1,210 brain samples in the Allen Human Brain Atlas (AHBA) database. A partial correlation analysis was performed on the brain expression data, and generated a $1,210 \times 1,210$ matrix indicating the ROI–ROI similarity based on the expression of selected genes. This matrix was later down-sampled to $68 \times 68$, where all 1,210 brain samples were mapped to amyloid ROIs and the median value was applied to aggregate the similarity

measures. For tau, there are eight genes found involved in tau phosphorylation pathway (Bekris et al., 2012). We went through the same process and generated a $73 \times 73$ co-expression matrix for tau. These two matrices were used as the prior knowledge in subsequent analysis.

## 3.3. Discriminative SCCA

In this part, let $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\} \subseteq \boldsymbol{R}^p$ be the imaging data and $Y = \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_n\} \subseteq \boldsymbol{R}^q$ be the genotype data, where n is the number of patients, $p$ and $q$ are the numbers of ROIs and SNPs, respectively. Sparse canonical correlation analysis (SCCA) aims to find the maximal correlation between $Xu$ and $Yv$ by adjusting these two weights, u and v, which indicates the significance of each feature of the imaging genetic associations. As shown in this formula:

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T \mathbf{X}^T \mathbf{Y} \mathbf{v}$$
$$\text{s.t.} \mathbf{u}^T \mathbf{X}^T \mathbf{X} \mathbf{u} = 1, \mathbf{v}^T \mathbf{Y}^T \mathbf{Y} \mathbf{v} = 1, \boldsymbol{P}_1(u) \leq c1, \boldsymbol{P}_2(v) \leq c2 \qquad (8)$$

where $\boldsymbol{P}_1(u) \leq c1 and \boldsymbol{P}_2(v) \leq c2$ are two penalty terms to control the sparsity of selected features. In this study, we used the PMA package (Witten et al., 2009) that applied the $L_1$ norm penalty for $P_1$ and $P_2$ constraints to perform the SCCA method. To ensure the selection of disease-relevant features, we used a novel discriminative SCCA (DSCCA) algorithm (Yan et al., 2017) to integrate imaging data, SNPs data and the prior knowledge for imaging genetics association. Prior knowledge can be diagnosis network, fusion network or ROI–ROI co-expression network.

As such, we can not only identify disease-relevant multi-modal biomarkers, but also reveal a strong association between them. Finally, we compare the performance of multiple DSCCA models guided by different prior knowledge.

For the original DSCCA algorithm, there are two constraints, $P_1$ and $P_2$, which are added for the multi-class discrimination, inspired by the application of locality preserving projection (LPP) in linear discriminant analysis (Ghamisi et al., 2018).

$$P_1(\mathbf{u}) = \|\mathbf{u}\|_D = \mathbf{u}^T\mathbf{X}^T\mathbf{L}_w\mathbf{X}\mathbf{u}$$
$$P_2(\mathbf{v}) = \|\mathbf{v}\|_D = \mathbf{v}^T\mathbf{Y}^T\mathbf{L}_w\mathbf{Y}\mathbf{v} \qquad (9)$$

Here, $L_w$ is the Laplacian graphs of prior knowledge graph.

The final objective function of DSCCA can be written as follows:

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} - \frac{\beta_1}{2}P_1(\mathbf{u}) - \frac{\beta_2}{2}P_2(\mathbf{v})$$
$$s.t.\,\mathbf{u}^T\mathbf{X}^T\mathbf{X}\mathbf{u} = 1, \mathbf{v}^T\mathbf{Y}^T\mathbf{Y}\mathbf{v} = 1, \|\mathbf{u}\|_1 \leq c_1, \|\mathbf{v}\|_1 \leq c_2 \qquad (10)$$

Using Lagrange multipliers, Equation (10) can be reformulated as follows:

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} - \frac{\gamma_1}{2}\|\mathbf{X}\mathbf{u}\|_2^2 - \frac{\gamma_2}{2}\|\mathbf{Y}\mathbf{v}\|_2^2 - \frac{\beta_1}{2}P_1(\mathbf{u}) - \frac{\beta_2}{2}P_2(\mathbf{v})$$
$$- \lambda_1\|\mathbf{u}\|_1 - \lambda_2\|\mathbf{v}\|_1 \qquad (11)$$

