



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Disease Progression Prediction Incorporating Genotype-Environment Interactions: A Longitudinal Neurodegenerative Disorder Study

Jin Zhang, Muheng Shang, Yan Yang, Lei Guo, Junwei Han, Lei Du*, and Alzheimer's Disease Neuroimaging Initiative†

School of Automation, Northwestern Polytechnical University, Xi'an 710072, China

Abstract. Disease progression prediction is a fundamental yet challenging task in neurodegenerative disorders. Despite extensive research endeavors, disease progression fitting on brain imaging data alone may yield suboptimal performance due to the effect of potential interactions between genetic variations, proteomic expressions and environmental exposures on the disease progression. To fill this gap, we draw on the idea of mutual-assistance (MA) learning and accordingly propose a fresh and powerful scheme, referred to as Mutual-Assistance Disease Progression fitting and Genotype-by-Environment interaction identification approach (MA-DPxGE). Specifically, our model jointly performs disease progression fitting using longitudinal imaging phenotypes and identification of genotype-by-environment interaction factors. To ensure stability and interpretability, we employ innovative penalties to discern significant risk factors. Moreover, we meticulously design adaptive mechanisms for loss-term reweighting, ensuring fair adjustments for each prediction task. Furthermore, due to high-dimensional genotype-by-environment interactions, we devise a rapid and efficient strategy to reduce runtime, ensuring practical availability and applicability. Experimental results on Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset reveal that MA-DPxGE demonstrates superior performance compared to state-of-the-art approaches, while maintaining exceptional interpretability. This outcome is pivotal in elucidating disease progression patterns and establishing effective strategies to mitigate or halt disease advancement.

Keywords: Imaging genetics · Disease progression · GE interactions.

*Correspondence to dulei@nwpu.edu.cn

†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 listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

1 Introduction

Alzheimer’s disease (AD) is a neurodegenerative disorder that arises from a complex interactions of genetic and environmental factors [14,20,22,7,6]. Understanding the genotype-by-environment interactions during disease progression, especially in the predementia stages, is pivotal in elucidating disease spreading patterns and devising effective strategies to mitigate or halt disease advancement.

Brain imaging genetics has become a powerful tool for uncovering the genetic underpinnings of intermediate brain phenotypes [13,3,21]. In the past decade, several approaches, such as univariate, multivariate and bi-multivariate methods, have been introduced to explore the complex interplay between genetic variations and endophenotypes [16,15,10,17,8]. These cross-sectional methods may have limitations in identifying risk factors, possibly due to the oversight of longitudinal imaging quantitative traits (QTs) that can capture the progression of neurodegenerative disorders [10]. Meanwhile, evidence suggests that longitudinal phenotypes are influenced by a combination of GE (Genotype-Environment) interactions, thus, understanding the interactions along dynamically changing trajectories can significantly contribute to disease progression prediction and prompt explain heritability of AD [19,9]. However, most studies to date have primarily focused on identifying genetic main effects, often neglecting potential GE interactions. Importantly, these non-additive effects could contribute to intermediate imaging phenotypes, finally leading to disease occurrence, but investigating GE interactions on longitudinal neuroimaging phenotypes poses statistical challenges due to high dimensionality of genetic variations [11,2].

To address the aforementioned challenges, we propose a simple yet versatile framework, MA-DPxGE. This is the first framework to achieve disease progression fitting with built-in interpretability. In particular, leveraging mutual-assistance learning, MA-DPxGE incorporates a feature interaction module and sparsity regularized modules into disease progression prediction framework, which could enhance interpretability while preserving predictability. First, disease progression fitting is employed to model disease advancement using longitudinal imaging phenotypes. The fitted disease progression is then correlated with genetic variations, environmental factors and their interactions. Second, to identify biologically risk factors, we introduce an interpretable regularized module that selects important features while eliminating inefficacious ones. Additionally, we separate the effects of normal aging from the disease progression, aiding in recovering the true relationship between GE interactions and disease-related phenotypes. Considering the computational complexity of high-dimensional GE interactions, we develop a fast optimization algorithm in spirit of divide-and-conquer. Experimentals on real neuroimaging genetic data demonstrate superior correlation coefficients and prediction performance compared to state-of-the-art methods. This outcome holds great promise for advancing our understanding of AD progression prediction and facilitating targeted therapeutic interventions.

