

RESEARCH ARTICLE

Multimodal data integration via mediation analysis with high-dimensional exposures and mediators

Yi Zhao¹  | Lexin Li² | Alzheimer's Disease Neuroimaging Initiative

¹Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, Indiana, USA

²Department of Biostatistics and Epidemiology, University of California, Berkeley, Berkeley, California, USA

Correspondence

Yi Zhao, Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN 46202, USA.
Email: yz125@iu.edu

Funding information

National Institute on Aging, Grant/Award Numbers: P30AG072976, R01AG034570, R01AG061303, R01AG062542, U54AG065181; National Science Foundation, Grant/Award Number: CIF-2102227

Abstract

Motivated by an imaging proteomics study for Alzheimer's disease (AD), in this article, we propose a mediation analysis approach with high-dimensional exposures and high-dimensional mediators to integrate data collected from multiple platforms. The proposed method combines principal component analysis with penalized least squares estimation for a set of linear structural equation models. The former reduces the dimensionality and produces uncorrelated linear combinations of the exposure variables, whereas the latter achieves simultaneous path selection and effect estimation while allowing the mediators to be correlated. Applying the method to the AD data identifies numerous interesting protein peptides, brain regions, and protein-structure-memory paths, which are in accordance with and also supplement existing findings of AD research. Additional simulations further demonstrate the effective empirical performance of the method.

KEYWORDS

Alzheimer's disease, mediation analysis, multimodal data integration, neuroimaging, principal component analysis

1 | INTRODUCTION

Alzheimer's disease (AD) is an irreversible neurodegenerative disorder and is characterized by progressive impairment of cognitive and bodily functions and ultimate death. It is currently affecting over 5.8 million American adults aged 65 years or older. Meanwhile, its prevalence continues to grow and is projected to reach 13.8 million by 2050 (Alzheimer's Association, 2020). Multimodal technologies have transformed AD research in recent years, by collecting different types of data from the same group of subjects and enabling the investigation

of complex interrelated mechanisms underlying AD development. Notable examples include multimodal neuroimaging studies of the joint impact of brain structure and function on the disorders (Higgins, Kundu, & Guo, 2018; Liu et al., 2015), and imaging genetics studies of the impact of genetic variants on the brain then the disease outcome (Nathoo et al., 2019), among others.

Our motivation is an imaging proteomics study, which is part of the Alzheimer's Disease Neuroimaging Initiative (ADNI) that aims to identify biomarkers for early detection and tracking of AD and to assist the development of prevention and intervention strategies. Amyloid- β is a microscopic brain protein fragment, denotes peptides of 36–43 amino acids, and is part of a larger protein called amyloid precursor protein. Tau is a group of microtubule-associated proteins predominantly found in brain cells and performs the function of stabilizing microtubules. Amyloid- β is the main component of amyloid plaques, while tau is the main component of neurofibrillary tangles,

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete list of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *Human Brain Mapping* published by Wiley Periodicals LLC.

both of which are commonly found in the brains of AD patients. Models of AD pathophysiology hypothesize a temporal sequence, in which accumulations of amyloid- β plaques and neurofibrillary tangles disrupt cell-to-cell communications and destroy brain cells, leading to brain structural atrophy in regions such as the hippocampus, and ultimately a clinical decline in cognition (Mormino et al., 2009). However, it remains unclear how these two proteins interact with each other and with other proteins in the cerebrospinal fluid (CSF), and how those proteins together subsequently affect brain atrophy and disease progression. In our study, we aim to investigate simultaneously the interrelations of multiple protein peptides in the CSF, along with multiple brain regions of the whole brain, and their impact on memory.

The problem can be formulated as a mediation analysis, where the goal is to identify and explain the mechanism, or path, that underlies an observed relationship between an exposure and an outcome variable, through the inclusion of an intermediate variable known as a mediator. It decomposes the effect of exposure on the outcome into a direct effect and an indirect effect, the latter of which indicates whether the mediator is on a path from the exposure to the outcome. In our multimodal AD study, the measurements of the amount of multiple protein peptides serve as the exposure variables, the volumetric measurements of multiple brain regions serve as the potential mediators, and a composite memory score serves as the outcome. See section 2 for more details about the study and the data. Our objective is to identify paths from proteins to brain regional atrophies that lead to memory decline.

Mediation analysis was first proposed with a single exposure and a single mediator (Baron & Kenny, 1986). See VanderWeele (2016) for a review of mediation analysis and many references therein. In our setting, both the exposure variables and mediators are multivariate and potentially high-dimensional. While there have been numerous extensions of mediation analysis to account for multiple mediators (see, e.g., Chén et al., 2017; Song et al., 2018; Zhao & Luo, 2022, among many others), there have been very few works studying multivariate exposures, or both multivariate exposures and mediators. Recently, Aung et al. (2020), Long, Irajizad, Doecke, Do, and Ha (2020), and Zhang (2021) proposed new approaches for mediation analysis of multivariate exposures and mediators. In particular, Zhang (2021) developed two regularization procedures and applied them to a mouse f2 dataset for diabetes, taking SNP genotypes as the exposures, islet gene expressions as the mediators, and insulin level as the outcome. However, they required the mediators to be independent, which hardly holds in our setting, as different brain regions are generally believed to influence each other. Aung et al. (2020) studied environmental toxicants on pregnancy outcomes, taking toxicants as the exposures, endogenous biomarkers such as inflammation and oxidative stress as the mediators, and gestational age at delivery as the outcome. A key strategy of their analysis was to reduce the exposure dimension by creating environmental risk scores for a small number of groups based on the domain knowledge. They showed that the between-group correlation in the reduced exposures is negligible. However, such prior domain knowledge may not always be available. Long et al. (2020) proposed a general mediation framework to identify

proteins that mediate the effect of metabolic gene expressions on survival for a type of kidney cancer, taking mRNA levels as the exposures, protein measures as the mediators, and survival time as the outcome. Nevertheless, they implicitly required the dimensions of the exposures and mediators cannot be too high, and thus their method is not directly applicable to our setting, where the number of exposures and mediators can both be potentially larger than the sample size.

In this article, we propose a mediation analysis approach, with both high-dimensional exposures and high-dimensional mediators, for multimodal data analysis. The method integrates principal components analysis (PCA) with penalized least squares estimation for a set of linear structural equation models. The former reduces the dimensionality and produces uncorrelated linear combinations of the exposure variables, whereas the latter achieves path selection and effect estimation while allowing the multivariate mediators to be potentially correlated. We apply this approach to the imaging proteomics study of AD to integrate CSF proteomics, brain volumes, and a memory measure of mild cognitive impairment (MCI) subjects in ADNI. We identify several interesting protein peptides, brain regions, and protein-structure-memory paths that are in accordance with and also supplement the existing knowledge of AD. Additional simulations further demonstrate the efficacy of the method. Similar to Aung et al. (2020), Long et al. (2020), and Zhang (2021), our approach is among the first attempts to conduct mediation analysis where both the exposures and mediators are high-dimensional. But unlike the existing solutions, we do not restrict the dimensionality or the correlation structures and do not require additional domain knowledge of the exposures or mediators. Moreover, although focusing on a multimodal neuroimaging study in this article, our proposed method is equally applicable to a wide range of multimodal data integration problems, for example, the multi-omics data analysis (Richardson, Tseng, & Sun, 2016), and the multimodal healthcare study (Cai, Wang, Li, & Liu, 2019). As such, our proposal makes a useful addition to the general toolbox of both mediation analysis and multimodal data integration.

The rest of the article is organized as follows. Section 2 introduces the motivating imaging proteomics data of AD. Section 3 presents the proposed model and estimation approach. Section 4 analyzes the AD dataset, with a detailed discussion on the identified protein peptides, brain regions, and path. Section 5 complements with additional simulation results to demonstrate the empirical performance of the method.

2 | AD IMAGING PROTEOMICS STUDY

While Alzheimer's disease is becoming a major public health challenge as the population ages, there is no effective treatment for AD that is capable of stopping or slowing the associated cognitive and neuronal degradation. Therefore, understanding the disease pathology, identifying biological markers, and finding early diagnosis and intervention strategies are of critical importance (Alzheimer's Association, 2020). Among numerous AD-related proteins in the CSF, amyloid- β and tau are two major proteins that are consistently identified in the brains of

AD patients, and their abnormal abundance generally indicates AD pathology (Jagust, 2018). Even though there has been evidence suggesting a pathological connection between amyloid- β deposition, hippocampus atrophy, and memory decline (Mormino et al., 2009), it remains largely unknown how amyloid- β and tau interact with each other, how they interact with other proteins in the CSF, and how these proteins together affect the downstream brain atrophy and cognitive outcome. In our study, we aim to delineate the regulatory relationships among multiple CSF proteins, structural atrophy of the whole brain, and cognitive behavior, and to identify important biological paths.

The data used in our study are obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI, adni.loni.usc.edu). The CSF proteomics data were obtained using targeted liquid chromatography multiple reaction monitoring mass spectrometry, which is a highly specific, sensitive, and reproducible technique for quantifying targeted proteins. A list of protein fragments, or peptides, was sent to the detector. The samples then went through peak integration, outliers detection, normalization, quantification, and quality control using test/re-test samples. This procedure results in the intensity measures of 320 peptides that are annotated from 142 proteins. The brain imaging data were obtained using anatomical magnetic resonance imaging (MRI). Each image was first preprocessed following the standard pipeline, then mapped to an atlas consisting of 145 brain regions-of-interest to extract the volumetric measures (Doshi et al., 2016). The atlas used in the study spans the entire brain and was actually built on multiple atlases. Individual atlases were first warped to the target image using a nonlinear registration method, followed by a spatially adaptive weighted voting strategy to fuse into a final segmentation. Moreover, the volume of each brain region was standardized by the total intracranial volume to account for variations of individual brain size. The cognitive outcome is a composite memory score, ADNI-MEM, that involves a battery of neuropsychological tests. In our study, we focus on 135 subjects diagnosed as mild cognitive impairment (MCI) patients at recruitment. MCI is a prodromal stage of AD, with a slight but noticeable and measurable decline in cognitive abilities. A person with MCI is at an increased risk of developing AD or other dementia. Understanding the pathologic mechanism underlying MCI provides important clues of onset of the disorder as well as a useful guide for early diagnosis and intervention.

3 | MODEL AND METHOD

We first present the proposed model, then an estimation method integrating principal components analysis and penalized estimation.

