IEEE Journal of Translational Engineering in Health and Medicine

Received 17 April 2024; revised 16 August 2024 and 13 September 2024; accepted 14 September 2024.
Date of publication 19 September 2024; date of current version 14 October 2024.

Digital Object Identifier 10.1109/JTEHM.2024.3463720

Integrating Multimodal Neuroimaging and Genetics: A Structurally-Linked Sparse Canonical Correlation Analysis Approach

JIWON CHUNG[1], (Graduate Student Member, IEEE), SUNGHUN KIM[1], JI HYE WON[2], AND HYUNJIN PARK[1,3]

[1]Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon 16419, Republic of Korea
[2]Department of Computer Engineering and Artificial Intelligence, Pukyong National University, Busan 48513, Republic of Korea
[3]Center for Neuroscience Imaging Research, Institute for Basic Science, Suwon 16419, Republic of Korea

CORRESPONDING AUTHORS: H. PARK (hyunjinp@skku.edu) AND J. H. WON (jhwon@pknu.ac.kr)

This work was supported in part by the National Research Foundation under Grant NRF-2020M3E5D2A01084892, in part by the Institute for Basic Science under Grant IBS-R015-D1, in part by AI Graduate School Support Program (Sungkyunkwan University) under Grant RS-2019-II190421, in part by ICT Creative Consilience Program under Grant RS-2020-II201821, in part by the Artificial Intelligence Innovation Hub Program under Grant RS-2021-II212068, and in part by the Pukyong National University Research fund in 2023 under Grant 202303840001.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board (IRB) of Sungkyunkwan University.

This article has supplementary downloadable material available at https://doi.org/10.1109/JTEHM.2024.3463720, provided by the authors.

ABSTRACT Neuroimaging genetics represents a multivariate approach aimed at elucidating the intricate relationships between high-dimensional genetic variations and neuroimaging data. Predominantly, existing methodologies revolve around Sparse Canonical Correlation Analysis (SCCA), a framework we expand to 1) encompass multiple imaging modalities and 2) promote the simultaneous identification of structurally linked features across imaging modalities. The structurally linked brain regions were assessed using diffusion tensor imaging, which quantifies the presence of neuronal fibers, thereby grounding our approach in biologically well-founded prior knowledge within the SCCA model. In our proposed structurally linked SCCA framework, we leverage T1-weighted MRI and functional MRI (fMRI) time series data to delineate both the structural and functional characteristics of the brain. Genetic variations, specifically single nucleotide polymorphisms (SNPs), are also incorporated as a genetic modality. Validation of our methodology was conducted using a simulated dataset and large-scale normative data from the Human Connectome Project (HCP). Our approach demonstrated superior performance compared to existing methods on simulated data and revealed interpretable gene-imaging associations in the real dataset. Thus, our methodology lays the groundwork for elucidating the genetic underpinnings of brain structure and function, thereby providing novel insights into the field of neuroscience. Our code is available at https://github.com/mungegg.

INDEX TERMS fMRI, human connectome project, neuroimaging genetics, sparse canonical correlation, T1 MRI.

Clinical and Translational Impact Statement— This study enhances the integration of genetic and neuroimaging data using an advanced multimodal model improving our understanding of how genetic variations influence brain structure and function. Thus, it could serve as the first step towards identifying genetic markers that correlate with neuroimaging patterns, aiding in the prediction, diagnosis, and treatment of neurological disorders.

I. INTRODUCTION

NEUROIMAGING genetics is an emerging field that aims to uncover intricate associations between genetic variants and neuroimaging data [1], [2]. While a conventional genome-wide association study (GWAS) directly links fine-grained genetic information with coarse-grained patient

information (e.g., diagnosis), neuroimaging genetics is more sensitive because it integrates rich imaging information. Imaging observations serve as an intermediate bridge between fine-grained genetic information and coarse-grained patient information. Through identifying associations between genetic factors and imaging measurements, neuroimaging genetics seeks to model and understand how genetic factors influence the structure and function of the human brain.

In previous studies, researchers have utilized various methodologies to ascertain gene-imaging associations, encompassing both pairwise univariate and multivariate regression approaches [3], [4], [5], [6]. Pairwise univariate methodologies, such as General Linear Models (GLM) and Reduced Rank Regression, are frequently employed to evaluate the correspondence between single-nucleotide polymorphisms (SNPs) and imaging data. However, these approaches are unable to account for multivariate interactions or intricate relationships among multiple variables. Conversely, multivariate regression techniques such as Principal Component Analysis (PCA), Independent Component Analysis (ICA), and clustering have been utilized to capture correlations between multiple SNPs and imaging characteristics. Nevertheless, the interpretability of these methods can be challenging, thereby restricting the practical implications of the findings. These methods are unable to effectively manage the complexity of multimodal datasets because PCA and ICA are fundamentally linear techniques. Despite yielding valuable insights, these methodologies have encountered difficulties in effectively managing the complexity of multimodal datasets. Multimodal datasets, such as those containing neuroimaging and genetic data, often involve heterogeneous data types and complex, non-linear relationships that PCA and ICA cannot adequately model.

Recently, neuroimaging genetics has increasingly adopted bivariate methods, with a notable emphasis on Sparse Canonical Correlation Analysis (SCCA) models [7]. SCCA is typically formulated to discern regularized multivariate associations between two distinct datasets: genetic markers and neuroimaging observations. The evolution to multi-view SCCA has extended the capability to accommodate additional modalities, including three or more [7]. Three-way SCCA (TSCCA) emerges as a specific instance of multi-view SCCA, offering a framework to investigate multivariate relationships between SNPs, imaging data, and clinical scores [8]. Gradient Kernel CCA (Grad KCCA) represents a CCA method operating within a transformed feature space via a kernel function. This approach enables the capture of nonlinear relationships, which traditional linear methods may overlook [9]. CCA-based methodologies have significantly enhanced the capacity to unveil intricate genetic-imaging associations. Nonetheless, some of the associations identified were noted to signify spurious relationships because they mainly reflected statistical correlation and did not agree with biological prior knowledge [7], [8]. For instance, Elliott et al. highlighted the risk of spurious correlations in large-scale neuroimaging-genetics studies due to the high dimensionality of the data, which can lead to findings that lack biological plausibility [9]. Similarly, another study by Grasby et al. pointed out that many of the associations identified in CCA-based analyses did not replicate in independent datasets, suggesting that these findings may be artifacts of the data rather than true biological relationships [10].