Equation (11) is known as a bi-convex problem, which can be solved using an alternating algorithm as discussed in Witten et al. (2009). By fixing $u$ and $v$, respectively, we will have the following two minimization problems shown in Equations (12) and (13).

$$\min_{\mathbf{u}} -\mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} + \frac{\gamma_1}{2}\mathbf{u}^T\mathbf{X}^T\mathbf{X}\mathbf{u} + \frac{\beta_1}{2}P_1(\mathbf{u}) + \lambda_1\|\mathbf{u}\|_1 \qquad (12)$$

$$\min_{\mathbf{v}} -\mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} + \frac{\gamma_2}{2}\mathbf{v}^T\mathbf{Y}^T\mathbf{Y}\mathbf{v} + \frac{\beta_2}{2}P_2(\mathbf{v}) + \lambda_2\|\mathbf{v}\|_1 \qquad (13)$$

We used the Nesterovs accelerated proximal gradient optimization algorithm to solve this objective function following original DSCCA paper (Liu et al., 2012; Yan et al., 2017). The convergence is based on the value changes of the objective function and we use $10^{-6}$ as stop criteria. A five-fold nested cross-validation was applied to automatically tune the parameters $\beta_1$, $\beta_2$, $\lambda_1$, and $\lambda_2$. According to Chen et al. (2012), the learned pattern and performance are insensitive to $\gamma_1$ and $\gamma_2$ settings. Therefore, in this article, we set both of them to 1 for simplicity.

# 4. Results

To test the effect of different prior knowledge on the performance of imaging genetics association, we performed four groups of experiments including the DSCCA algorithm with different prior knowledge (fusion network, diagnosis network, and ROI–ROI network) and the simple SCCA method as the baseline. For SCCA, the parameters were automatically tuned using a permutation method provided in the PMA package. For DSCCA algorithm, we applied a five-fold nested cross-validation to tune the parameters that can also help avoid the overfitting problem. For a fair comparison, the training/test partition was kept exactly the same across methods and ratios of diagnosis groups inside each partition are also identical. All methods went through the same nested cross-validation for parameter tuning.

## 4.1. Imaging genetic associations for AV45

We first tested the association between brain-wide amyloid deposition and top AD-risk SNPs. The performance of test data including the DSCCA algorithm with fused network, diagnosis network, co-expression network as prior knowledge, and the original SCCA method are shown in Table 2. As expected, DSCCA algorithms with the guidance of prior knowledge all outperformed traditional SCCA, confirming that prior knowledge does help reveal stronger brain imaging genetics associations. More specifically, out of all three types of prior, fused network and diagnosis network led to similar association performance, which is much better than the co-expression network.

For all ROIs and SNPs, we averaged their weights across five-folds for feature selection. Figure 2 shows the top 10 ROIs selected by DSCCA guided by the fused network, including left precentral, right parahippocampal, right rostral middle frontal, right precentral, right bankssts, left rostral anterior gingulate, right caudal middle frontal, left postcentral, right postcentral, and left parstriangularis. Among these, right bankssts, right caudal middle frontal, parstriangularis, and left rostral anterior gingulate are part of the default mode network (DMN), and right rostral middle frontal is part of the frontoparietal network. Both of them have consistently shown early accumulation of cortical A$\beta$ fibrils in previous studies (Palmqvist et al., 2017). Amyloid deposition of these regions is strongly associated with only one SNP, *APOE SNP (rs429358)*, which is known as the major risk factor for Alzheimer's disease. There are strong evidences suggesting that *APOE* could inhibit amyloid-$\beta$ (A$\beta$) clearance and promote A$\beta$ aggregation to increase AD risk (Polvikoski et al., 1995; Kim et al., 2009; Kok et al., 2009; Wirths, 2010).

TABLE 2 Test performance of brain imaging genetics association on AV45.

| Prior knowledge | Testing results | | | | | |
|---|---|---|---|---|---|---|
| | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Average |
| **Fused network** | **0.4872** | **0.5844** | **0.5959** | **0.5586** | **0.5043** | **0.5461** |
| Diagnosis network | 0.4863 | 0.5848 | 0.5919 | 0.5655 | 0.4945 | 0.5446 |
| Co-expression network | 0.3486 | 0.4401 | 0.4185 | 0.4203 | 0.37 | 0.3995 |
| SCCA (no prior) | 0.2838 | 0.4041 | 0.4015 | 0.3588 | 0.3865 | 0.367 |

Bold values denote the best performance.