3.1 | Model

Suppose there are totally n subjects. Let $\mathbf{X}_i = (X_{i1}, \dots, X_{ir})^T \in \mathbb{R}^r$ denote the r -dimensional vector of exposure variables, $\mathbf{M}_i = (M_{i1}, \dots, M_{ip})^T \in \mathbb{R}^p$ denote the p -dimensional vector of mediators, and $Y_i \in \mathbb{R}$ denote the

univariate outcome variable, for subjects $i = 1, \dots, n$. In our imaging proteomics study, \mathbf{X}_i denotes the protein peptide measures with $r = 320$, \mathbf{M}_i denotes the brain volumetric measures with $p = 145$, Y_i denotes the memory score, and the sample size $n = 135$.

The first step of our method is to perform a principal components analysis on \mathbf{X}_i to produce uncorrelated composite exposures. If \mathbf{X}_i further follows a multivariate normal distribution, then the produced composite exposures are independent. Let $\tilde{\mathbf{X}}_i = (\tilde{X}_{i1}, \dots, \tilde{X}_{iq})^T \in \mathbb{R}^q$ denote the first q principal components. We then continue to model the path relations among $\tilde{\mathbf{X}}_i, \mathbf{M}_i$ and Y_i via the following set of linear structural equation models,

$$\begin{aligned} \mathbf{M} &= \tilde{\mathbf{X}}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \\ \mathbf{Y} &= \tilde{\mathbf{X}}\boldsymbol{\gamma} + \mathbf{M}\boldsymbol{\beta} + \boldsymbol{\eta}, \end{aligned} \quad (1)$$

where $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_1, \dots, \tilde{\mathbf{X}}_n)^T \in \mathbb{R}^{n \times q}$, $\mathbf{M} = (\mathbf{M}_1, \dots, \mathbf{M}_n)^T \in \mathbb{R}^{n \times p}$, $\mathbf{Y} = (Y_1, \dots, Y_n)^T \in \mathbb{R}^n$ stack the composite exposures, mediators, and outcome across all subjects, respectively, $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^T \in \mathbb{R}^{n \times p}$, with $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{ip})^T \in \mathbb{R}^p$, and $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^T \in \mathbb{R}^n$ are measurement errors. Suppose both error terms follow some zero mean normal distribution, and $\boldsymbol{\epsilon}$ is independent of $\tilde{\mathbf{X}}$, $\boldsymbol{\eta}$ is independent of $\tilde{\mathbf{X}}$ and \mathbf{M} , and $\boldsymbol{\epsilon}$ and $\boldsymbol{\eta}$ are independent of each other. The parameters $\boldsymbol{\alpha} = (\alpha_{jk}) \in \mathbb{R}^{q \times p}$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T \in \mathbb{R}^p$, and $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_q)^T \in \mathbb{R}^q$ capture the path effects. Model 1 is similar to that used in Zhao, Li, and Caffo (2021) and Zhao and Luo (2022), but none of those can handle multivariate exposure variables. Besides, we introduce some different forms of penalty functions in our parameter estimation.

Figure 1 shows a schematic description of Model 1. Under this model, we define the direct effect of \tilde{X}_j on Y as $\text{DE}(\tilde{X}_j) = \gamma_j$, the indirect effect of \tilde{X}_j on Y through M_k as $\text{IE}(\tilde{X}_j, M_k) = \alpha_{jk}\beta_k$, and the total indirect effect of \tilde{X}_j on Y as $\text{IE}(\tilde{X}_j) = \sum_{k=1}^p \alpha_{jk}\beta_k$, for $j = 1, \dots, q$. The total effect of \tilde{X}_j satisfies that $\text{TE}(\tilde{X}_j) = \text{IE}(\tilde{X}_j) + \text{DE}(\tilde{X}_j) = \sum_{k=1}^p \alpha_{jk}\beta_k + \gamma_j$.

A key characteristic of Model 1 is that it allows the multivariate mediators to be conditionally dependent given the exposures. To better illustrate this, we consider a simple example of Model 1, where $q = 1, p = 3$, as shown in Figure 2. In this example, Figure 2a outlines the sequential influences among all the mediators, while Figure 2b is the proposed Model 1. We see that, for the first mediator, M_1 , $\alpha_{11} = a_{11}, \beta_1 = b_1$; for the second mediator, M_2 , $\alpha_{12} = a_{11}d_{12} + a_{12}, \beta_2 = b_2$; and for the third mediator, M_3 , $\alpha_{13} = a_{11}d_{13} + a_{11}d_{12}d_{23} + a_{12}d_{23} + a_{13}, \beta_3 = b_3$. As such, α_{1k} consolidates the effects through the k th mediator M_k , and the indirect effect $\text{IE}(\tilde{X}_1, M_k) = \alpha_{1k}\beta_k$ can be viewed as the consolidated indirect effect through $M_k, k = 1, 2, 3$.

3.2 | Estimation

We propose to estimate the parameters in Model 1 through the penalized ordinary least squares,

$$\underset{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}}{\text{minimize}} \quad \frac{1}{2} \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) + \lambda_1 \mathcal{R}_1(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda_2 \mathcal{R}_2(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda_3 \mathcal{R}_3(\boldsymbol{\gamma}), \quad (2)$$

where the loss function is the usual least squares loss,

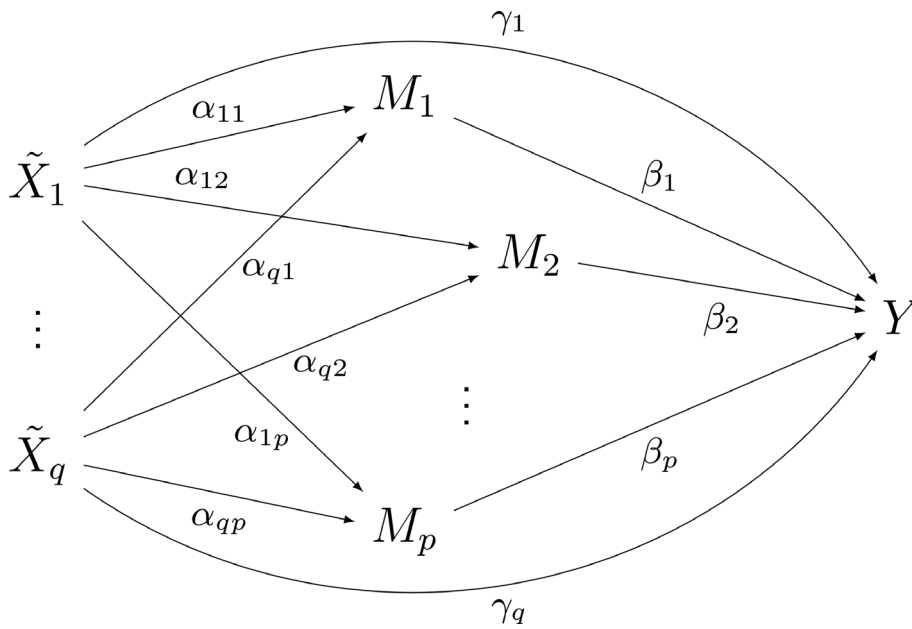


FIGURE 1 The schematic diagram of the proposed model with q exposure variables $\tilde{X}_1, \dots, \tilde{X}_q$, p mediators M_1, \dots, M_p , and the outcome variable Y

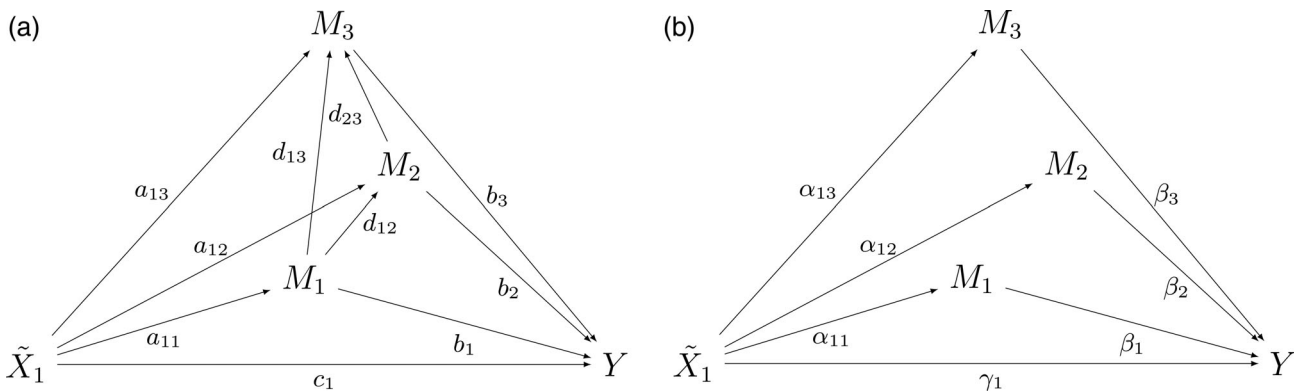


FIGURE 2 A model example with $q = 1$ exposure variable and $p = 3$ sequentially ordered mediators

$$\mathcal{L}(\alpha, \beta, \gamma) = \text{tr} \left\{ (\mathbf{M} - \tilde{\mathbf{X}}\alpha)^T (\mathbf{M} - \tilde{\mathbf{X}}\alpha) \right\} + (\mathbf{Y} - \tilde{\mathbf{X}}\gamma - \mathbf{M}\beta)^T (\mathbf{Y} - \tilde{\mathbf{X}}\gamma - \mathbf{M}\beta).$$

$\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$ are three penalty functions, with the tuning parameters $\lambda_1, \lambda_2, \lambda_3$, respectively. We next discuss each penalty function in detail.

The first penalty function \mathcal{R}_1 is of the form,

$$\mathcal{R}_1(\alpha, \beta) = \sum_{j=1}^q \sum_{k=1}^p \left\{ |\alpha_{jk}\beta_k| + c_0 (\alpha_{jk}^2 + \beta_k^2) \right\} + c_1 \left(\sum_{j=1}^q \sum_{k=1}^p |\alpha_{jk}| + \sum_{k=1}^p |\beta_k| \right),$$

for some parameters c_0 and c_1 . It is a generalization of the pathway Lasso penalty of Zhao and Luo (2022) to q exposure variables, and is to facilitate selection of individual mediators. Specifically, for a given mediator M_k , the term $\sum_{j=1}^q |\alpha_{jk}\beta_k|$ is a product Lasso penalty, and encourages all the paths going through M_k to be shrunk to zero,

which in effect achieves the goal of mediator selection. The term $c_0 (\alpha_{jk}^2 + \beta_k^2)$ is to make the penalty a convex function, with a proper choice of the parameter c_0 . It is straightforward to show that, when $c_0 \geq 1/2$, the sum $|\alpha_{jk}\beta_k| + c_0 (\alpha_{jk}^2 + \beta_k^2)$ is convex. In our implementation, we fix $c_0 = 2$. The last term in \mathcal{R}_1 is the sum of usual Lasso penalty that further penalizes individual path effects α_{jk}, β_k , with c_1 being an additional tuning parameter. It is found that this additional penalty helps further improves the selection accuracy (Zhao & Luo, 2022).