II. RELATED WORK

A. SPARSE CANONICAL CORRELATION ANALYSIS (SCCA)

Let $X \in R^{n\times p}$ consist of n patients and p genetic features, and $Y \in R^{n\times q}$ consist of n patients and q imaging features. Uppercase notation indicates a matrix, and lower-case notation indicates a vector. SCCA aims to determine a linear combination that maximizes the correlation between two datasets while controlling the sparsity of the model using the L1 penalty.

$$\begin{aligned} &\min_{u,v} -u^T X^T Y v \\ &s.t.\ \|u\|_2 \le 1, \|v\|_2 \le 1, \|u\|_1 \le c1, \|v\|_1 \le c2 \end{aligned} \tag{1}$$

Here, u and v represent the corresponding canonical loading vectors and the L1 penalty introduces sparsity to the vector, facilitating the interpretation of canonical variables in high-dimensional data [7].

B. MULTI-VIEW SCCA

Bimultivariate techniques have been proposed to detect complex associations between genetic features and multimodal neuroimaging datasets. Multi-view SCCA (mSCCA) methods have been developed to handle multimodal imaging data. The mSCCA method is a simple extension of conventional SCCA, which cannot model structurally linked features across modalities [7], [12].

$$\begin{aligned} &\min_{u_k v_k} \sum_{k=1}^{K} \|Xu_k - Yv_k\|_2^2 \\ &\text{s.t. } \|Xu_k\|_2^2 = 1, \|Yv_k\|_2^2 = 1, \end{aligned} \tag{2}$$

where K is the number of imaging modalities.

III. METHOD AND PROCEDURE

A. STRUCTURALLY-LINKED SCCA (S^2CCA)

Using the SCCA method may yield the identification of features that, despite lacking direct structural connections, exhibit a substantial correlation. In this study, we present an algorithm grounded in SCCA, integrating structural constraints to simultaneously identify linked features. We ensured the concurrent extraction of features known to exhibit high correlation reflecting their underlying structural connectivity. Figure 1 provides an overview of the proposed method.

Dataset X represents the genetic data consisting of n patients and p SNP features. Y_1 and Y_2 represent two different imaging modalities with q neuroimaging features. The SCCA

IEEE Journal of Translational Engineering in Health and Medicine

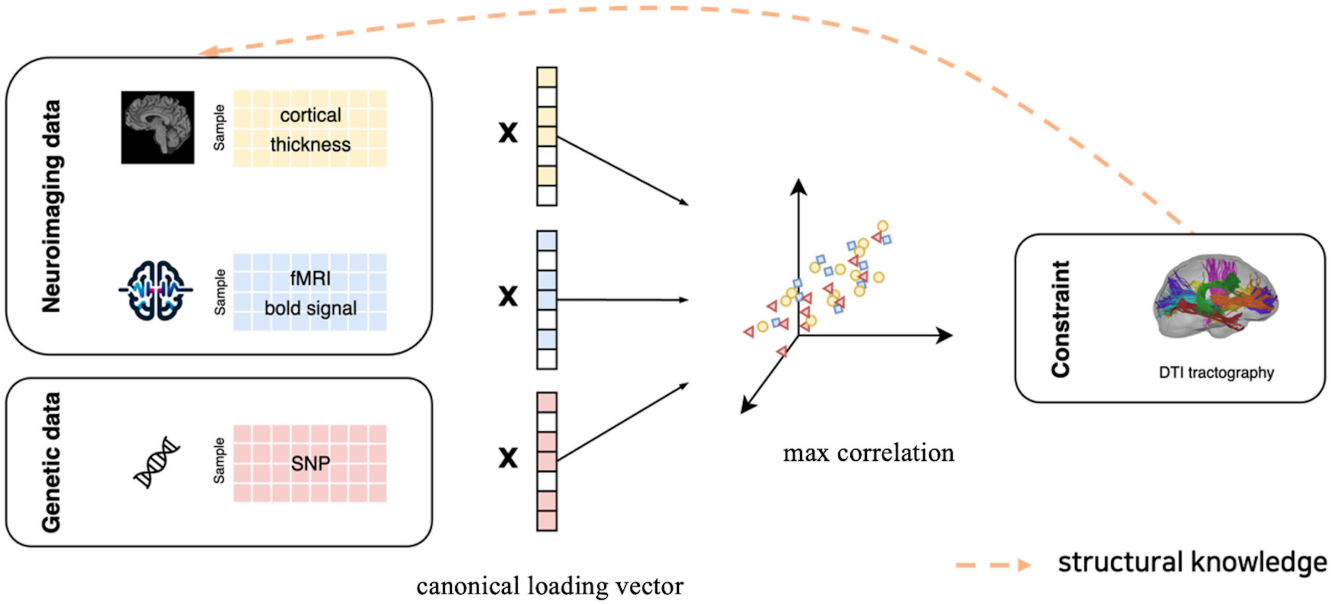

FIGURE 1. Proposed Methodology for Integrating Multimodal Data. Our methodology is specifically designed for analyzing multimodal data encompassing neuroimaging and genetic information. We utilize neuroimaging data obtained from T1-weighted MRI in the form of cortical thickness and functional MRI scans measuring the amplitude of low-frequency fluctuation, to provide a comprehensive understanding of both structural and functional aspects of the brain. Additionally, we employ DTI tractography to facilitate the simultaneous identification of structurally linked features across two MRI modalities.

formulations for the three modalities are as follows:

$$\begin{aligned}\min_{u,v} &-\frac{1}{n}u^{T}X^{T}Y_1v_1-\frac{1}{n}u^{T}X^{T}Y_2v_2-\frac{1}{n}v_1^{T}Y_1^{T}Y_2v_2,\\ \text{s.t. }&\|u\|_2^2\le 1,\ \|v_1\|_2^2\le 1,\ \|v_2\|_2^2\le 1,\ \|u\|_1\le c_1,\\ &\|v_1\|_1,\ \|v_2\|_1\le c_2,\ L(u)\le c_3,\\ &P(v_1),P(v_2)\le c_4,\ \|v_2-v_1\|\le c_5\end{aligned}\tag{3}$$

In our investigation, we utilized T1-weighted MRI and functional MRI (fMRI) time-series data to delineate the structural and functional attributes of the brain [13], [14]. These modalities can be linked by assessing the connectivity of white matter tracts between brain regions using DTI. To facilitate the simultaneous identification of structurally connected features across modalities, we expanded Equation (1) to the following objective function:

$$\begin{aligned}\min_{u,v} &-\frac{1}{n}u^{T}X^{T}Y_1v_1-\frac{1}{n}u^{T}X^{T}Y_2v_2-\frac{1}{n}v_1^{T}Y_1^{T}Y_2v_2\\ &+\beta_1\|u\|_1+\beta_2\|v_1\|_1+\beta_2\|v_2\|_1+\lambda_1u^{T}Lu\\ &+\lambda_2v_1^{T}P^{T}v_1+\lambda_2v_2^{T}Pv_2+\tau\|v_2-v_1\|\end{aligned}\tag{4}$$

Here, u and $V=[V_1, V_2]$ denote the corresponding canonical vectors of X, Y_1 and Y_2, respectively and β_1, β_2, λ_1, λ_2, τ are regularization parameters. β_1 and β_2 correspond to L1 regularization and affect the sparsity of the genetic and imaging canonical vectors. L encourages connected SNPs to be identified together in the form of a graph Laplacian matrix and λ_1 acts as a connectivity-based penalty. P is a newly proposed constraint applicable across neuroimaging data encouraging structurally connected features to be identified using probability value and λ_2 is the associated penalty. τ is the fused least absolute shrinkage and selection operator (LASSO) penalty that encourages canonical vectors from different modalities to share similar weights between imaging modalities.