The main contributions can be summarized as follows:

- Drawing on the idea of mutual-assistance (MA), we propose a robust yet practical disease progression prediction incorporating GE interactions, i.e, MA-DPxGE, which benefits disease progression fitting and biomarker identification.
- We introduce feature-interaction module to identify biologically meaningful genetic variations, environment as well as their GE interactions.
- We further extend our proposed approach to the chromosome-wide setting, which significantly enhances the practical applicability of our method.
- We separate the normal aging effect from the disease trajectory. This separation ensures more accurate identification of risk factors. The results demonstrate that MA-DPxGE attains superior performance in terms of biomarker identification and disease progression prediction, establishing a new state-of-the-art.

2 Method

2.1 Dataset

We obtain the SNPs and endophenotypes, including longitudinal imaging data (Voxel-based morphometry (VBM) and FreeSurfer) of AD-related areas (left and right Hippocampus, Parahippocampal, MidTemporal at baseline, Month 6, 12, and 24), proteomics (146 proteomic markers), environmental factors (16 risk factors such as age, visual and auditory impairment, body mass index, alcohol abuse, drug sensitivities, blood pressure, current smoking status, education, gender, stroke, etc.), and genotyping data (10000 SNPs) from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu)[15]. The dataset consists 276 non-Hispanic Caucasian participants, including 39 healthy controls (HCs), 182 individuals with mild cognitive impairments (MCIs), and 55 individuals with AD. Our aim is to build the disease progression and the related genetic variations, proteomic markers, environmental factors, as well as their interactions underpinning the progression.

2.2 The MA-DPxGE

Description As mentioned above, we draw on the idea of mutual-assistance learning and accordingly and integrate disease progression and biomarker identification tasks into a whole, which could benefit both tasks mutually.

For ease of presentation, \mathbf{y}_{ij} denotes the longitudinal imaging phenotype of the i -th subject at time-point j . The SNPs are denoted as $\mathbf{X} \in \mathbb{R}^{n \times p}$, proteins and environmental exposures are denoted as $\mathbf{Z} \in \mathbb{R}^{n \times q}$ and $\mathbf{E} \in \mathbb{R}^{n \times r}$. n is the number of subjects, m is the times of time points. p , q and r represent the number of SNPs, proteomics biomarkers and environmental exposures respectively. MA-DPxGE constructs a disease progression model and incorporates the genetic risk factors, proteomics markers, and environmental exposures by considering both the baseline status (intercept) and the changing rate (slope) as responses.

Therefore, the formulation of MA-DPxGE is as follows:

$$\begin{aligned} \min_{\mathbf{U}, \mathbf{V}, \mathbf{Q}} \sum_{i=1}^n \sum_{j=1}^m & \|\mathbf{a} + \mathbf{x}_i \mathbf{u}_1 + \mathbf{x}_i \mathbf{Q}_1 \mathbf{e}_i^T + \mathbf{z}_i \mathbf{v}_1 + (\mathbf{b} + \mathbf{x}_i \mathbf{u}_2 + \mathbf{x}_i \mathbf{Q}_2 \mathbf{e}_i^T + \mathbf{z}_i \mathbf{v}_2) t_j - y_{ij}\|_2^2 \\ & + \eta \sum_{i=1}^n \sum_{f=1}^2 \|\mathbf{x}_i \mathbf{u}_f - \mathbf{z}_i \mathbf{v}_f\|_2^2 + \Omega(\mathbf{U}) + \Omega(\mathbf{V}) + \Omega(\mathbf{Q}) \end{aligned} \quad (1)$$