The second penalty function \mathcal{R}_2 is of the form,

$$\mathcal{R}_2(\alpha, \beta) = \sum_{j=1}^q \sqrt{p} \sqrt{\sum_{k=1}^p (\alpha_{jk}\beta_k)^2}.$$

It is a group Lasso penalty and is to facilitate the selection of individual exposure. Specifically, for a given exposure \tilde{X}_j , the penalty $\left\{ \sum_{k=1}^p (\alpha_{jk}\beta_k)^2 \right\}^{1/2}$ encourages all the paths originating from \tilde{X}_j to be

shrunk to zero, which in effect achieves the goal of exposure selection.

The third penalty function \mathcal{R}_3 is of the form,

$$\mathcal{R}_3(\boldsymbol{\gamma}) = \sum_{j=1}^q |\gamma_j|.$$

This is simply the usual Lasso penalty and is to facilitate selection of direct effects between the exposures and the outcome.

We next discuss how to solve the minimization problem (2). We note that (2) involves the penalties on the product terms $\alpha_{jk}\beta_k$, making it difficult to derive the analytical solutions. As such, we first introduce a new parameter, $\mu_{jk} = \alpha_{jk}\beta_k$, which turns (2) to an equivalent problem of solving a sparse group lasso that has an explicit form of solution (Simon, Friedman, Hastie, & Tibshirani, 2013). That is, letting $\boldsymbol{\mu} = (\mu_{jk}) \in \mathbb{R}^{q \times p}$, we turn to the equivalent optimization problem,

$$\begin{aligned} & \text{minimize}_{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}} \frac{1}{2} \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) + \lambda_1 \mathcal{R}_1(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda_2 \mathcal{R}_2(\boldsymbol{\mu}) + \lambda_3 \mathcal{R}_3(\boldsymbol{\gamma}), \\ & \text{suchthat } \mu_{jk} = \alpha_{jk}\beta_k, \text{ for } j = 1, \dots, q \text{ and } k = 1, \dots, p. \end{aligned} \quad (3)$$

Let $\boldsymbol{\mu}_j = (\mu_{j1}, \dots, \mu_{jp})^T \in \mathbb{R}^p$, $\boldsymbol{\alpha}_j = (\alpha_{j1}, \dots, \alpha_{jp})^T \in \mathbb{R}^p$, and introduce the augmented Lagrangian parameter $\boldsymbol{\tau}_j = (\tau_{j1}, \dots, \tau_{jp})^T \in \mathbb{R}^p$, for $j = 1, \dots, q$, and $\boldsymbol{\tau} = (\boldsymbol{\tau}_j) \in \mathbb{R}^{q \times p}$. Then, the augmented Lagrangian form of (3) is

$$\begin{aligned} & \text{minimize}_{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\tau}} \frac{1}{2} \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) + \lambda_1 \mathcal{R}_1(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda_2 \mathcal{R}_2(\boldsymbol{\mu}) + \lambda_3 \mathcal{R}_3(\boldsymbol{\gamma}) \\ & + \sum_{j=1}^q \left(\langle \boldsymbol{\mu}_j - \boldsymbol{\alpha}_j \circ \boldsymbol{\beta}, \boldsymbol{\tau}_j \rangle + \frac{\rho}{2} \|\boldsymbol{\mu}_j - \boldsymbol{\alpha}_j \circ \boldsymbol{\beta}\|_2^2 \right), \end{aligned} \quad (4)$$

where $\rho > 0$ is the augmented Lagrangian constant that we set $\rho = 1$ in our implementation, \circ is the Hadamard product, $\langle \cdot, \cdot \rangle$ is the inner product, and $\|\cdot\|_2$ is the L_2 -norm. We next solve (4) by updating $\boldsymbol{m}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{g}$ and \boldsymbol{t} iteratively.

More specifically, we first fix $\boldsymbol{\alpha}^{(s)}, \boldsymbol{\beta}^{(s)}, \boldsymbol{\gamma}^{(s)}, \boldsymbol{\tau}^{(s)}$ at iteration s , and update $\boldsymbol{\mu}_j$ by solving

$$\text{minimize}_{\boldsymbol{\mu}_j} \frac{\rho}{2} \|\boldsymbol{\mu}_j - \boldsymbol{\alpha}_j^{(s)} \circ \boldsymbol{\beta}^{(s)}\|_2^2 + \boldsymbol{\tau}_j^{(s)T} (\boldsymbol{\mu}_j - \boldsymbol{\alpha}_j^{(s)} \circ \boldsymbol{\beta}^{(s)}) + \lambda_1 \|\boldsymbol{\mu}_j\|_1 + \lambda_2 \sqrt{\bar{\rho}} \|\boldsymbol{\mu}_j\|_2,$$

for $j = 1, \dots, q$, where $\|\cdot\|_1$ is the L_1 -norm. There is a closed-form solution,

$$\boldsymbol{\mu}_{jk}^{(s+1)} = \begin{cases} \left\{ \|\mathcal{S}(\boldsymbol{\nu}_j, \lambda_1/\rho)\|_2 - \lambda_2 \sqrt{\bar{\rho}}/\rho \right\}_+ \frac{\mathcal{S}(\boldsymbol{\nu}_j, \lambda_1/\rho)}{\|\mathcal{S}(\boldsymbol{\nu}_j, \lambda_1/\rho)\|_2}, & \text{if } \|\mathcal{S}(\boldsymbol{\nu}_j, \lambda_1/\rho)\|_2 \neq 0, \\ 0 & \text{otherwise,} \end{cases} \quad (5)$$

for $j = 1, \dots, q, k = 1, \dots, p$, where $\boldsymbol{\nu}_j = \boldsymbol{\alpha}_j^{(s)} \circ \boldsymbol{\beta}^{(s)} - \boldsymbol{\tau}_j^{(s)}/\rho$, $\mathcal{S}(a, \lambda) = \text{sgn}(a) \max\{|a| - \lambda, 0\}$ is the soft-thresholding function with $\text{sgn}(a)$ denoting the sign of a and $a_+ = \max\{a, 0\}$, and $\mathcal{S}(\boldsymbol{a}, \lambda)$ denotes the element-wise soft-thresholding of a vector \boldsymbol{a} .

We next fix $\boldsymbol{m}^{(s+1)}, \boldsymbol{b}^{(s)}, \boldsymbol{g}^{(s)}, \boldsymbol{t}^{(s)}$, and update \boldsymbol{a}_j by solving

$$\text{minimize}_{\boldsymbol{a}_j} \boldsymbol{V}_j \boldsymbol{a}_j + \lambda_1 c_1 \text{sgn}(\boldsymbol{a}_j) - \boldsymbol{w}_j,$$

where $\boldsymbol{V}_j = \rho \boldsymbol{D}_{\boldsymbol{\beta}^{(s)}}^2 + (4\lambda_1 + \tilde{\boldsymbol{x}}_j^T \tilde{\boldsymbol{x}}_j) \boldsymbol{I}_p$, $\boldsymbol{w}_j = (\boldsymbol{M} - \sum_{l \neq j} \tilde{\boldsymbol{x}}_l \boldsymbol{\alpha}_l^{(s)T})^T \tilde{\boldsymbol{x}}_j + \boldsymbol{D}_{\boldsymbol{\beta}^{(s)}} \boldsymbol{\tau}_j^{(s)} + \rho \boldsymbol{D}_{\boldsymbol{\beta}^{(s)}} \boldsymbol{\mu}_j^{(s+1)}$, $\boldsymbol{D}_{\boldsymbol{\beta}^{(s)}}$ is a diagonal matrix with $\boldsymbol{\beta}^{(s)}$ as the diagonal elements, $\tilde{\boldsymbol{x}}_j \in \mathbb{R}^n$ is the j th column of $\tilde{\boldsymbol{X}}$, and \boldsymbol{I}_p is the p -dimensional identity matrix. The solution is

$$\boldsymbol{a}_j^{(s+1)} = \boldsymbol{V}_j^{-1} \mathcal{S}(\boldsymbol{w}_j, \lambda_1 c_1), \quad j = 1, \dots, q. \quad (6)$$

We next fix $\boldsymbol{\mu}^{(s+1)}, \boldsymbol{\alpha}^{(s+1)}, \boldsymbol{\gamma}^{(s)}, \boldsymbol{\tau}^{(s)}$, and update $\boldsymbol{\beta}$ by solving

$$\text{minimize}_{\boldsymbol{\beta}} \boldsymbol{V}_{\boldsymbol{\beta}} \boldsymbol{\beta} + \lambda_1 \text{sgn}(\boldsymbol{\beta}) - \boldsymbol{w}_{\boldsymbol{\beta}},$$

where $\boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{M}^T \boldsymbol{M} + \rho \sum_{j=1}^q \boldsymbol{D}_{\boldsymbol{\alpha}_j^{(s+1)}}^2 + 4\lambda_1 q \boldsymbol{I}_p$, $\boldsymbol{w}_{\boldsymbol{\beta}} = \boldsymbol{M}^T (\boldsymbol{Y} - \tilde{\boldsymbol{X}} \boldsymbol{\gamma}^{(s)}) + \sum_{j=1}^q \boldsymbol{D}_{\boldsymbol{\alpha}_j^{(s+1)}} \boldsymbol{\tau}_j^{(s)} + \rho \sum_{j=1}^q \boldsymbol{D}_{\boldsymbol{\alpha}_j^{(s+1)}} \boldsymbol{\mu}_j^{(s+1)}$, and $\boldsymbol{D}_{\boldsymbol{\alpha}_j^{(s+1)}}$ is a diagonal matrix with $\boldsymbol{\alpha}_j^{(s+1)}$ as the diagonal elements. The solution is

$$\boldsymbol{\beta}^{(s+1)} = \boldsymbol{V}_{\boldsymbol{\beta}}^{-1} \mathcal{S}(\boldsymbol{w}_{\boldsymbol{\beta}}, \lambda_1 c_1). \quad (7)$$

We then fix $\boldsymbol{\mu}^{(s+1)}, \boldsymbol{\alpha}^{(s+1)}, \boldsymbol{\beta}^{(s+1)}, \boldsymbol{\tau}^{(s)}$, and update \boldsymbol{g} by solving

$$\text{minimize}_{\boldsymbol{\gamma}} \boldsymbol{V}_{\boldsymbol{\gamma}} \boldsymbol{\gamma} + \lambda_3 \text{sgn}(\boldsymbol{\gamma}) - \boldsymbol{w}_{\boldsymbol{\gamma}},$$

where $\boldsymbol{V}_{\boldsymbol{\gamma}} = \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}}$ and $\boldsymbol{w}_{\boldsymbol{\gamma}} = \tilde{\boldsymbol{X}}^T (\boldsymbol{Y} - \boldsymbol{M} \boldsymbol{b}^{(s+1)})$. The solution is

$$\boldsymbol{\gamma}^{(s+1)} = \boldsymbol{V}_{\boldsymbol{\gamma}}^{-1} \mathcal{S}(\boldsymbol{w}_{\boldsymbol{\gamma}}, \lambda_3). \quad (8)$$

Finally, we fix $\boldsymbol{\mu}^{(s+1)}, \boldsymbol{\alpha}^{(s+1)}, \boldsymbol{\beta}^{(s+1)}, \boldsymbol{\gamma}^{(s+1)}$, and update $\boldsymbol{\tau}$ by

$$\boldsymbol{\tau}_j^{(s+1)} = \boldsymbol{\tau}_j^{(s)} + \rho (\boldsymbol{\mu}_j^{(s+1)} - \boldsymbol{\alpha}_j^{(s+1)} \circ \boldsymbol{\beta}^{(s+1)}), \quad j = 1, \dots, q. \quad (9)$$

We stop the iterations until some stopping criterion is met. In our implement, we stop when the difference of two consecutive objective values is smaller than 10^{-6} . We summarize the above optimization procedure in Algorithm 1.