B. OPTIMIZATION

Equation (3) represents the objective function aimed at minimizing u and V. Due to its non-convex nature, direct optimization within our algorithm presents a challenge. Therefore, we opted for an alternative convex search method for optimization [15]. Fixing one of the variables, either u or V, renders the corresponding objective function convex, allowing for sequential variable fixation for minimization. Initially, both u and V are initialized. Subsequently, at each iteration, a block of V is fixed while optimizing a block of u, and vice versa. This iterative process continues until convergence is attained.

C. TUNING OF HYPERPARAMETERS

Our model relies on several hyperparameters to operate and the correct combination of these parameters strongly determines the performance of the model. In this study, we focused on five hyperparameters (i.e., β_1, β_2, λ_1, λ_2, τ) for tuning. We systematically analyzed the performance variation of each hyperparameter of the model using the cross-validation method as follows.

$$\begin{aligned}\mathrm{CV}=\frac{1}{5}\sum_{i=1}^{5}\frac{1}{3}\{&\mathrm{corr}\left(X_iu_{-i},Y_{1_i}v_{1_{-i}}\right)\\ &+\mathbf{corr}\left(X_iu_{-i},Y_{2_i}v_{2_{-i}}\right)\\ &+\mathbf{corr}\left(Y_{1_i}v_{1_{-i}},Y_{2_i}v_{2_{-i}}\right)\}\end{aligned}\tag{5}$$

In particular, a grid search method was adopted to systematically explore the hyperparameters. β_1 and β_2 were tuned in ranges [5, 2500] and [5, 200] respectively in 5 increments. λ_1, λ_2 were tuned in ranges [0.1, 0.5]. in 0.05 increments. τ was tuned in ranges [0.05, 0.2] in 0.05 increments.

D. DATA PREPROCESSING

Brain imaging data were obtained from the Human Connectome Project (HCP) [16]. We obtained quality-controlled data from the HCP website [17]. This study was a retrospective analysis and institutional review board (IRB) approval was obtained from Sungkyunkwan University. Out of the 1206 subjects in the HCP S1200 dataset, we filtered for those with available genetic data and multimodal imaging, identified as White, and whose ethnicity was not Hispanic/Latino. The last filtering criterion was to reduce the likelihood of ethnic stratification effects in the genetic analysis. This filtering process resulted in 525 subjects. We used parts of the HCP data, where brain structural MRI scans were conducted using Siemens scanner, specifically 3 T Trio or 3T Prisma models. In this study, we utilized the HCP database, which includes longitudinal data. However, for our analysis, we only used the baseline data (i.e., the first visit) for each participant.

The structural images were captured using a T1-weighted 3D magnetization-prepared rapid acquisition gradient echo sequence. Specific parameters were as follows: voxel size $= 1.0 \times 1.0 \times 1.0$ mm^3, FoV $= 256 \times 256$ mm^2, matrix size $= 256 \times 256$, TR $=$ 2300ms, TE $=$ 2.98ms, and flip angle $= 9°$. Every T1-weighted image was first evaluated for artifacts or excessive motion. These imaging data were corrected for gradient nonlinearity and b0 distortions, followed by co-registration using a rigid-body transformation. The processed data were nonlinearly registered to the MNI152 standard space. The white and pial surfaces were generated by following the boundaries between the different tissues [18], [19], [20]. These surfaces were averaged to generate the mid-thickness surface, which was used to generate the inflated surface. The generated spherical surface was registered to the Conte69 template with 164k vertices using MSMAll [21], [22].

Cortical thickness values were derived using the HCP-MMP1.0 (HCP Multi-Modal Parcellation version 1.0) parcellation. This atlas, created through a combination of multimodal brain images and an objective semi-automated neuroanatomical approach divides the human brain into a total of 360 regions (180 in each hemisphere) [23].

The fMRI data were obtained using a multiband gradient-echo EPI imaging sequence with TR $=$ 1,000 ms, TE $=$ 22.2 ms, flip angle $= 45°$, FoV $= 208 \times 208$, matrix $=$ 130×130, 85 slices, voxel size $= 1.6 \times 1.6 \times 1.6$ mm^3, and a multiband factor of 5. The entire scanning duration time for the fMRI protocol was approximately 16 minutes, producing 900 volumes. The preprocessing for the fMRI was performed by the HCP based on the updated data pipeline (v3.21.0, https://github.com/WashingtonUniversity/HCPpipelines). The data were registered onto the T1w structural data and then onto the MNI 152 standard space. Magnetic field bias correction, non-brain tissue removal, and intensity normalization were performed. They were also corrected for gradient distortions and head motion. Noise components attributed to head movement, white matter, cardiac pulsation, arterial, and large vein-related contributions were removed using FIX [24], [25]. We calculated the average framewise displacement (FD) for the fMRI data to quantify head motion. FD is defined as the sum of the absolute values of the differentiated realignment estimates (translations and rotations) across all time points. The average FD was computed for each participant and the mean and standard deviation of FD across participants were reported. The mean and standard deviation of FD were 0.13mm and 0.0038mm for our data. The fMRI data from the HCP 1200 release were used. The preprocessed rs-fMRI data were mapped to standard grayordinate surface space with a cortical ribbon-constrained volume-to-surface mapping algorithm. We used the averaged and cleaned time series of all the grayordinates data for each region of the HCP-MMP 1.0 atlas [23].