$\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2] \in \mathcal{R}^{p \times 2}$ is the weight carrying the main effects of SNPs, $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2] \in \mathcal{R}^{q \times 2}$ carries the main effects of proteomic markers, $\mathbf{Q}_1 \in \mathcal{R}^{p \times r}$ and $\mathbf{Q}_2 \in \mathcal{R}^{p \times r}$ is the interaction effects between SNPs and environment markers with respect to intercept and slope respectively. f measures the number of intermediate phenotypes, i.e. intercept and slope. \mathbf{a} and \mathbf{b} are the intercept and slope for normal aging effect. η is a nonnegative tuning parameter to balance the influence of genotype-protein interactions and extract co-expression patterns that contribute to progression prediction.

Motivated by sparse learning techniques, we use $\Omega(\mathbf{Q}) = \|\mathbf{Q}\|_{1,1} = \sum_i \sum_j |\mathbf{Q}_{i,j}|$ to identify interpretable genotype-environment interactions. $\Omega(\mathbf{U})$ and $\Omega(\mathbf{V})$ control the sparsity of the main effects of genetics and proteomic markers. We employ FGL_{2,1}-norm^{*}, $\ell_{2,1}$ -norm, and ℓ_1 -norm to identify a small subset of relevant SNPs at group and individual levels. Thus, $\Omega(\mathbf{U}) = \lambda_{u_1} \|\mathbf{U}\|_{\text{FGL}_{2,1}} + \lambda_{u_2} \|\mathbf{U}\|_{2,1} + \lambda_{u_3} \|\mathbf{U}\|_{1,1}$. In addition, $\ell_{2,1}$ -norm and ℓ_1 -norm are introduced to identify meaningful proteomics expression, i.e. $\Omega(\mathbf{V}) = \lambda_{v_1} \|\mathbf{V}\|_{2,1} + \lambda_{v_2} \|\mathbf{V}\|_{1,1}$, where $\lambda_{v_1}, \lambda_{v_2}$ are nonnegative tuning parameters. Consequently, our method facilitates the tracking of disease progression while identifying the associated genetic factors, encompassing both main and interaction effects. According to (1), MA-DPxGE is multi-convex and thus can be optimized by alternating convex search (ACS) strategy. We first fix \mathbf{V} and \mathbf{Q} and solve for \mathbf{U} using gradient descent, and then iteratively update each variable while treating the others as constants. Through mathematical derivation, we deduce that MA-DPxGE has a lower bound of zero, guaranteeing the convergence to a local optimum. In general, neurodegenerative disorders often affect multiple brain areas, indicating that using a single imaging QT or treating multiple imaging QTs separately may be inadequate for disease progression modeling. Therefore, jointly learning multiple disease progressions can enhance the predictive capability. However, treating all prediction tasks equally may not be optimal, as a simple fusion strategy could prioritize easy sub-objective and neglect the optimization of more challenging ones. To address this issue, we meticulously devise adaptive mechanisms that facilitate loss-term reweighting, i.e. dynamic task balancing module. This mechanism encourages the prioritization of challenging tasks within the multiple prediction tasks, ensuring that they receive adequate optimization and preventing the dominance of easier tasks.

Moreover, the direct application of MA-DPxGE to chromosome-wide or genome-wide analysis poses challenges due to the computationally intensive of genotype matrixes. To handle the high-dimensional SNPs and genotype-environmental interactions, we partition them into L non-intersecting subsets, denoted as $\mathbf{U} =$

$$^* \|\mathbf{U}\|_{\text{FGL}_{2,1}} = \sum_{i=1}^{p-1} \sqrt{\|\mathbf{u}^i\|_2^2 + \|\mathbf{u}^{i+1}\|_2^2}.$$