We tune the parameters in (4) using the Bayesian information criterion (BIC),

$$\text{BIC} = -2 \log \mathcal{L}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{\gamma}}) + \log(n) |\hat{\mathcal{A}}|,$$

where $\hat{\boldsymbol{a}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}$ are the estimates under a given set of tuning parameters $\lambda_1, \lambda_2, \lambda_3$ and c_1 , $\mathcal{A} = \{(j, k) : \alpha_{jk} \beta_k \neq 0\}$ denotes the active set, and $|\mathcal{A}|$ is the cardinality. In our implementation, we adopt the tuning strategy of Zou and Hastie (2005), by tuning the ratios $\lambda_2/\lambda_1, \lambda_3/\lambda_1$ along with c_1 in a grid search, and choose the best set of parameters that minimizes the BIC.

Algorithm The optimization algorithm for (4)

Input: (\tilde{X}, M, Y) and the tuning parameters $\lambda_1, \lambda_2, \lambda_3$ and c_1

- 1: **initialization:** $\{\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)}, \mu^{(0)}, \tau^{(0)}\}$
- 2: **repeat**
- 3: update $\mu_{jk}^{(s+1)}$ given $\alpha^{(s)}, \beta^{(s)}, \gamma^{(s)}, \tau^{(s)}$ by (5),
for $j = 1, \dots, q, k = 1, \dots, p$
- 4: update $\alpha_j^{(s+1)}$ given $\mu^{(s+1)}, \beta^{(s)}, \gamma^{(s)}, \tau^{(s)}$ by (6),
for $j = 1, \dots, q$
- 5: update $\beta^{(s+1)}$ given $\mu^{(s+1)}, \alpha^{(s+1)}, \gamma^{(s)}, \tau^{(s)}$ by (7)
- 6: update $\gamma^{(s+1)}$ given $\mu^{(s+1)}, \alpha^{(s+1)}, \beta^{(s+1)}, \tau^{(s)}$ by (8)
- 7: update $\tau_j^{(s+1)}$ given $\mu^{(s+1)}, \alpha^{(s+1)}, \beta^{(s+1)}, \gamma^{(s+1)}$ by (9),
for $j = 1, \dots, q$
- 8: **until** the stopping criterion is met

Output: $\{\hat{\alpha}, \hat{\beta}, \hat{\gamma}\}$

4 | AD IMAGING PROTEOMICS STUDY REVISITED

We apply the proposed method to the ADNI imaging proteomics data, taking the CSF peptide measures as the exposures, the brain volumetric measures as the mediators, and the memory score as the outcome. Moreover, we adjust the exposures, mediators, and outcome for age, gender, ApoE4, and years of education to remove potential confounding effects (Rosenbaum, 2002). We first summarize the identified paths with nonzero effects, then discuss the relevant proteins and brain regions in detail. In summary, our findings are consistent with the existing knowledge of AD. Moreover, our method also suggests a few potentially interesting protein–structure–memory paths that may deserve further examination and verification.

4.1 | Paths with nonzero effects

We first apply principal components analysis to the peptide data. The top 20 principal components (PCs) account for about 85% of total data variation. We thus focus on those $q = 20$ top PCs and feed them as the exposure variables into the subsequent penalized path analysis. Figure 3 presents all the identified paths with a nonzero indirect path effect. Table 1 presents the estimated path effects including the estimated α and β of each path, and Table 2 presents the indirect, direct, and total effect of each exposure PC.

4.2 | Proteins

Among the 20 PCs, seven have nonzero indirect effects on memory. Next, we focus on PC1, PC4, and PC5 as they account for a higher proportion of total data variation and demonstrate a

relatively higher indirect path effect on the outcome. To better interpret the PCs, the loading profiles are sparsified following the sparse PCA approach (Zou, Hastie, & Tibshirani, 2006). The fused lasso regularization (Tibshirani, Saunders, Rosset, Zhu, & Knight, 2005) is considered to impose local consistency and smoothness within the same protein. Table 3 lists the top proteins in PC1, PC4, and PC5, and the corresponding gene name. We also include the regulation directions found in the AD literature, where an upregulation compared to cognitive normal controls indicates a higher protein abundance in MCI/AD patients, as well as the direction of correlations with the CSF amyloid- β and tau, the two well-established AD protein biomarkers (Wesenhagen, Teunissen, Visser, & Tijms, 2020). We next discuss the identified proteins by their relevance in the amyloid- β and tau pathology.

4.2.1 | Proteins related to amyloid pathology

Among the top-loaded proteins, SPON1, SORCS1, PTGDS, CST3, NPTX2, VGF, and CHGA have been found to be related to amyloid- β pathology in AD. The accumulation of amyloid- β is generally considered a hallmark of AD, which is derived from the amyloid precursor protein (APP) through sequential cleavages by beta-site amyloid precursor protein cleaving enzyme 1 (BACE1) and γ -secretase (Vassar et al., 1999). Blocking BACE1 can potentially reduce the abundance in amyloid- β , however, this may prohibit the other functions of BACE1 in psychological activities. For SPON1, using an in vivo AD mouse model, it was found that, by injecting SPON1, the amount of amyloid- β was significantly reduced, and subsequently, the ameliorated cognitive dysfunction and memory impairment were improved, suggesting SPON1 to be a potential AD therapy target (Park et al., 2020). Interacting with APOE, human SPON1 suppresses amyloid- β level through the APP transgene, and has an impact on working memory performance through the activation of the triangular part of the right inferior frontal gyrus (Liu et al., 2018). For NPTX1 and NPTX2, both belong to the family of long neuronal pentraxins. Together with NPTXR, they bind AMPA type glutamate receptors and contribute to multiple forms of developmental and adult synaptic plasticity. Using an AD mouse model, reduction in NPTX2 together with amyloidosis was found to induce a synergistic reduction of inhibitory circuit function. In AD subjects, the level of NTPX2 was found to be related to hippocampal volume, as well as cognitive decline (Xiao et al., 2017). For CST3, cysteine proteases, including cathepsin B (CatB), is a recently discovered amyloid- β -degrading enzyme. Using a mouse model, CST3 was discovered to be a key inhibitor of CatB-induced amyloid- β degradation in vivo. Genetic ablation of CST3 significantly reduced soluble amyloid- β levels, and attenuated associated cognitive deficits and behavioral abnormalities, and restored synaptic plasticity in hippocampus (Sun et al., 2008). For VGF, through a mouse model, over-expression of neuropeptides precursor VGF was found to partially rescue amyloid- β -mediated memory impairment and neuropathology, suggesting a possible causal role of VGF in protecting

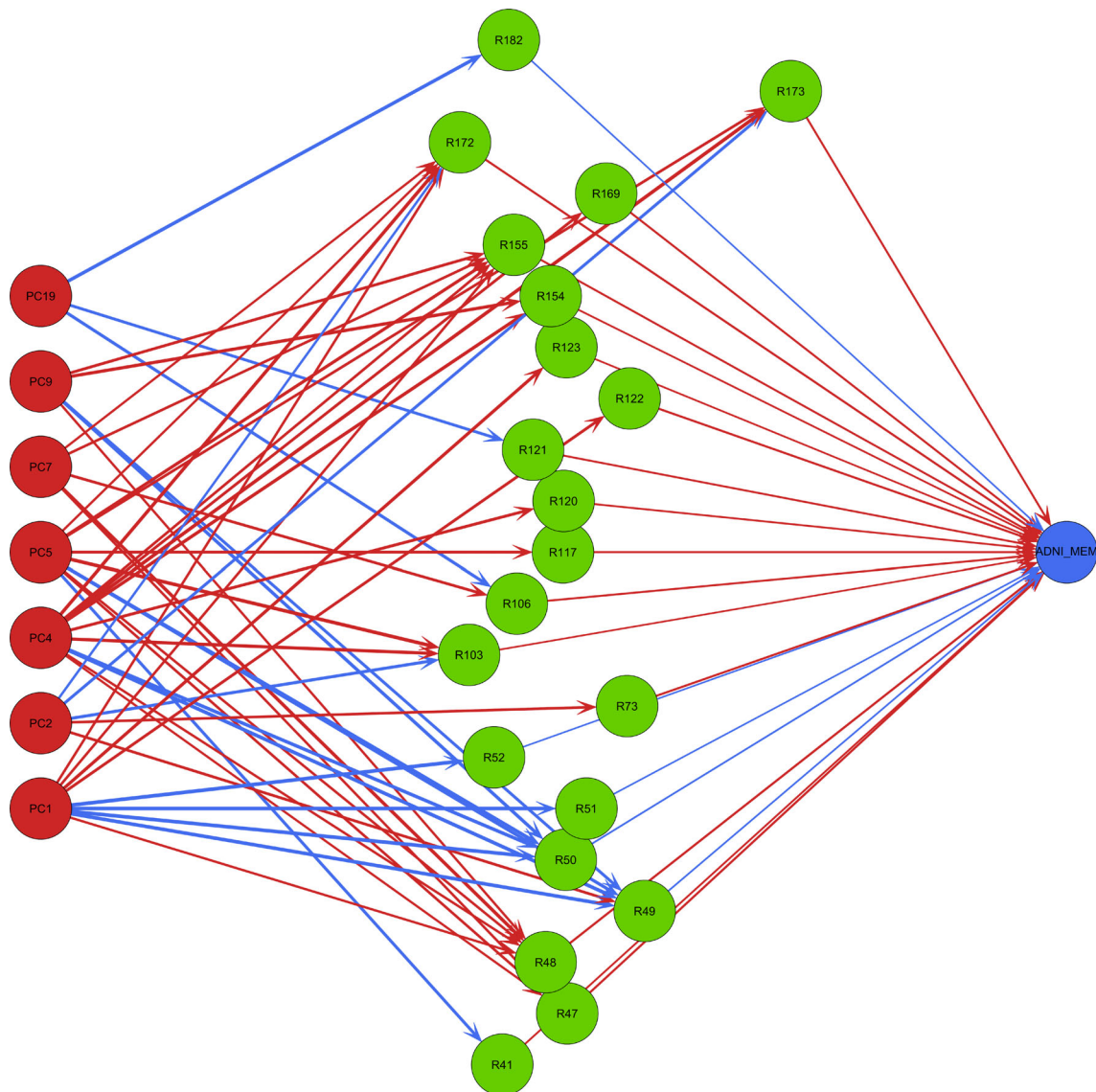


FIGURE 3 The estimated paths for the AD imaging proteomics study. The red nodes denote the principal components of the peptides as exposures, the green nodes the brain regions as mediators, and the blue node the memory score as outcome. The red arrows indicate positive path effects, and the blue arrows negative path effects