DTI data were also from HCP and obtained using a spin-echo EPI sequence with 1.05 mm isotropic voxels, TR $=$ 7000ms, TE $=$ 71ms, 65 unique diffusion gradient directions, and 6 b0 images obtained for each phase encoding direction pair. Preprocessing included B0 intensity normalization, eddy current-induced field inhomogeneity correction, and head motion correction in the volume space [26]. We further conducted diffusion tractography of DTI data using methods detailed in a previous study to obtain neuronal fiber information connection brain regions [23]. The B0 portion of the DTI was volumetrically registered to the volume version of the MMP atlas [22]. The motion-corrected fMRI data were mapped to the standard grayordinate surface space where the MMP atlas resides as described in the fMRI preprocessing procedure [28]. Combining the two procedures led to a spatial alignment of DTI data with fMRI data.

We computed the Amplitude of Low-Frequency Fluctuations (ALFF) feature to quantify spontaneous brain activity from preprocessed fMRI data [29]. After the standard preprocessing of fMRI data, Fourier transform was applied followed by band-pass filtering (0.01 – 0.1 Hz) and power spectrum computation for each region [30]. The values were further normalized by dividing by the global mean ALFF of each subject. Thus, a total of 360 ALFF values were computed for the fMRI data.

The Yeo 7 network is a widely used parcellation of the human brain that categorizes it into seven distinct functional networks based on resting-state functional MRI data [31]. These networks include the Visual, Somatomotor, Dorsal Attention, Ventral Attention, Limbic, Frontoparietal, and Default Mode networks. Each network in the Yeo 7 parcellation is linked to specific cognitive and sensory functions, creating a detailed map of brain organization.

IEEE Journal of Translational Engineering in Health and Medicine

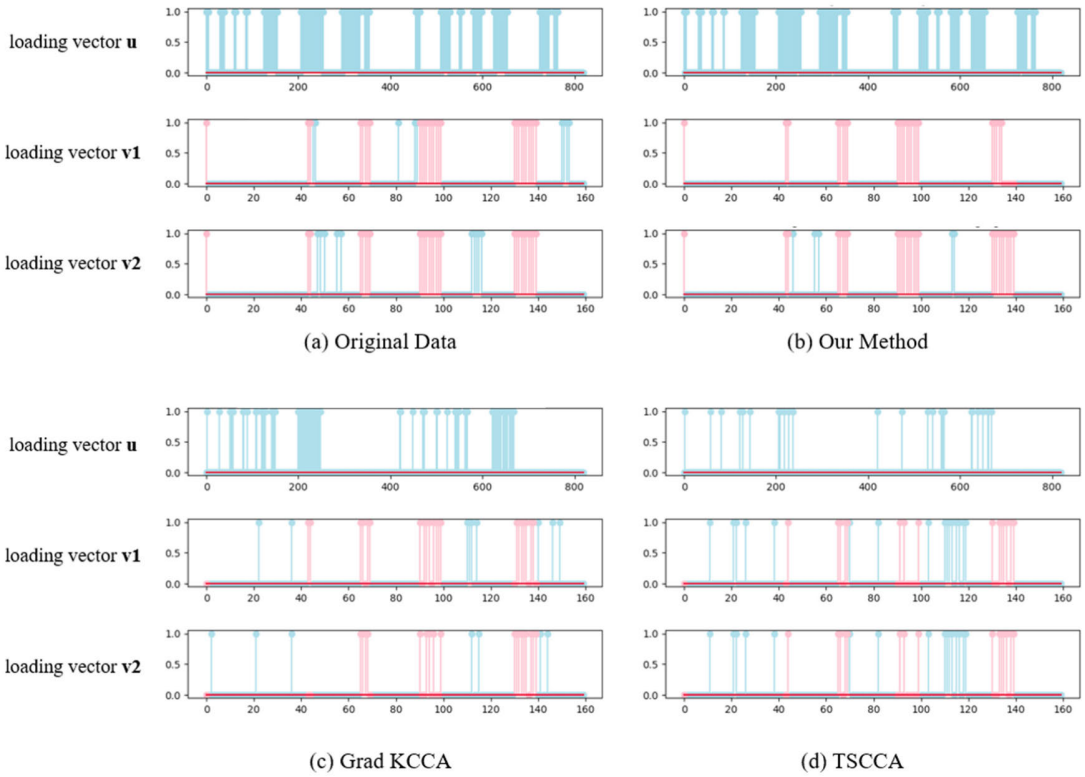

FIGURE 2. Comparison of loading vectors for various methods at the noise level of 1.5. (a) ground truth (b) proposed method (c) Grad KCCA (d) TSCCA Non-zero elements (both pink and blue) denote significant elements present. The pink elements are structurally linked components occurring across imaging modalities.

The genotyping data released by HCP was obtained from the dbGAP website, specifically under the designation phs001364.v1.p1. These data were collected using the Illumina Multi-Ethnic Global Array (MEGA) SNP array. Through this technique, genotypic information of 1,580,642 SNPs was obtained across all participating subjects. SNPs that exhibited a minor allele frequency of less than 1% were discarded. Additionally, any SNP that did not conform to the Hardy–Weinberg equilibrium criteria of less than 10^{-6} or had a genotype missing rate exceeding 5% was excluded. This stringent criterion ensures the reliability and validity of the genotyping data. Once the genetic data were curated and prepared, they were further analyzed to investigate any potential associations between SNPs and measures of association with imaging features. For this task, a GWAS was conducted using the widely-recognized PLINK software [27]. The methodology chosen for this analysis was a linear regression model, which effectively adjusts for potential confounding variables, namely sex and age. After the mentioned analysis, we obtained 4,259 candidate SNPs.

IV. EXPERIMENT AND RESULT

A. SIMULATION STUDY

To evaluate the performance of our method, we performed a simulation experiment. First, our simulation experiment started with a data generation step. We generated the genetic dataset $A \in \mathbb{R}^{n\times p}$ and imaging datasets $B_1, B_2 \in \mathbb{R}^{n\times q}$ where ground truth is known. The data size was set to n = 100, p = 820, and q = 160, where n is the number of patients, p is the number of SNPs, and q is the number of imaging features. The number of SNP and imaging features were chosen to be similar to existing studies [32]. For the genetic dataset $A = a\ell + e$, let ℓ be a latent variable that was randomly sampled from a normal distribution, e be the noise that was added to the ground truth value of a latent variable, and a be the binary indicator variable to denote the significant element. For neuroimaging data B1 and B2, the same method in genetic data generation was used.

We experimented with different noise levels (from 0 to 5) to ensure diversity in the experiment. We assumed structurally linked features across the neuroimage datasets B1 and B2 with matrix P where non-zero elements denote linked features. For example, if P_{ij} is non-zero, it implies that i-th element of the first modality is linked with the j-th element of the second modality. We also simulated the graph Laplacian matrix L for the genetic data as $L = D - C$, where D represents the degree matrix of the connectivity matrix C [32].