$\oplus_{l=1}^L \mathbf{U}^l$ and $\mathbf{Q} = \oplus_{l=1}^L \mathbf{Q}^l$ respectively. L can be either user-defined or tuned based on the data. Then, we adopt an effective strategy that avoids the direct calculation of main and interaction terms, i.e, computing main and interaction effects within each subset and then combine the results across all genotypes. This approach considerably reduces the computational complexity. In the spirit of divide-and-conquer principle, we reformulate MA-DPxGE as follows:

$$\begin{aligned}
& \min_{\mathbf{U}, \mathbf{V}, \mathbf{Q}} \sum_{i=1}^n \sum_{k=1}^c \sum_{j=1}^m \Lambda_k \| \mathbf{a}_k + \text{intercept}_{ki} + (\mathbf{b}_k + \text{slope}_{ki}) t_j - \mathbf{y}_{ikj} \|_2^2 \\
& + \eta \sum_{i=1}^n \sum_{k=1}^c \sum_{f=1}^2 \| \mathbf{x}_i (\mathbf{u}_{fk_1} \oplus \dots \oplus \mathbf{u}_{fk_L}) - \mathbf{z}_i \mathbf{v}_{fk} \|_2^2 + \Omega(\mathbf{U}) + \Omega(\mathbf{V}) + \Omega(\mathbf{Q}) \quad (2) \\
\text{st. } & \text{intercept}_{ki} = \mathbf{x}_i (\mathbf{u}_{1k_1} \oplus \dots \oplus \mathbf{u}_{1k_L}) + \mathbf{x}_i (\mathbf{Q}_{1_1} \oplus \dots \oplus \mathbf{Q}_{1_L}) \mathbf{e}_i^T + \mathbf{z}_i \mathbf{v}_{1k}, \\
& \text{slope}_{ki} = \mathbf{x}_i (\mathbf{u}_{2k_1} \oplus \dots \oplus \mathbf{u}_{2k_L}) + \mathbf{x}_i (\mathbf{Q}_{2_1} \oplus \dots \oplus \mathbf{Q}_{2_L}) \mathbf{e}_i^T + \mathbf{z}_i \mathbf{v}_{2k}.
\end{aligned}$$

where $\Lambda_k = -(1 - \text{pc}_k)^{\gamma_k} \log \text{pc}_k$, (pc , prediction criterion; we use correlation coefficients in this work), which is inspired by the focal loss. c represents the number of longitudinal phenotypic markers, γ_k is nonnegative tuning parameters. This allows the model to dynamically prioritize challenging tasks during the training process, as the more challenging target response prediction tasks contribute more to the overall loss and are consequently assigned greater “weight”. The matrix concatenation operator, denoted as \oplus , is used to represent the merging of SNPs and interaction terms. This decoupling enables parallel processing, as SNPs and interaction terms can be independently processed. This approach reduces memory requirements, as fast MA-DPxGE only needs to store small SNP matrices during iterations.

Experiments To evaluate MA-DPxGE, we select most related state-of-the-art models as benchmark methods, including (SMTR, G-SMuRFS and LMTFL) [16,15,10,17]. SMTR and G-SMuRFS model cross-sectional brain imaging QTs at different time points and calculates their average values. LMTFL models the relationship between SNPs and longitudinal imaging QTs for all time points in parallel. To adjust the parameters, we employ a nested 5-fold cross-validation method, exploring a range of 10^i ($i = -3, -2, \dots, 0, \dots, 2, 3$). The performance of all methods are evaluated using correlation coefficient (CCC), root mean square errors (RMSEs), and feature selection ability.