against AD pathogenesis and progression (Beckmann et al., 2020). For SORCS1, through a meta-analysis of 16 SORCS1-single nucleotide polymorphisms (SNPs) in six independent datasets, it was found that over-expression of SORCS1 can reduce γ -secretase activity and amyloid- β levels, and the suppression of SORCS1 can increase γ -secretase processing of APP and the levels of amyloid- β (Reitz et al., 2011). For PTGDS, it is one of the most abundant proteins in the CSF, which binds and transports small lipophilic molecules such as amyloid- β , and thus has been considered as the endogenous amyloid- β chaperone (Kanekiyo et al., 2007), and is believed to play an important role in AD development. For CHGA, compared to the normal controls, the level of CHGA was significantly higher in the CSF of patients with MCI, especially with MCI progressing to AD (Duits et al., 2018). CHGA is the major soluble protein in catecholamine storage vesicles, abnormalities of which may play a central role in memory deficits in

AD. Elevation of CHGA was observed in AD brains, and was believed to play a role in amyloid- β pathology (Mattsson et al., 2013; O'Connor, Kailasam, & Thal, 1993). It has also been found that CHGA is negatively associated with hippocampal and entorhinal volume (Khan et al., 2015).

4.2.2 | Proteins related to tau pathology

For IGF2BP2, it is an abundant cerebral insulin-like growth factor signaling protein associated with the AD biomarkers. In both AD mouse models and AD patients, IGF2BP2 was observed to be associated with CSF tau levels and brain atrophy in nonhippocampal regions, suggesting that it is relevant in neurodegeneration through tau pathology (Bonham et al., 2018).

TABLE 1 Brain regions with nonzero indirect effect (IE = $\alpha\beta$) in the AD imaging proteomics study

	Brain regions as mediators		Principal components of peptides as exposures						$\beta (\times 10^{-2})$
			PC1	PC2	PC4	PC5	PC7	PC9	
R41	Left cerebellum white matter	α				-0.17			
		IE ($\times 10^{-3}$)				-1.30			0.76
R47	Right hippocampus	α			0.11	0.13	0.13		
		IE ($\times 10^{-3}$)			1.12	1.60	1.52		1.17
R48	Left hippocampus	α	0.11		0.13	0.13	0.22	0.12	
		IE ($\times 10^{-3}$)	1.34		1.59	1.76	3.40	1.55	1.20
R49	Temporal horn of right lateral ventricle	α	-0.25	0.15	-0.26	-0.18		-0.16	
		IE ($\times 10^{-3}$)	2.06	-1.01	2.03	1.28		1.08	-0.66
R50	Temporal horn of left lateral ventricle	α	-0.29		-0.23	-0.25		-0.21	
		IE ($\times 10^{-3}$)	2.55		1.78	2.05		1.74	-0.71
R51	Right lateral ventricle	α	-0.36						
		IE ($\times 10^{-3}$)	1.06						-0.30
R52	Left lateral ventricle	α	-0.36						
		IE ($\times 10^{-3}$)	1.15						-0.27
R73	Cerebellar vermal lobules VIII-X	α		0.14					
		IE ($\times 10^{-3}$)		1.88					1.00
R103	Left anterior insula	α		-0.17	0.20	0.25			
		IE ($\times 10^{-3}$)		-1.13	1.27	1.76			0.56
R106	Right angular gyrus	α					0.15		-0.18
		IE ($\times 10^{-3}$)					1.03		-1.41
R117	Left entorhinal areas	α				0.17			
		IE ($\times 10^{-3}$)				1.15			0.76
R120	Right frontal pole	α			0.16				
		IE ($\times 10^{-3}$)			1.12				0.75
R121	Left frontal pole	α						-0.16	
		IE ($\times 10^{-3}$)						-1.12	0.77
R122	Right fusiform gyrus	α	0.16						
		IE ($\times 10^{-3}$)	1.32						1.02
R123	Left fusiform gyrus	α	0.19						
		IE ($\times 10^{-3}$)	1.12						0.66
R154	Right middle temporal gyrus	α			0.21			0.19	
		IE ($\times 10^{-3}$)			1.68			1.66	0.74
R155	Left middle temporal gyrus	α	0.13		0.14	0.18	0.14	0.15	
		IE ($\times 10^{-3}$)	1.01		1.09	1.63	1.05	1.35	0.79
R169	Left precuneus	α			0.15				
		IE ($\times 10^{-3}$)			1.10				0.82
R172	Right posterior insula	α	0.13	-0.12	0.22	0.11	0.11		
		IE ($\times 10^{-3}$)	1.44	-1.30	2.67	1.03	1.00		1.03
R173	Left posterior insula	α		-0.17	0.24	0.15			
		IE ($\times 10^{-3}$)		-1.46	2.18	1.09			0.82
R182	Right precentral gyrus	α							-0.24
		IE ($\times 10^{-3}$)							1.91

TABLE 2 The estimated indirect effects (IE), direct effects (DE), and total effects (TE) of the top principal components

	PC1	PC2	PC4	PC5	PC6	PC7	PC9	PC11	PC14	PC15	PC16	PC19	Total
IE	0.013	-0.003	0.018	0.012		0.008	0.007					-0.001	0.054
DE	0.138			0.066	-0.035	0.168	0.065	-0.018	0.102	-0.007	0.156		0.634
TE	0.151	-0.003	0.018	0.078	-0.035	0.176	0.072	-0.018	0.102	-0.007	0.156	-0.001	0.688

Note: The PCs with zero IE and DE are not presented in the table.

TABLE 3 Proteins with top loading magnitude in PC1, PC4, and PC5

Protein	Loading	Gene	Direction	Correlation	
				tau	amyloid
PC1					
Neuroblastoma suppressor of tumorigenicity 1	0.283	NBL1	↑		
Spondin-1	0.160	SPON1	↑	↑	↓
VPS10 domain-containing receptor SorCS1	0.152	SORCS1		↑	↓
ProSAAS	0.116	PCSK1N	⇕		
Prostaglandin-H2 D-isomerase	0.110	PTGDS	↓		↓
Neuronal growth regulator 1	0.110	NEGR1	↓		
Monocyte differentiation antigen CD14	0.109	CD14	↑		
Cell adhesion molecule 3	0.103	CADM3	↓		
PC4					
Beta-2-microglobulin	-0.252	B2M	⇕	↓	
Neuronal pentraxin-2	0.190	NPTX2	↓		↑
Insulin-like growth factor-binding protein 2	-0.147	IGFBP2	⇕	↑	
Neuronal pentraxin-1	0.137	NPTX1	↓		
Kallikrein-6	-0.129	KLK6	↑	↑	↑
Apolipoprotein D	-0.121	APOD	⇕	↑	
Neurexin-2	0.117	NRXN2	⇕		
Cystatin-C	-0.116	CST3	⇕	⇕	↑
PC5					
Superoxide dismutase (Cu-Zn)	0.236	SOD1	↓	↑	↓
Neurosecretory protein VGF	0.195	VGF	↓		↓
Ectonucleotide pyrophosphatase/phosphodiesterase family member 2	-0.152	ENPP2	↑	↓	
Complement C4-A	-0.152	C4A	↑		
Complement factor B	0.121	CFB	↑		
Glial fibrillary acidic protein	-0.120	GFAP	↑		
Mimecan	-0.105	OGN	⇕		
Chromogranin-A	0.103	CHGA	⇕	↑	↑
Alpha-1B-glycoprotein	0.102	A1BG	⇕		

Note: For each protein, direction of protein level in MCI/AD compared to normal control and correlation with CSF tau and amyloid reported in the literature are provided. ↑, consistently upregulated in MCI/AD or positively correlated; ↓, consistently downregulated in MCI/AD or negatively correlated; ⇕, inconsistent reports.

4.2.3 | Proteins related to both amyloid and tau pathology

There was evidence showing that proteins KLK6 and SOD1 were relevant in both amyloid and tau pathology. For SOD1, using an APP-

overexpressing mouse model, SOD1 deficiency was found to accelerate amyloid- β oligomerization, induce tau phosphorylation and lower levels of synaptophysin, and consequently memory impairment (Murakami et al., 2011). Kallikrein-related peptidases (KLKs) represent the largest family of secreted serine proteases. Human KLK6 is the

most abundant KLKs in the spinal cord, brain stem, cerebral cortex including the hippocampus and thalamus. It has been found that KLK6 cleaves APP and mediates cleavage of laminin and collagen, which has implications for APP processing and amyloid- β mediated neurotoxicity (Angelo et al., 2006; Small, Nurcombe, Clarris, Beyreuther, & Masters, 1993). In AD patients, the level of KLK6 in CSF is significantly elevated and is associated with levels of CSF tau suggesting a potential marker of tau pathology (Goldhardt et al., 2019).

4.2.4 | Other AD-related proteins

NRXN2 is another protein marker that was found to be up-regulated among MCI patients, especially with MCI progression to AD (Duits et al., 2018). APOD was found to be elevated in the prefrontal cortex associated with cognitive decline (Thomas et al., 2003). GFAP immunohistochemistry is a marker to assess the oxidative stress and glial cell activation expressed in astrocytes. Focusing on the human entorhinal cortex and hippocampus, the GFAP expression was observed in the hippocampus of AD patients (Hol et al., 2003). B2M is a component of major histocompatibility complex class 1 molecules. Increased soluble B2M has been discovered in the CSF of patients with AD, and was associated with cognitive decline (Carrette et al., 2003). Using mouse models, elevated B2M was observed in the hippocampus of aged mice. Injecting exogenous B2M locally in the hippocampus, impaired hippocampal-dependent cognitive function and neurogenesis were observed in young mice. The findings suggest that the accumulation of B2M increases the risk of age-related cognitive dysfunction and neurogenesis impairment (Smith et al., 2015).