B. REAL DATA

We evaluated the performance of our approach on real data of healthy individuals from the HCP database. The regional fMRI time series data were averaged in the

TABLE 1. Comparison of detection performances among methods.

Model	Loading vector	AUC		
		Noise e =1.5	Noise e =3	Noise e = 5
TSCCA	u	0.52	0.58	0.67
	v0	0.90	0.78	0.68
	v1	0.89	0.66	0.54
Grad KCCA	u	0.55	0.46	0.51
	v0	0.62	0.44	0.52
	v1	0.70	0.62	0.57
Ours	u	**0.92**	**0.92**	**0.80**
	v0	**0.99**	**0.98**	**0.91**
	v1	**0.99**	**0.95**	**0.89**

AUC; area under the curve. Bold denotes the best performance among comparisons.

temporal dimensional yielding 360 regional values similar to the Amplitude of Low-Frequency Fluctuation (ALFF) approach [33]. We applied our S^{2}CCA to temporally averaged fMRI times series feature (from fMRI) and cortical thickness features (from T1-weighted images), which were constrained by the probabilistic DTI tractography so that we can explore the association between genes and brain imaging data to extract not only significant features but also structurally connected features in fMRI and T1-weighted images. In our experiments, we examined the associations between 360 brain regions (fMRI and cortical thickness features) and 4,259 SNPs extracted from the preprocessing steps as described before to identify highly correlated and simultaneously structurally connected regions. For this purpose, 10, 200, 0.01, 0.3, and 0.5 were chosen as the five hyperparameters of the algorithm, we obtained an average canonical correlation value of 0.1771 across cross-validations.

Table 2 shows the top 5% (18 regions out of 360 regions) features selected by applying S^{2}CCA in each imaging modality. There are 6,480 (i.e., top 1% of 129,600 possible connections) strong connections in the tractography matrix and the identified imaging features are all subsets of the strong connections. This shows the benefits of constraining the algorithm with tractography.

Figure 3 (a) shows a comparison of selected features in Table 2 based on the Yeo 7-network parcellation [31]. Regional weights were summed to generate the network-level maps. The results of both modalities are dominated by the frontoparietal and the default mode networks. Previous study indicates that the default mode and frontoparietal networks are heavily connected playing a central role in integrating various cognitive and functional processes [34]. Unlike approaches that focus solely on functional connectivity, our S^{2}CCA approach allowed us to integrate both structural connectivity (via DTI) and genetic influences (via SNP correlations). These constraints likely accentuated the prominence of the frontoparietal and default mode networks, as these networks not only demonstrate robust neuronal fiber connections but also exhibit significant heritability in their connectivity patterns. The existing study supports the notion that genetic factors can exert a substantial influence on regional gray matter, particularly within the prefrontal and parietal cortices,

TABLE 2. The top 5% selected neuroimaging features with non-zero canonical weights.

Cortical thickness	Weight	fMRI ALFF	Weight
Right Primary Visual Cortex	0.0429	Right Superior Temporal Visual Area	0.1172
Right Area 8C	0.0422	Right Area 8C	0.2073
Right Area IFSp	0.0436	Right Area IFSp	0.2128
Right Area IFSa	0.0421	Right Area posterior 9-46v	0.1605
Right Area posterior 9-46v	0.0439	Right Area 46	0.2225
Right Area 46	0.0433	Right RetroInsular Cortex	0.1058
Right Auditory 5 Complex	0.0426	Right Auditory 5 Complex	0.1206
Right STSd posterior	0.0435	Right STSd posterior	0.2811
Right Area TE1 posterior	0.0451	Right STSv posterior	0.1064
Right Area PHT	0.0434	Right Area TE1 posterior	0.1419
Right Area TemporoParietoOccipital Junction 1	0.0450	Right Area TemporoParietoOccipital Junction 1	0.3794
Right Area PGp	0.0441	Right Area TemporoParietoOccipital Junction 2	0.131
Right Area IntraParietal 1	0.0444	Right Area PGp	0.1565
Right Area IntraParietal 0	0.0434	Right Area IntraParietal 1	0.0976
Right Area PFm Complex	0.0468	Right Area PFm Complex	0.2863
Right Area PGi	0.0448	Right Area PGi	0.3225
Right Area PGs	0.0470	Right Area PGs	0.3231
Right Area TE1 Middle	0.0436	Right Area TE1 Middle	0.1596

Names of the regions follow this study [15].

areas that are central to the frontoparietal network [35], [36]. These genetic influences, in turn, could underlie the strong connectivity observed within the default mode network, especially during resting-state conditions [37]. This genetic mediation may explain why networks such as the dorsal attention or ventral attention networks, despite their known functional connectivity with the default mode network, were not highlighted in our results. Figure 3 (b) shows the canonical weight from the result of S^{2}CCA confirming that the frontoparietal network and default mode regions have high canonical weights. Figure 3(c) shows the summed weights for the 22 parcellated cortical regions from HCP-MMP 1.0 atlas [22], a finer level than the Yeo-7 network.

We found 24 gene features with non-zero weights from our S^{2}CCA method. The identified SNPs were mapped to 18 genes and they were mostly in the SORC2, SRGAP1,