FIGURE 2
Top 10 brain regions with amyloid deposition associated with *APOE SNP rs429358*.
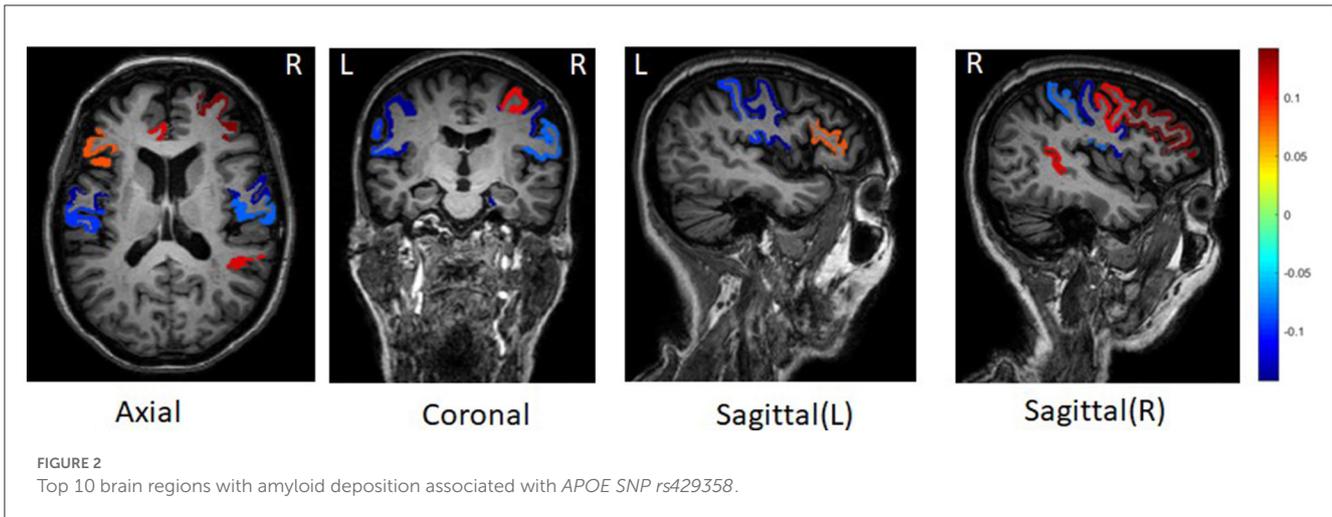
TABLE 3 Tau 5-fold cross-validation results.

| Prior knowledge | Testing results | | | | | |
|---|---|---|---|---|---|---|
| | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Average |
| **Fused network** | **0.4186** | **0.3948** | **0.4882** | **0.1725** | **0.5443** | **0.4025** |
| Diagnosis network | 0.4237 | 0.3786 | 0.4809 | 0.1722 | 0.5458 | 0.4002 |
| Co-expression network | 0.4251 | 0.4246 | 0.4785 | 0.1775 | 0.5424 | 0.4096 |
| SCCA (no prior) | 0.3084 | 0.2674 | 0.321 | 0.1491 | 0.581 | 0.3254 |

Bold values denote the best performance.