4.2.5 | Proteins related to brain structure/atrophy

NEGR1 is a member of the immunoglobulin superfamily of cell adhesion molecules, and is involved in cortical layering. Using a NEGR1-targeted mouse model, brain morphological analysis revealed NEGR1-related neuroanatomical abnormalities, including enlargement of ventricles and decrease in the volume of the whole brain, corpus callosum, globus pallidus, and hippocampus (Singh et al., 2019). CST3 was discovered to be related to a higher hippocampal atrophy rate (Paterson et al., 2014), and atrophy in the entorhinal cortex (Mattsson et al., 2014). APOD and NPTX2 were found to be related to medial temporal lobe atrophy (Mattsson et al., 2014; Swanson et al., 2016).

4.3 | Brain regions

While Table 1 lists the brain regions with nonzero path effects induced by PC1, PC4, and PC5, Figure 4 visualizes those regions on a template brain. The identified brain regions include the hippocampus, the entorhinal cortex, cortical regions on the temporal, parietal and frontal lobes, the lateral ventricles, and the cerebellum. Brain structural atrophy occurs early in the medial temporal lobe, including the hippocampus and entorhinal cortex, then extends soon after to the

rest of the cortical areas, usually following a temporal, parietal, frontal trajectory, whereas the motor areas are affected toward late stages. (Pini et al., 2016). We next discuss those identified brain regions roughly following this trajectory.

4.3.1 | The hippocampus and entorhinal cortex

The hippocampus is a major component of the human brain located in the medial temporal lobe, and is functionally involved in response inhibition, episodic memory, and spatial cognition. Hippocampal atrophy is the best established and validated biomarker across the entire disease spectrum (Jack Jr et al., 2011). The entorhinal cortex also locates in the medial temporal lobe. It connects the neocortex and the hippocampus that receives information from the neocortex and projects to the hippocampus through the perforant pathway (Insausti, Tunon, Sobreviela, Insausti, & Gonzalo, 1995). It has been consistently reported that, compared to the healthy controls, entorhinal atrophy was observed in the MCI patients, and more severe atrophy in the AD patients (Pini et al., 2016). The hippocampus and entorhinal cortex, as well as the anatomically related parahippocampal and perirhinal cortices, are parts of the medial temporal lobe memory system. Impairments of this system are responsible for the deficit in episodic memory, and are early hallmark of AD (Nadel & Hardt, 2011).

4.3.2 | The lateral temporal, parietal, and frontal cortex

The gray matter loss in the lateral temporal cortex, dorsal parietal, parietal angular and frontal cortex occurs during the progression from incipient to mild AD. During this period, cognitive deficits have been observed in both memory and nonmemory domains, including language, visuo-spatial and executive function (Frisoni, Prestia, Rasser, Bonetti, & Thompson, 2009). Moreover, a higher amount of tau deposition has been observed in the middle temporal cortex, fusiform gyrus, and entorhinal cortex (Schultz et al., 2018). The fusiform gyrus is critical in facial recognition. Alterations of gene expression specific to the fusiform gyrus were discovered in AD patients (Ma et al., 2020). The left middle temporal gyrus is related to the recognition of known faces and accessing word meaning while reading (Acheson & Hagoort, 2013). The precuneus, a hub of the default mode network, has been found to be related to episodic memories (Sadigh-Eteghad, Majdi, Farhoudi, Talebi, & Mahmoudi, 2014). Atrophy in the entorhinal cortex, fusiform, middle temporal gyrus, precuneus, and precentral has been noted in AD (Parker et al., 2018). The association between atrophy in the insular cortex and memory deficits in AD has been reported too (Lin et al., 2017).

4.3.3 | The lateral ventricles

The ventricles are one of the interests in brain atrophy research as the volumetric measurement is robust to automatic segmentation due to the sharp contrast between the CSF in the ventricles and surrounding

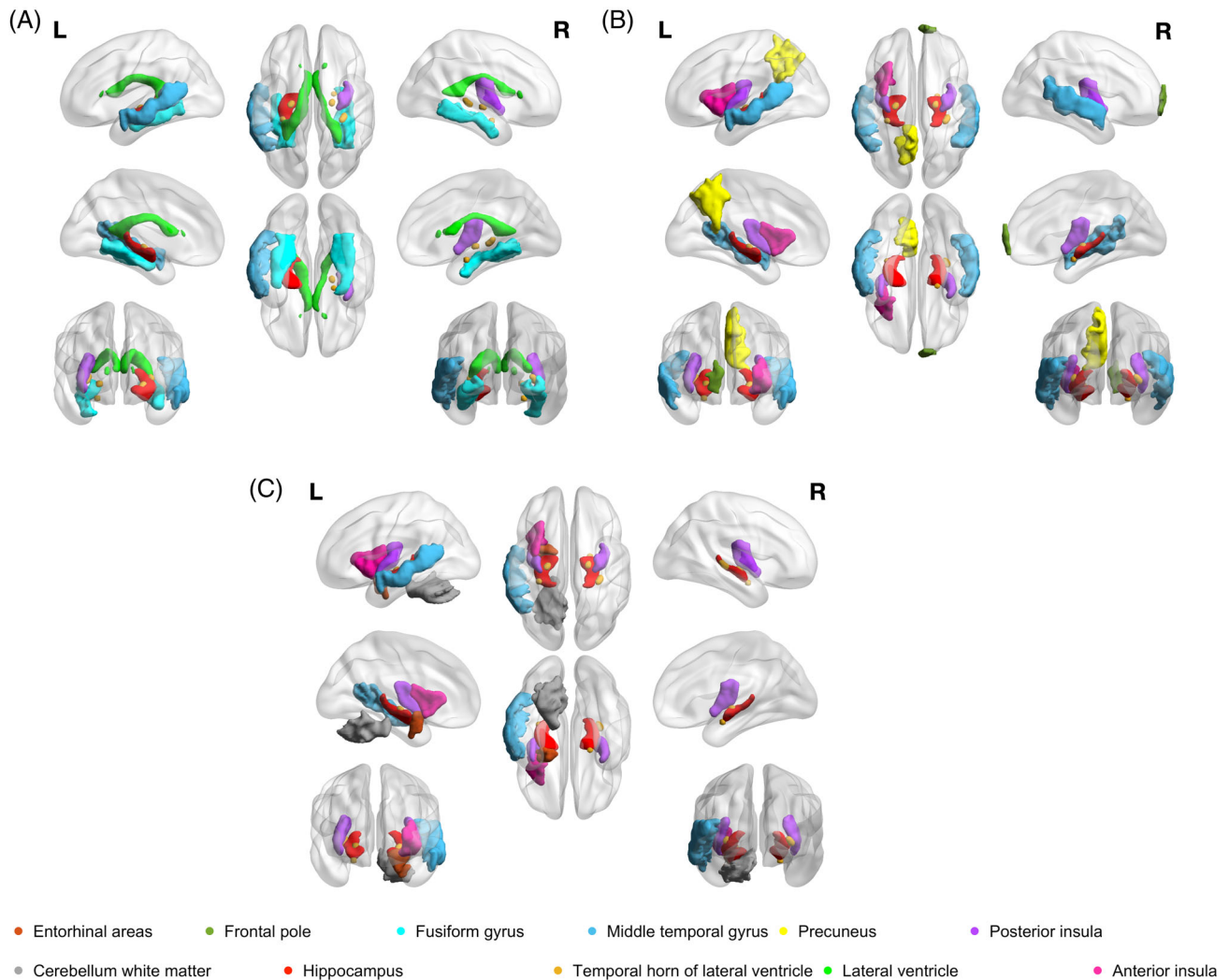


FIGURE 4 Brain regions with a nonzero mediation effect in (a) PC1, (b) PC4, and (c) PC5

tissue in T1-weighted images. Thus, as a complement metric of hemispheric atrophy rates, enlargement in the lateral ventricles is an important marker of AD progression (Kruthika et al., 2019).

4.3.4 | The cerebellum

The cerebellum is involved in cognition and emotion and communicates with cerebral cortices in a topographically organized manner. Based on existing evidence of cerebellar modulation of cognition and emotion, it was hypothesized that there exists cerebellar contribution to the cognitive and neuropsychiatric deficits in AD. However, more research is required to validate the hypothesis and to understand cerebrotocerebellar interactions in AD pathology (Jacobs et al., 2018).

5 | SIMULATION STUDY

We complement our data analysis with some additional simulation studies to further examine the empirical performance of the proposed method.

We generate $\mathbf{X}_i \in \mathbb{R}^f$ ($i = 1, \dots, n$) from a multivariate normal distribution with mean zero and a covariance matrix whose eigenvalues exponentially decay. After applying PCA, we obtain $\tilde{\mathbf{X}} \in \mathbb{R}^{n \times q}$, where q is chosen such that the top q PCs account for over 80% of total data variation. We then generate \mathbf{M} and \mathbf{Y} following Model 1 given $\tilde{\mathbf{X}}$. We set 5% of the path effects to be nonzero. We consider two sets of data dimension, $r = 100$, $p = 100$, and $r = 350$, $p = 150$, the latter of which has a similar data dimension as in the ADNI dataset. We also consider three sample sizes, $n = 100, 500, 1000$. We compare the proposed approach with an approach based on the univariate mediation analysis (Imai, Keele, & Tingley, 2010). After the PCs are obtained, univariate mediation analysis is performed for each mediator and each exposure PC and finish with a p -value correction (Benjamini & Hochberg, 1995).