IEEE Journal of Translational Engineering in Health and Medicine

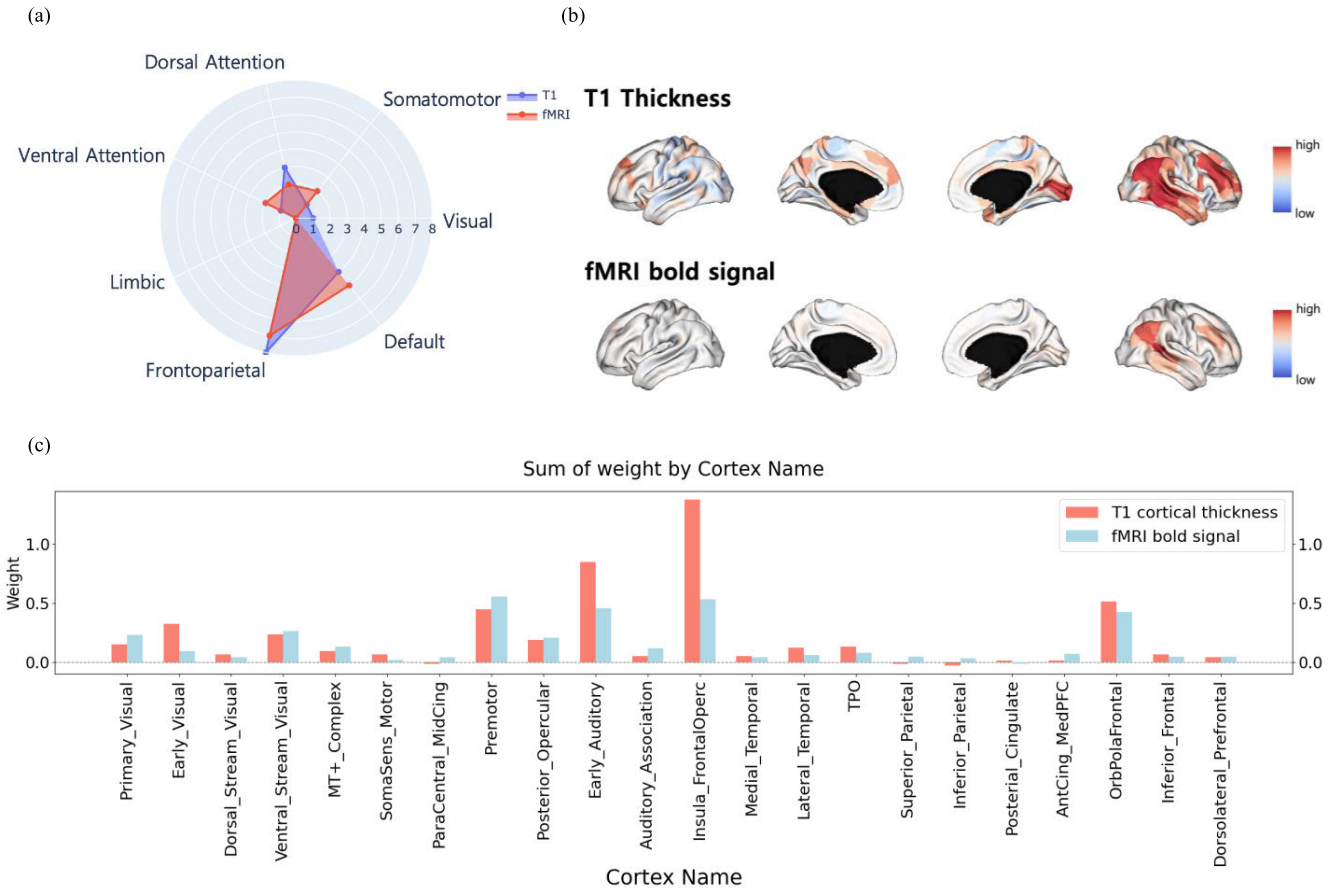

FIGURE 3. **(a) Comparison of selected features for each imaging modality at the network level. (b) Canonical weights visualized on the brain from the result of S^{2}CCA. (c) Canonical weights according to 22 parcellated cortical regions.**

SYT1, and WDR21A genes (Supplement Table 1). Further discussions of these SNPs are given in the Discussion and Conclusion section.

V. DISCUSSION AND CONCLUSION

The SCCA method is a multivariate technique used in neuroimaging genetics for identifying features simultaneously linked with imaging and genetic data [7]. To enhance the selection of relevant features, several regularization techniques have been introduced [7], [8], [11]. We propose an algorithm that leverages the SCCA, referred to as the S^{2}CCA, incorporating prior knowledge of structurally linked regions. This approach guarantees the simultaneous extraction of features that are recognized to have high correlation and connectivity. We argue that our S^{2}CCA framework with structural connectivity constraint identifies features that are consistent with biological prior knowledge and thus improves the existing SCCA-based methods. Results of the simulation data indicated that our algorithm is highly effective at modeling and detecting intricate gene-imaging relationships, even under conditions of arbitrary data designed to mimic real challenges. When applied to real data of SNP, T1-MRI, and fMRI, our method revealed SNPs that are significantly correlated with key neuroimaging markers, thus demonstrating the practical applicability of the approach in uncovering genetic influences on both the structural and functional aspects of the human brain.

Our study distinguishes itself by not only integrating structural T1-weighted and fMRI measurements with genetic markers within a normal population but also by harnessing the capabilities of DTI tractography. This addition of tractography has allowed us to assign weights to the physical connections between brain structures, thereby enriching our analysis with the dimension of actual neural connectivity offering insights into how structural and functional brain networks are modulated by genetic factors. This integration is particularly crucial for unraveling the complex interactions that underlie neurodevelopmental and neurodegenerative processes, potentially opening new avenues for the diagnosis and treatment of neurological and psychiatric conditions. While our findings offer a broad overview rather than disease-specific insights, they notably expand the understanding of potential neurobiological variations, setting a foundation for further exploration of genetic and neuroimaging correlations.

We chose the cortical thickness of T1 MRI and ALFF of fMRI as two neuroimaging measures. Cortical thickness is a well-established marker of brain morphology [38]. Variations in cortical thickness are associated with a variety of cognitive

functions, developmental stages, and neurological conditions. Thus, we can identify structural traits in the brain that may have genetic underpinnings, providing clues about how genes influence brain development, aging, and susceptibility to diseases [39]. ALFF quantifies low-frequency oscillations in fMRI signal, representing spontaneous brain activity. ALFF analysis enables the identification of functional abnormalities and variations in brain activity that could be genetically mediated. This helps in understanding how genetic factors may influence brain function and connectivity patterns, offering insights into the genetic basis of cognitive functions and neuropsychiatric disorders [40]. Together, cortical thickness and ALFF are complementary to one another and provide a comprehensive view of the brain's structural and functional status.

We identified 22 unique regions mostly concentrated in frontoparietal network and default mode network from two neuroimaging modalities that are closely related. For the genetic features, we identified 18 genes annotated from 4,259 SNPs. A few of the resulting genes, ARHGAP26, SORCS2, and SRGAP1, have been reported to be influential in the structural and functional integrity of the brain. For instance, ARHGAP26 is known for its role in cell signaling and cytoskeletal organization, which could be pivotal in the development and maintenance of neural pathways as evidenced by DTI tractography [41]. Similarly, SORCS2, with its involvement in neuronal growth and guidance, may play a crucial role in shaping the connectivity patterns we observed among the highlighted brain regions [42]. In addition, our findings suggest a potential genetic basis for the connectivity between areas such as the Primary Visual Cortex and the Superior Temporal Visual Area, areas crucial for visual processing and integration. The genetic markers identified in our study, including those associated with synaptic transmission (e.g., SYT1), might explain the functional synchrony and structural coherence observed in these regions, offering a genetic lens through which to understand the complex web of neural connections [42]. Furthermore, the observed associations between genetic markers and regions involved in auditory processing, executive functions, and spatial navigation (e.g., Area 46, Intraparietal Areas) emphasize the genetic contributions to cognitive and sensory processing capacities. This underscores the importance of considering genetic data alongside neuroimaging to fully grasp the neurobiological substrates of brain function [42].