FIGURE 3
Top tau brain regions selected by DSCCA guided by fused network.
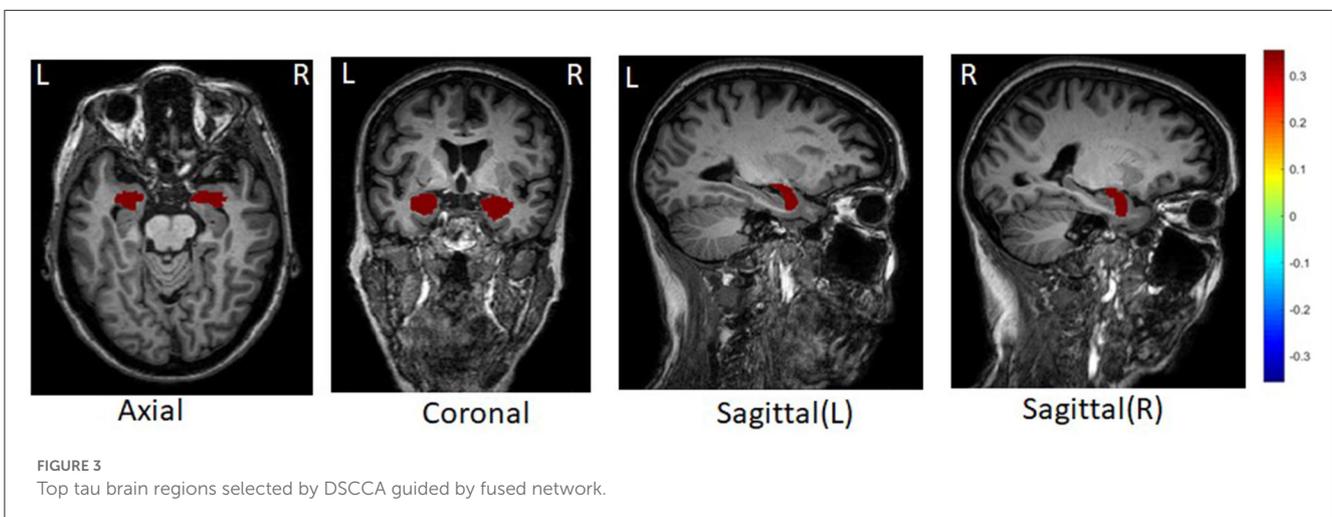
## 4.2. Imaging genetic associations for tau

We next tested the association between brain-wide tau deposition and top AD risk SNPs. The performance on test data across all the methods is shown in Table 3. Similar to amyloid, DSCCA guided by three types of prior knowledge performed significantly better than SCCA. Also, all tree DSCCA models showed very similar performance. Considering that the co-expression network only showed moderate performance in amyloid data, we speculate that the selection of candidate genes has a major effect on the prior co-expression network and later lead to the fluctuation of association performance.

After averaging the weight across 5-folds, we found the tau deposition in the left and right amygdala are strongly associated with 56 SNPs (Figure 3). Amygdala is one of the earliest sites showing tau deposition and neurofibrillary tangles, as reported in previous post-mortem and neuroimaging studies (Vogt et al., 1990; Abiose et al., 2020; Insel et al., 2020). Early tau position in Amygdala is associated with reduced volume and worse cognition performance as well in the preclinical stage of AD (Abiose et al., 2020; Berron et al., 2021). Selected 56 SNPs associating with the amygdala tau deposition are located in or near genes *TOMM40, AC011481.3, NECTIN2, APOC1, AC011481.2, AC015687.1, CLU. TOMM40* poly-T lengths have a significant relationship with

the higher medial temporal plaque and tangle burden in the living brain of non-demented older adults within individuals not carrying the *APOE*-4 allele (Siddarth et al., 2018). We further performed pathway enrichment analysis of these genes using EnrichR (Kuleshov et al., 2016). Top enriched biological process in gene ontology include regulation of receptor-mediated endocytosis (adjusted $p = 0.01$) and cholesterol transport (adjusted $p = 0.01$).

## 5. Discussion

We proposed a new data-driven prior knowledge and tested whether it could help improve the performance of association between brain imaging and genetic features. Our result confirmed the value of all types of prior knowledge in improving association identification. In addition, the fused network representing the subject relationship supported by multi-modal data showed consistently the best or equally best performance compared to the diagnosis network and the co-expression network. By incorporating information from multiple modalities, the fused network more accurately captures the similarity of subjects in disease severity. With the guidance of such a network, it is more likely to reveal the imaging genetic associations related to AD. Co-expression network showed moderate performance in amyloid but was among the top performers in Tau, suggesting the potential effect of gene selection. The construction of the co-expression network relies on a careful selection of gene set that is specific to certain aspects of AD. While our knowledge of the biological mechanism underlying AD is still very limited, there is no optimal strategy to select relevant genes and different selection processes may lead to varying performance in association. Therefore, the value of co-expression network as prior is compromised. Finally, considering fused network and diagnosis network consistently demonstrated similarly good performance, fused network can be valuable prior knowledge to leverage when there is no diagnosis information or in case of studies using a single diagnosis group, e.g., control only or case only.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## Author contributions

BH: conceptualization, methodology, visualization, formal analysis, validation, writing—original draft, and writing—review and editing. LX: investigation, visualization, and formal analysis. PV: conceptualization and visualization. KN: data curation and writing—review and editing. SR: methodology, supervision, and writing—review and editing. AS: data curation, resource, and writing—review and editing. JY: conceptualization, methodology, visualization, writing—original draft, writing—review and editing, supervision, and funding acquisition. All authors contributed to the article and approved the submitted version.