Table 4 presents the estimated total indirect effects and the indirect effects of the top six PCs, and Table 5 presents the estimated number of PCs and the sensitivity and specificity of the identified nonzero path effects. Among all cases, the estimated number of PCs is 6, which agrees with the truth. From the tables, we observe that the proposed method achieves a competitive performance, and the performance improves, with a lower estimation error and a higher

TABLE 4 The estimation bias and mean squared error (MSE) of estimating the total indirect effect and indirect effect of top PCs in the simulation study

r	p	Truth	n = 100						n = 500						n = 1000					
			UniMed			PathLasso			UniMed			PathLasso			UniMed			PathLasso		
			Bias	MSE		Bias	MSE		Bias	MSE		Bias	MSE		Bias	MSE		Bias	MSE	
100	100	Total	29.203	8829.276	9.030	128.080	12.375	14868.680	-1.827	16.593	9.937	19518.110	-0.921	13.508						
		PC1	9.219	1740.035	7.172	61.226	9.754	3045.273	0.033	1.843	10.971	3015.491	0.035	2.883						
		PC2	7.474	4086.099	-0.883	15.244	-3.086	11896.018	-1.410	9.968	-5.698	15284.011	-0.856	4.364						
		PC3	10.254	1249.422	3.205	20.100	6.314	1609.520	-0.168	2.531	4.989	1764.686	-0.032	2.248						
		PC4	0.900	410.946	-0.477	11.943	-0.830	88.795	-0.272	4.242	-0.270	44.037	-0.067	1.963						
		PC5	1.043	229.951	-0.049	5.168	-0.084	40635	-0.078	1.328	-0.077	18.711	0.002	0.611						
		PC6	0.369	185.175	0.085	2.809	0.139	36.633	0.054	0.794	0.024	17.296	0.019	0.370						
350	150	Total	7.246	24906.450	-6.317	80.520	12.668	50394.860	-1.164	37.593	39.456	84070.740	-1.284	19.271						
		PC1	17.762	9155.026	6.461	54.644	29.582	28896.549	0.723	13.823	24.086	39594.888	-0.511	3.616						
		PC2	-7.415	5417.007	-9.528	102.378	-5.064	13073.687	-0.933	20.122	-1.226	14461.470	0.816	3.045						
		PC3	5.364	4131.577	-3.446	27.204	-10.482	11714.378	-1.398	10.500	17.494	24773.663	-1.740	7.489						
		PC4	-4.088	2540.851	-0.093	8.015	1.605	485.008	0.147	2.595	-0.202	230.990	-0.033	1.403						
		PC5	3.947	2552.043	0.153	5.445	-0.355	627.776	0.060	1.615	-0.271	237.362	0.066	0.781						
		PC6	-8.322	4271.384	0.137	6.489	-2.618	610.476	0.238	1.054	-0.424	387.267	0.117	0.750						

Note: UniMed is an approach based on univariate mediation analysis. PathLasso is the proposed approach.

TABLE 5 The estimated number of PC (and the standard error, SE) in the PCA step and sensitivity and specificity of identifying paths with a nonzero path effect in the simulation study

<i>r</i>	<i>p</i>	<i>n</i>	# PC (SE)	UniMed		PathLasso	
				Sensitivity	Specificity	Sensitivity	Specificity
100	100	100	6.03 (0.17)	0.55	0.98	0.84	0.53
		500	5.21 (0.41)	0.78	0.97	1.00	0.89
		1000	6.26 (0.44)	0.86	0.97	1.00	0.91
350	150	100	5.99 (0.10)	0.62	0.97	0.80	0.57
		500	6.00 (0.00)	0.85	0.96	1.00	0.89
		1000	6.00 (0.00)	0.87	0.95	1.00	0.91

Note: UniMed is an approach based on univariate mediation analysis. PathLasso is the proposed approach.

selection accuracy, as the sample size increases. For the univariate-based approach (UniMed), the performance in estimating the effects does not improve as the sample size increases and the power of identifying nonzero mediation effects is much lower.

6 | DISCUSSION

In this study, we propose a mediation framework with high-dimensional exposures and high-dimensional mediators. The framework integrates the PCA with marginal linear SEMs, where the PCA leads to multiple independent exposures and the marginal SEMs allow the mediators to be dependent. A regularization combining the Group Lasso and the Pathway Lasso is considered to achieve simultaneous exposure and mediator selection. Through simulation studies, the proposed approach yields competitive estimation performance and selection accuracy. The proposed framework is applied to integrate the CSF proteomics data, the brain volumetric data, and a memory measurement acquired from MCI subjects in ADNI. Several protein–imaging–memory pathways are identified, which are in accordance with existing knowledge about AD.

The proposed framework is among the first attempts to conduct mediation analysis where both the exposure and mediator are high dimensional. It also fits in the context of integrating multiview data. In this study, pathology of deficits in memory among MCI patients induced by CSF protein deposition and mediated by brain atrophy is articulated. Integrating proteomics with neuroimaging data on a large scale is not commonly seen in the existing literature. One can apply the proposed approach to integrate other types of data under mechanistic and causal assumptions (Data S1, Supporting Information). For example, in an imaging-genetics study, the genetic/genomic data are the exposures and the neuroimaging data are the mediators. Another example is to integrate multimodal neuroimaging data with the structural imaging data as the exposures and the functional imaging data as the mediators based on Hebb's law (Hebb, 2005). Another direction of application is in a longitudinal study, where imaging (or omics) data collected at two (consecutive) time points can be considered as the exposures and mediators, respectively, and a phenotyping measurement at the end of study is the outcome. The temporal ordering in the measurements intrinsically infers the causality.

In order to account for the dependence between the exposures, the PCA is employed. It not only rotates the exposures into independent components, but also significantly reduces the data dimension. However, one drawback of PCA is that the loadings are not sign identifiable. Thus, both the estimated indirect and direct effect are sign sensitive. In the analysis, we keep the highest loading in each component to be positive.

The current study focuses on selecting exposures and their induced mediation pathways and estimating the indirect/direct effects. Post-selection inference is also an important question. We leave the study of drawing statistical inference to future research. In the current study, PCA is considered as an initial step to decorrelate the exposures. A next step will be merging this decomposition into the mediation optimization.

ACKNOWLEDGMENTS

Yi Zhao was partially supported by NIH grants P30AG072976 and U54AG065181. Lexin Li was partially supported by NSF grant CIF-2102227, and NIH grants R01AG061303, R01AG062542, and R01AG034570.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database at adni.loni.usc.edu.

ORCID

Yi Zhao  <https://orcid.org/0000-0003-4766-5934>

REFERENCES

- Acheson, D. J., & Hagoort, P. (2013). Stimulating the brain's language network: Syntactic ambiguity resolution after TMS to the inferior frontal gyrus and middle temporal gyrus. *Journal of Cognitive Neuroscience*, 25(10), 1664–1677.
- Alzheimer's Association. (2020). 2020 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 16(3), 391–460.
- Angelo, P. F., Lima, A. R., Alves, F. M., Blaber, S. I., Scarisbrick, I. A., Blaber, M., ... Juliano, M. A. (2006). Substrate specificity of human Kallikrein 6 salt and glycosaminoglycan activation effects. *Journal of Biological Chemistry*, 281(6), 3116–3126.
- Aung, M. T., Song, Y., Ferguson, K. K., Cantonwine, D. E., Zeng, L., McElrath, T. F., ... Mukherjee, B. (2020). Application of an analytical