Neurosynth [43] meta-analysis revealed that the ARHGAP 26 and SYT1 genes show strong expression in the frontoparietal and default mode networks (e.g., Area 8C, Area posterior 9-46v, Area 46, Area PGs, Area PGi, Temporo-Parieto-Occipital Junction1, STSd posterior, IntraParietal 0, Area TE1 posterior, Auditory 5 Complex, Area PFm Complex). Additionally, ARHGAP26, EPC2, and CCDC147 genes are associated with regions within the visual, dorsal attention, and ventral attention networks (e.g., Primary Visual Cortex, Superior Temporal Visual Area, Intraparietal Area, and Temporo-Parieto-Occipital Junction 2) [43]. Furthermore, the C20orf26, RSRC1, SRGAP1, CBX5, and CD82 genes show high expression in the thalamus, a central brain region pivotal for sensory information transmission and processing. This indicates a vital connection to the functionality of this region, underscoring the potential importance of these genes in maintaining normal brain function. However, we could not find existing reports on the remaining regions. Additional gene ontology analysis on the 18 genes revealed potentially different roles (Supplement Fig.1). Thus, all the identified genetic features could have been reported before. In sum, the identified gene – brain region associations emphasize the close relationship between genes and the brain functionally. These findings suggest the significant roles these identified genes may play in cognitive and attentional functions, further emphasizing the intricate relationship between genetic expressions and brain functionality. Still, the identified regions and genetic features need further validation studies to fully explore their role.

Our study has several limitations. This is a single-center study and thus our findings should be interpreted with care. Future studies are needed to see if our findings are replicable across other normal cohorts. We adopted ALFF as the fMRI measurement. There are other measures such as functional connectivity. Integrating other fMRI-related measurements might offer novel insights into the brain and this is left for future work. We considered only White and non-Hispanic subjects in our study and thus our findings might not generalize to other ethnic groups.

Our study shares the idea of leveraging multiple data modalities to decipher the intricacies of brain function and genetics, akin to disease-focused research, but it distinguishes itself by its broad applicability and potential to inform on the vast heterogeneity observed in the general population. Future research should aim to build upon these initial findings, incorporating longitudinal studies to further elucidate the developmental and aging processes as influenced by genetic and neuroanatomical variations. This study exemplifies the importance of integrating diverse scientific methods to enhance our understanding of the human brain, potentially leading to personalized medical interventions that consider the unique genetic and neurobiological makeup of individuals.

REFERENCES

[1] A. R. Hariri and D. R. Weinberger, "Imaging genomics," *Brit. Med. Bull.*, vol. 65, pp. 259–270, Mar. 2003.

[2] R. Hashimoto et al., "Imaging genetics and psychiatric disorders," *Current Mol. Med.*, vol. 15, no. 2, pp. 168–175, 2015.

[3] D. Liu, X. Lin, and D. Ghosh, "Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models," *Biometrics*, vol. 63, no. 4, pp. 1079–1088, Dec. 2007, doi: 10.1111/j.1541-0420.2007.00799.x.

[4] X. Zhu, H.-I. Suk, H. Huang, and D. Shen, "Low-rank graph-regularized structured sparse regression for identifying genetic biomarkers," *IEEE Trans. Big Data*, vol. 3, no. 4, pp. 405–414, Dec. 2017.

[5] H. Wang et al., "Identifying quantitative trait loci via group-sparse multitask regression and feature selection: An imaging genetics study of the ADNI cohort," *Bioinformatics*, vol. 28, no. 2, pp. 229–237, Jan. 2012.

IEEE Journal of Translational Engineering in Health and Medicine

[6] M. Vounou, T. E. Nichols, and G. Montana, "Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach," *NeuroImage*, vol. 53, no. 3, pp. 1147–1159, Nov. 2010, doi: 10.1016/j.neuroimage.2010.07.002.

[7] D. M. Witten and R. J. Tibshirani, "Extensions of sparse canonical correlation analysis with applications to genomic data," *Stat. Appl. Genet. Mol. Biol.*, vol. 8, no. 1, pp. 1–27, Jan. 2009.

[8] X. Hao et al., "Mining outcome-relevant brain imaging genetic associations via three-way sparse canonical correlation analysis in Alzheimer's disease," *Sci. Rep.*, vol. 7, p. 44272, Aug. 2017.

[9] L. T. Elliott et al., "Genome-wide association studies of brain imaging phenotypes in U.K. biobank," *Nature*, vol. 562, no. 7726, pp. 210–216, Oct. 2018, doi: 10.1038/s41586-018-0571-7.

[10] K. L. Grasby et al., "The genetic architecture of the human cerebral cortex," *Science*, vol. 367, no. 6484, Mar. 2020, doi: 10.1126/science.aay6690.

[11] V. Uurtio, S. Bhadra, and J. Rousu, "Large-scale sparse kernel canonical correlation analysis," in *Proc. 36th Int. Conf. Mach. Learn.*, in Proceedings of Machine Learning Research, vol. 97, 2019, pp. 6383–6391. [Online]. Available: https://proceedings.mlr.press/v97/uurtio19a.html

[12] J. Fang, D. Lin, S. C. Schulz, Z. Xu, V. D. Calhoun, and Y.-P. Wang, "Joint sparse canonical correlation analysis for detecting differential imaging genetics modules," *Bioinformatics*, vol. 32, no. 22, pp. 3480–3488, Nov. 2016.

[13] M. Milchenko and D. Marcus, "Obscuring surface anatomy in volumetric imaging data," *Neuroinformatics*, vol. 11, no. 1, pp. 65–75, Jan. 2013.

[14] S. Moeller et al., "Multiband multislice GE-EPI at 7 Tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI," *Magn. Reson. Med.*, vol. 63, no. 5, pp. 1144–1153, May 2010.

[15] J. Gorski, F. Pfeuffer, and K. Klamroth, "Biconvex sets and optimization with biconvex functions: A survey and extensions," *Math. Methods Oper. Res.*, vol. 66, no. 3, pp. 373–407, Nov. 2007, doi: 10.1007/s00186-007-0161-1.

[16] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. J. Behrens, E. Yacoub, and K. Ugurbil, "The WU-Minn human connectome project: An overview," *NeuroImage*, vol. 80, pp. 62–79, Oct. 2013.

[17] D. S. Marcus et al., "Human connectome project informatics: Quality control, database services, and data visualization," *NeuroImage*, vol. 80, pp. 202–219, Oct. 2013, doi: 10.1016/j.neuroimage.2013.05.077.

[18] A. M. Dale, B. Fischl, and M. I. Sereno, "Cortical surface-based analysis: I. Segmentation and surface reconstruction," *NeuroImage*, vol. 9, no. 2, pp. 179–194, 1999.