## Conflict of interest

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

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

# References

Abiose, O., Deters, K. D., Young, C., and Mormino, E. C. (2020). Amygdala tau in preclinical Alzheimer's disease: neuroimaging/normal brain aging. *Alzheimers Dement.* 16, e046762. doi: 10.1002/alz.046762

Bekris, L. M., Millard, S., Lutz, F., Li, G., Galasko, D. R., Farlow, M. R., et al. (2012). Tau phosphorylation pathway genes and cerebrospinal fluid tau levels in alzheimer's disease. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 159, 874–883. doi: 10.1002/ajmg.b.32094

Berron, D., Vogel, J. W., Insel, P. S., Pereira, J. B., Xie, L., Wisse, L. E., et al. (2021). Early stages of tau pathology and its associations with functional connectivity, atrophy and memory. *Brain* 144, 2771–2783. doi: 10.1093/brain/awab114

Chen, X., Han, L., and Carbonell, J. (2012). "Structured sparse canonical correlation analysis," in *Artificial Intelligence and Statistics*. PMLR, 199–207.

Chi, E. C., Allen, G. I., Zhou, H., Kohannim, O., Lange, K., Thompson, P. M., et al. (2013). "Imaging genetics via sparse canonical correlation analysis," in *2013 IEEE 10th International Symposium on Biomedical Imaging* (San Francisco, CA: IEEE), 740–743. doi: 10.1109/ISBI.2013.6556581

Du, L., Huang, H., Yan, J., Kim, S., Risacher, S. L., Inlow, M., et al. (2016). Structured sparse canonical correlation analysis for brain imaging genetics: an improved graphnet method. *Bioinformatics* 32, 1544–1551. doi: 10.1093/bioinformatics/btw033

Fischl, B. (2012). Freesurfer. *Neuroimage* 62, 774–781. doi: 10.1016/j.neuroimage.2012.01.021

Fleisher, A. S., Pontecorvo, M. J., Devous, M. D., Lu, M., Arora, A. K., Truocchio, S. P., et al. (2020). Positron emission tomography imaging with [18f] flortaucipir and postmortem assessment of alzheimer disease neuropathologic changes. *JAMA Neurol.* 77, 829–839. doi: 10.1001/jamaneurol.2020.0528

Ghamisi, P., Maggiori, E., Li, S., Souza, R., Tarablaka, Y., Moser, G., et al. (2018). New frontiers in spectral-spatial hyperspectral image classification: the latest advances based on mathematical morphology, markov random fields, segmentation, sparse representation, and deep learning. *IEEE Geosci. Remote Sens. Mag.* 6, 10–43. doi: 10.1109/MGRS.2018.2854840

Insel, P. S., Mormino, E. C., Aisen, P. S., Thompson, W. K., and Donohue, M. C. (2020). Neuroanatomical spread of amyloid $\beta$ and tau in Alzheimer's disease: implications for primary prevention. *Brain Commun.* 2, fcaa007. doi: 10.1093/braincomms/fcaa007

Kim, J., Basak, J. M., and Holtzman, D. M. (2009). The role of apolipoprotein e in Alzheimer's disease. *Neuron* 63, 287–303. doi: 10.1016/j.neuron.2009.06.026

Kok, E., Haikonen, S., Luoto, T., Huhtala, H., Goebeler, S., Haapasalo, H., et al. (2009). Apolipoprotein e-dependent accumulation of alzheimer disease-related lesions begins in middle age. *Ann. Neurol.* 65, 650–657. doi: 10.1002/ana.21696

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90'W97. doi: 10.1093/nar/gkw377

Landau, S., Ward, T. J., Murphy, A., and Jagust, W. (2016). *Flortaucipir (AV-1451) Processing Methods.* Alzheimer's Disease Neuroimaging Initiative. Available online at: https://adni.bitbucket.io/reference/docs/UCBERKELEYAV1451/UCBERKELEY_AV1451_Methods_2021-01-14.pdf