- framework for multivariate mediation analysis of environmental data. *Nature Communications*, 11(1), 5624.
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), 1173–1182.
- Beckmann, N. D., Lin, W.-J., Wang, M., Cohain, A. T., Charney, A. W., Wang, P., ... Schadt, E. E. (2020). Multiscale causal networks identify VGF as a key regulator of Alzheimer's disease. *Nature Communications*, 11(1), 3942.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300.
- Bonham, L. W., Geier, E. G., Steele, N. Z., Holland, D., Miller, B. L., Dale, A. M., ... Alzheimer's Disease Neuroimaging Initiative. (2018). Insulin-like growth factor binding protein 2 is associated with biomarkers of Alzheimer's disease pathology and shows differential expression in transgenic mice. *Frontiers in Neuroscience*, 12, 476.
- Cai, Q., Wang, H., Li, Z., & Liu, X. (2019). A survey on multimodal data-driven smart healthcare systems: Approaches and applications. *IEEE Access*, 7, 133583–133599.
- Carrette, O., Demalte, I., Scherl, A., Yalkinoglu, O., Corthals, G., Burkhard, P., ... Sanchez, J.-C. (2003). A panel of cerebrospinal uid potential biomarkers for the diagnosis of Alzheimer's disease. *PROTEOMICS: International Edition*, 3(8), 1486–1494.
- Chén, O. Y., Crainiceanu, C., Ogburn, E. L., Caffo, B. S., Wager, T. D., & Lindquist, M. A. (2017). High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics*, 19(2), 121–136.
- Doshi, J., Erus, G., Ou, Y., Resnick, S. M., Gur, R. C., Gur, R. E., ... Alzheimer's Disease Neuroimaging Initiative. (2016). MUSE: Multi-atlas region segmentation utilizing ensembles of registration algorithms and parameters, and locally optimal atlas selection. *NeuroImage*, 127, 186–195.
- Duits, F. H., Brinkmalm, G., Teunissen, C. E., Brinkmalm, A., Scheltens, P., Van der Flier, W. M., ... Blennow, K. (2018). Synaptic proteins in CSF as potential novel biomarkers for prognosis in prodromal Alzheimer's disease. *Alzheimer's Research & Therapy*, 10(1), 1–9.
- Frisoni, G. B., Prestia, A., Rasser, P. E., Bonetti, M., & Thompson, P. M. (2009). In vivo mapping of incremental cortical atrophy from incipient to overt Alzheimer's disease. *Journal of Neurology*, 256(6), 916–924.
- Goldhardt, O., Warnhoff, I., Yakushev, I., Begcevic, I., Förstl, H., Magdolen, V., ... Grimmer, T. (2019). Kallikrein-related peptidases 6 and 10 are elevated in cerebrospinal uid of patients with Alzheimer's disease and associated with CSF-TAU and FDG-PET. *Translational Neurodegeneration*, 8(1), 1–13.
- Hebb, D. O. (2005). *The organization of behavior: A neuropsychological theory*. East Sussex, England: Psychology Press.
- Higgins, I. A., Kundu, S., & Guo, Y. (2018). Integrative Bayesian analysis of brain functional networks incorporating anatomical knowledge. *NeuroImage*, 181, 263–278.
- Hol, E., Roelofs, R., Moraal, E., Sonnemans, M., Sluijs, J., Proper, E., ... Van Leeuwen, F. (2003). Neuronal expression of GFAP in patients with Alzheimer pathology and identification of novel GFAP splice forms. *Molecular Psychiatry*, 8(9), 786–796.
- Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4), 309–334.
- Insausti, R., Tunon, T., Sobreviela, T., Insausti, A., & Gonzalo, L. (1995). The human entorhinal cortex: A cytoarchitectonic analysis. *Journal of Comparative Neurology*, 355(2), 171–198.
- Jack, C. R., Jr., Barkhof, F., Bernstein, M. A., Cantillon, M., Cole, P. E., DeCarli, C., ... Foster, N. L. (2011). Steps to standardization and validation of hippocampal volumetry as a biomarker in clinical trials and diagnostic criterion for Alzheimer's disease. *Alzheimer's & Dementia*, 7(4), 474–485.
- Jacobs, H. I., Hopkins, D. A., Mayrhofer, H. C., Bruner, E., van Leeuwen, F. W., Raaijmakers, W., & Schmahmann, J. D. (2018). The cerebellum in Alzheimer's disease: Evaluating its role in cognitive decline. *Brain*, 141(1), 37–47.
- Jagust, W. (2018). Imaging the evolution and pathophysiology of Alzheimer disease. *Nature Reviews. Neuroscience*, 19, 687–700.
- Kanekiyo, T., Ban, T., Aritake, K., Huang, Z.-L., Qu, W.-M., Okazaki, I., ... Urade, Y. (2007). Lipocalin-type prostaglandin D synthase/ β -trace is a major amyloid β -chaperone in human cerebrospinal fluid. *Proceedings of the National Academy of Sciences of the United States of America*, 104(15), 6412–6417.
- Khan, W., Aguilar, C., Kiddle, S. J., Doyle, O., Thambisetty, M., Muehlboeck, S., ... Alzheimer's Disease Neuroimaging Initiative. (2015). A subset of cerebrospinal uid proteins from a multi-analyte panel associated with brain atrophy, disease classification and prediction in Alzheimer's disease. *PLoS One*, 10(8), e0134368.
- Kruthika, K., Maheshappa, H., & Alzheimer's Disease Neuroimaging Initiative. (2019). Multistage classifier-based approach for Alzheimer's disease prediction and retrieval. *Informatics in Medicine Unlocked*, 14, 34–42.
- Lin, F., Ren, P., Lo, R. Y., Chapman, B. P., Jacobs, A., Baran, T. M., ... Foxe, J. J. (2017). Insula and inferior frontal gyrus' activities protect memory performance against Alzheimer's disease pathology in old age. *Journal of Alzheimer's Disease*, 55(2), 669–678.
- Liu, S., Cai, W., Liu, S., Zhang, F., Fulham, M., Feng, D., ... Kikinis, R. (2015). Multimodal neuroimaging computing: A review of the applications in neuropsychiatric disorders. *Brain Informatics*, 2(3), 167–180.
- Liu, Z., Dai, X., Tao, W., Liu, H., Li, H., Yang, C., ... Zhang, Z. (2018). APOE influences working memory in non-demented elderly through an interaction with SPON1 rs2618516. *Human Brain Mapping*, 39(7), 2859–2867.
- Long, J. P., Irajzad, E., Doecke, J. D., Do, K.-A., & Ha, M. J. (2020). A framework for mediation analysis with multiple exposures, multivariate mediators, and non-linear response models. arXiv preprint: arXiv:2011.06061.
- Ma, D., Fetahu, I. S., Wang, M., Fang, R., Li, J., Liu, H., ... Shi, Y. G. (2020). The fusiform gyrus exhibits an epigenetic signature for Alzheimer's disease. *Clinical Epigenetics*, 12(1), 1–16.
- Mattsson, N., Insel, P., Nosheny, R., Trojanowski, J. Q., Shaw, L. M., Jack, C. R., Jr., ... Alzheimer's Disease Neuroimaging Initiative. (2014). Effects of cerebrospinal uid proteins on brain atrophy rates in cognitively healthy older adults. *Neurobiology of Aging*, 35(3), 614–622.
- Mattsson, N., Insel, P., Nosheny, R., Zetterberg, H., Trojanowski, J., Shaw, L., ... Weiner, M. (2013). CSF protein biomarkers predicting longitudinal reduction of CSF β -amyloid42 in cognitively healthy elders. *Translational Psychiatry*, 3(8), e293–e293.
- Mormino, E. C., Kluth, J. T., Madison, C. M., Rabinovici, G. D., Baker, S. L., Miller, B. L., ... Alzheimer's Disease Neuroimaging Initiative*. (2009). Episodic memory loss is related to hippocampal-mediated β -amyloid deposition in elderly subjects. *Brain*, 132(5), 1310–1323.
- Murakami, K., Murata, N., Noda, Y., Tahara, S., Kaneko, T., Kinoshita, N., ... Shimizu, T. (2011). SOD1 (copper/zinc superoxide dismutase) deficiency drives amyloid β protein oligomerization and memory loss in mouse model of Alzheimer disease. *Journal of Biological Chemistry*, 286(52), 44557–44568.
- Nadel, L., & Hardt, O. (2011). Update on memory systems and processes. *Neuropsychopharmacology*, 36(1), 251–273.
- Nathoo, F. S., Kong, L., Zhu, H., & Alzheimer's Disease Neuroimaging Initiative. (2019). A review of statistical methods in imaging genetics. *Canadian Journal of Statistics*, 47(1), 108–131.
- O'Connor, D. T., Kailasam, M. T., & Thal, L. J. (1993). Cerebrospinal uid chromogranin A is unchanged in Alzheimer dementia. *Neurobiology of Aging*, 14(3), 267–269.
- Park, S. Y., Kang, J. Y., Lee, T., Nam, D., Jeon, C.-J., & Kim, J. B. (2020). SPON1 can reduce amyloid beta and reverse cognitive impairment

- and memory dysfunction in Alzheimer's disease mouse model. *Cell*, 9(5), 1275.
- Parker, T. D., Slattery, C. F., Zhang, J., Nicholas, J. M., Paterson, R. W., Foulkes, A. J., ... Schott, J. M. (2018). Cortical microstructure in young onset Alzheimer's disease using neurite orientation dispersion and density imaging. *Human Brain Mapping*, 39(7), 3005–3017.
- Paterson, R., Bartlett, J., Blennow, K., Fox, N., Shaw, L., Trojanowski, J., ... Schott, J. M. (2014). Cerebrospinal uid markers including trefoil factor 3 are associated with neurodegeneration in amyloid-positive individuals. *Translational Psychiatry*, 4(7), e419.
- Pini, L., Pievani, M., Bocchetta, M., Altomare, D., Bosco, P., Cavedo, E., ... Frisoni, G. B. (2016). Brain atrophy in Alzheimer's disease and aging. *Ageing Research Reviews*, 30, 25–48.
- Reitz, C., Tokuhiro, S., Clark, L. N., Conrad, C., Vonsattel, J.-P., Hazrati, L.-N., ... Mayeux, R. (2011). SORCS1 alters amyloid precursor protein processing and variants may increase Alzheimer's disease risk. *Annals of Neurology*, 69(1), 47–64.
- Richardson, S., Tseng, G. C., & Sun, W. (2016). Statistical methods in integrative genomics. *Annual Review of Statistics and its Application*, 3, 181–209.
- Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3), 286–327.
- Sadigh-Eteghad, S., Majdi, A., Farhoudi, M., Talebi, M., & Mahmoudi, J. (2014). Different patterns of brain activation in normal aging and Alzheimer's disease from cognitive sight: Meta analysis using activation likelihood estimation. *Journal of the Neurological Sciences*, 343(1–2), 159–166.
- Schultz, S. A., Gordon, B. A., Mishra, S., Su, Y., Perrin, R. J., Cairns, N. J., ... Benzinger, T. L. (2018). Widespread distribution of tauopathy in pre-clinical Alzheimer's disease. *Neurobiology of Aging*, 72, 177–185.
- Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2), 231–245.
- Singh, K., Jayaram, M., Kaare, M., Leidmaa, E., Jagomäe, T., Heinla, I., ... Vasar, E. (2019). Neural cell adhesion molecule Negr1 deficiency in mouse results in structural brain endophenotypes and behavioral deviations related to psychiatric disorders. *Scientific Reports*, 9(1), 1–15.
- Small, D. H., Nurcombe, V., Clarris, H., Beyreuther, K., & Masters, C. L. (1993). The role of extracellular matrix in the processing of the amyloid protein precursor of Alzheimer's disease. *Annals of the New York Academy of Sciences*, 695(1), 169–174.
- Smith, L. K., He, Y., Park, J.-S., Bieri, G., Snelthage, C. E., Lin, K., ... Villeda, S. A. (2015). β 2-microglobulin is a systemic pro-aging factor that impairs cognitive function and neurogenesis. *Nature Medicine*, 21(8), 932–937.
- Song, Y., Zhou, X., Zhang, M., Zhao, W., Liu, Y., Kardia, S. L., ... Mukherjee, B. (2018). Bayesian shrinkage estimation of high dimensional causal mediation effects in omics studies. *Biometrics*, 76, 700–710.
- Sun, B., Zhou, Y., Halabisky, B., Lo, I., Cho, S.-H., Mueller-Stieber, S., ... Gan, L. (2008). Cystatin C-cathepsin B axis regulates amyloid beta levels and associated neuronal deficits in an animal model of Alzheimer's disease. *Neuron*, 60(2), 247–257.
- Swanson, A., Willette, A., & Alzheimer's Disease Neuroimaging Initiative. (2016). Neuronal pentraxin 2 predicts medial temporal atrophy and memory decline across the Alzheimer's disease spectrum. *Brain, Behavior, and Immunity*, 58, 201–208.
- Thomas, E. A., Laws, S. M., Sutcliffe, J. G., Harper, C., Dean, B., McClean, C., ... Martins, R. N. (2003). Apolipoprotein D levels are elevated in prefrontal cortex of subjects with Alzheimer's disease: No relation to apolipoprotein E expression or genotype. *Biological Psychiatry*, 54(2), 136–141.
- Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1), 91–108.
- VanderWeele, T. J. (2016). Mediation analysis: A practitioner's guide. *Annual Review of Public Health*, 37(1), 17–32.
- Vassar, R., Bennett, B. D., Babu-Khan, S., Kahn, S., Mendiaz, E. A., Denis, P., ... Citron, M. (1999). β -secretase cleavage of Alzheimer's amyloid precursor protein by the transmembrane aspartic protease BACE. *Science*, 286(5440), 735–741.
- Wesenhagen, K. E., Teunissen, C. E., Visser, P. J., & Tijms, B. M. (2020). Cerebrospinal uid proteomics and biological heterogeneity in Alzheimer's disease: A literature review. *Critical Reviews in Clinical Laboratory Sciences*, 57(2), 86–98.
- Xiao, M.-F., Xu, D., Craig, M. T., Pelkey, K. A., Chien, C.-C., Shi, Y., ... Worley, P. F. (2017). NPTX2 and cognitive dysfunction in Alzheimer's disease. *eLife*, 6, e23798.
- Zhang, Q. (2021). High-dimensional mediation analysis with applications to causal gene identification. *Statistics in Biosciences*, 1–20.
- Zhao, Y., Li, L., & Caffo, B. S. (2021). Multimodal neuroimaging data integration and pathway analysis. *Biometrics*, 77(3), 879–889.
- Zhao, Y., & Luo, X. (2022). Pathway lasso: Pathway estimation and selection with high dimensional mediators. *Statistics and Its Interface*, 15(1), 39–50.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.
- Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2), 265–286.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Zhao, Y., Li, L., & Alzheimer's Disease Neuroimaging Initiative (2022). Multimodal data integration via mediation analysis with high-dimensional exposures and mediators. *Human Brain Mapping*, 43(8), 2519–2533. <https://doi.org/10.1002/hbm.25800>