[19] B. Fischl, M. I. Sereno, and A. M. Dale, "Cortical surface-based analysis: II: Inflation, flattening, and a surface-based coordinate system," *NeuroImage*, vol. 9, no. 2, pp. 195–207, 1999.

[20] B. Fischl, M. I. Sereno, R. B. H. Tootell, and A. M. Dale, "High-resolution intersubject averaging and a coordinate system for the cortical surface," *Hum. Brain Mapping*, vol. 8, no. 4, pp. 272–284, 1999.

[21] D. C. Van Essen, M. F. Glasser, D. L. Dierker, J. Harwell, and T. Coalson, "Parcellations and hemispheric asymmetries of human cerebral cortex analyzed on surface-based atlases," *Cerebral Cortex*, vol. 22, no. 10, pp. 2241–2262, Oct. 2012, doi: 10.1093/cercor/bhr291.

[22] M. F. Glasser et al., "A multi-modal parcellation of human cerebral cortex," *Nature*, vol. 536, no. 7615, pp. 171–178, Aug. 2016, doi: 10.1038/nature18933.

[23] C.-C. Huang, E. T. Rolls, C.-C.-H. Hsu, J. Feng, and C.-P. Lin, "Extensive cortical connectivity of the human hippocampal memory system: Beyond the 'what' and 'where' dual stream model," *Cerebral Cortex*, vol. 31, no. 10, pp. 4652–4669, Aug. 2021.

[24] R. Vos de Wael, F. Hyder, and G. J. Thompson, "Effects of tissue-specific functional magnetic resonance imaging signal regression on resting-state functional connectivity," *Brain Connectivity*, vol. 7, no. 8, pp. 482–490, Oct. 2017, doi: 10.1089/brain.2016.0465.

[25] K. Murphy and M. D. Fox, "Towards a consensus regarding global signal regression for resting state functional connectivity MRI," *NeuroImage*, vol. 154, pp. 169–173, Jul. 2017, doi: 10.1016/j.neuroimage.2016.11.052.

[26] J. C. Haselgrove and J. R. Moore, "Correction for distortion of echo-planar images used to calculate the apparent diffusion coefficient," *Magn. Reson. Med.*, vol. 36, pp. 960–964, May 1996, doi: 10.1002/mrm.1910360620.

[27] S. Purcell et al., "PLINK: A tool set for whole-genome association and population-based linkage analyses," *Amer. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–575, Sep. 2007, doi: 10.1086/519795.

[28] M. F. Glasser et al., "The minimal preprocessing pipelines for the human connectome project," *NeuroImage*, vol. 80, pp. 105–124, Oct. 2013, doi: 10.1016/j.neuroimage.2013.04.127. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811913005053

[29] Z. Yu-Feng et al., "Altered baseline brain activity in children with ADHD revealed by resting-state functional MRI," *Brain Develop.*, vol. 29, no. 2, pp. 83–91, Mar. 2007, doi: 10.1016/j.braindev.2006.07.002.

[30] M. P. van den Heuvel, C. J. Stam, M. Boersma, and H. E. H. Pol, "Small-world and scale-free organization of voxel-based resting-state functional connectivity in the human brain," *NeuroImage*, vol. 43, no. 3, pp. 528–539, Nov. 2008, doi: 10.1016/j.neuroimage.2008.08.010.

[31] B. T. Thomas Yeo et al., "The organization of the human cerebral cortex estimated by intrinsic functional connectivity," *J. Neurophysiol.*, vol. 106, no. 3, pp. 1125–1165, Sep. 2011, doi: 10.1152/jn.00338.2011.

[32] M. Kim, J. H. Won, J. Youn, and H. Park, "Joint-Connectivity-Based sparse canonical correlation analysis of imaging genetics for detecting biomarkers of Parkinson's disease," *IEEE Trans. Med. Imag.*, vol. 39, no. 1, pp. 23–34, Jan. 2020, doi: 10.1109/TMI.2019.2918839.

[33] Q.-H. Zou et al., "An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: Fractional ALFF," *J. Neurosci. Methods*, vol. 172, no. 1, pp. 137–141, Jul. 2008, doi: 10.1016/j.jneumeth.2008.04.012.

[34] L. J. Hearne, J. B. Mattingley, and L. Cocchi, "Functional brain networks related to individual differences in human intelligence at rest," *Sci. Rep.*, vol. 6, no. 1, p. 32328, Aug. 2016, doi: 10.1038/srep32328.

[35] K. H. Karlsgodt et al., "A multimodal assessment of the genetic control over working memory," *J. Neurosci.*, vol. 30, no. 24, pp. 8197–8202, Jun. 2010.

[36] P. M. Thompson et al., "Genetic influences on brain structure," *Nature Neurosci.*, vol. 4, no. 12, pp. 1253–1258, 2001.

[37] M. D. Greicius, K. Supekar, V. Menon, and R. F. Dougherty, "Resting-state functional connectivity reflects structural connectivity in the default mode network," *Cerebral Cortex*, vol. 19, no. 1, pp. 72–78, Jan. 2009.

[38] V. Warrier et al., "Genetic insights into human cortical organization and development through genome-wide analyses of 2,347 neuroimaging phenotypes," *Nature Genet.*, vol. 55, no. 9, pp. 1483–1493, Sep. 2023, doi: 10.1038/s41588-023-01475-y.

[39] S. Berto et al., "Association between resting-state functional brain connectivity and gene expression is altered in autism spectrum disorder," *Nature Commun.*, vol. 13, no. 1, p. 3328, Jun. 2022, doi: 10.1038/s41467-022-31053-5.

[40] *The Human Protein Atlas*. Accessed: Sep. 15, 2024. [Online]. Available: https://www.proteinatlas.org/ENSG00000145819-ARHGAP26/brain

[41] S. Glerup et al., "SorCS2 regulates dopaminergic wiring and is processed into an apoptotic two-chain receptor in peripheral glia," *Neuron*, vol. 82, no. 5, pp. 1074–1087, Jun. 2014, doi: 10.1016/j.neuron.2014.04.022.

[42] K. Baker et al., "*SYT*1-associated neurodevelopmental disorder: A case series," *Brain*, vol. 141, no. 9, pp. 2576–2591, Sep. 2018, doi: 10.1093/brain/awy209.

[43] T. Yarkoni, R. Poldrack, T. Nichols, D. Van Essen, and T. Wager, "NeuroSynth: A new platform for large-scale automated synthesis of human functional neuroimaging data," in *Proc. 4th INCF Congr. Neuroinf.*, 2011, doi: 10.3389/conf.fninf.2011.08.00058.