Landau, S. M., Breault, C., Joshi, A. D., Pontecorvo, M., Mathis, C. A., Jagust, W. J., et al. (2013). Amyloid-$\beta$ imaging with Pittsburgh compound b and florbetapir: comparing radiotracers and quantification methods. *J. Nucl. Med.* 54, 70–77. doi: 10.2967/jnumed.112.109009

Lin, D., Cao, H., Calhoun, V. D., and Wang, Y.-P. (2014). Sparse models for correlative and integrative analysis of imaging and genetic data. *J. Neurosci. Methods* 237, 69–78. doi: 10.1016/j.jneumeth.2014.09.001

Liu, J., Ji, S., and Ye, J. (2012). Multi-task feature learning via efficient l2, 1-norm minimization. *arXiv.* [preprint]. 2009, 339–348. doi: 10.48550/arXiv.1205.2631

Okamura, N., and Yanai, K. (2010). Florbetapir (18f), a pet imaging agent that binds to amyloid plaques for the potential detection of Alzheimer's disease. *IDrugs* 13, 890–899.

Palmqvist, S., Schöll, M., Strandberg, O., Mattsson, N., Stomrud, E., Zetterberg, H., et al. (2017). Earliest accumulation of $\beta$-amyloid occurs within the default-mode network and concurrently affects brain connectivity. *Nat. Commun.* 8, 1214. doi: 10.1038/s41467-017-01150-x

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Burlington, MA: Morgan Kaufmann. doi: 10.1016/B978-0-08-051489-5.50008-4

Polvikoski, T., Sulkava, R., Haltia, M., Kainulainen, K., Vuorio, A., Verkkoniemi, A., et al. (1995). Apolipoprotein e, dementia, and cortical deposition of $\beta$-amyloid protein. *N. Engl. J. Medi.* 333, 1242–1248. doi: 10.1056/NEJM199511093331902

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Schellenberg, G. D., and IGAP. (2012). International genomics of alzheimer's disease project (IGAP) genome-wide association study. *Alzheimers Dement.* 8, P101. doi: 10.1016/j.jalz.2012.05.254

Shen, L., Chepelev, I., Liu, J., and Wang, W. (2010). Prediction of quantitative phenotypes based on genetic networks: a case study in yeast sporulation. *BMC Syst. Biol.* 4, 1–10. doi: 10.1186/1752-0509-4-128

Siddarth, P., Burggren, A. C., Merrill, D. A., Ercoli, L. M., Mahmood, Z., Barrio, J. R., et al. (2018). Longer tomm40 poly-t variants associated with higher fddnp-pet medial temporal tau and amyloid binding. *PLoS ONE* 13, e0208358. doi: 10.1371/journal.pone.0208358

Swaminathan, S., Shen, L., Risacher, S. L., Yoder, K. K., West, J. D., Kim, S., et al. (2012). Amyloid pathway-based candidate gene analysis of [11C] pib-pet in the alzheimer's disease neuroimaging initiative (ADNI) cohort. *Brain Imaging Behav.* 6, 1–15. doi: 10.1007/s11682-011-9136-1

Vogt, L. K., Hyman, B., Van Hoesen, G., and Damasio, A. (1990). Pathological alterations in the amygdala in Alzheimer's disease. *Neuroscience* 37, 377–385. doi: 10.1016/0306-4522(90)90408-v

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11, 333–337. doi: 10.1038/nmeth.2810

Wirths, O. (2010). Accumulation of intraneuronal abeta correlates with apoe4 genotype. *Acta Neuropathol.* 119, 555566. doi: 10.1007/s00401-010-0666-1

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 10, 515–534. doi: 10.1093/biostatistics/kxp008

Yan, J., Du, L., Kim, S., Risacher, S. L., Huang, H., Moore, J. H., et al. (2014). Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm. *Bioinformatics* 30, i564–i571. doi: 10.1093/bioinformatics/btu465

Yan, J., Risacher, S. L., Nho, K., Saykin, A. J., Shen, L., and Alzheimer's Disease Neuroimaging Initiative (2017). Identification of discriminative imaging proteomics associations in Alzheimer's disease via a novel sparse correlation model. *Pac. Symp. Biocomput.* 22, 94–104. doi: 10.1142/9789813207813_0010

Zeng, H., Shen, E. H., Hohmann, J. G., Oh, S. W., Bernard, A., Royall, J. J., et al. (2012). Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. *Cell* 149, 483–496. doi: 10.1016/j.cell.2012.02.052