3 Evaluation and Results

3.1 Comparison with state-of-the-art

In Figure 1, we present the CCCs and RMSEs on testing data sets. A lower RMSE and a higher CCC indicates better performance. It is evident that MA-DPxGE outperforms the benchmark methods on both VBM- and Freesurfer-derived data sets, i.e, the best average CCC and RMSE, with the smallest standard error. In addition, intercept and slope of hippocampus and midtemporal (VBM) areas exhibit substantial difference ($p < 0.05$) among different diagnostic

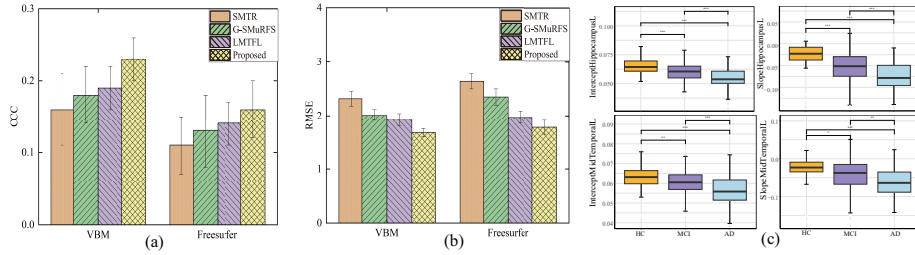


Fig. 1: The testing results in VBM and FreeSurfer (a). The CCC (mean \pm std.) (b). The RMSE (mean \pm std.) obtained from 5-fold cross-validation. (c). Visualization of intercept, slope estimated from qualitative examples. t-test between diagnostic groups are presented. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

groups [18,4]. This indicates the progression speed of endophenotypes is also an important marker for distinguishing AD, MCI and HC. This would also explain why considering disease progression are generally outperform by cross-sectional methods, as the former explicitly models the morphological change between time points. All these results highlight the effectiveness of the mutual-assistance learning module in fitting a more accurate disease progression trajectory.

3.2 Identification and Interpretation of Biomarkers

Accurately and comprehensively identifying underlying pathogenic factors is a crucial objective in longitudinal brain imaging genetics. This endeavor not only enhances our understanding of the disease's pathogenesis but also facilitates the development of effective treatments. To streamline the presentation, we focus on showcasing the results from the VBM dataset.

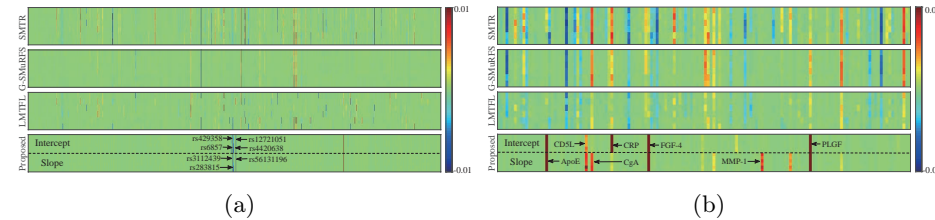


Fig. 2: (a). Weights (heatmap) of SNPs from five-fold cross-validation. (b). Weights (heatmap) of proteomic markers from five-fold cross-validation. (1) SMTR; (2) G-SMuRFS; (3) LMTFL; (4) Proposed.

Main effects: Figure 2a displays the weight coefficients of SNPs. Interestingly, MA-DP \times GE successfully identify several AD-risk loci, including the well-known rs429358 (*APOE*), rs6857 (*TOMM40*), rs4420638 (*APOC1*), rs56131196

(*APOC1*) and rs12721051 (*APOC1*) [12,5]. To test whether the selected SNPs significantly affect AD, we conduct ANOVA to examine their main effects on the diagnostic phenotype. As anticipated, all p -values reach the significance level ($p < 0.05$). Due to FGL_{2,1}-norm, meaningful groups of SNPs are identified, such as rs4420638 ($p = 1.89 \times 10^{-9}$ *), rs56131196 ($p = 1.89 \times 10^{-9}$ *) and rs12721051 ($p = 1.89 \times 10^{-9}$ *) that locate in *APOC1*. These findings could be attributed to and confirm the oligogenic or polygenic characteristic of AD.

In Figure 2b, our method successfully identify several AD-associated proteomic markers, including ApoE, CRP, CgA, FGF-4 [12]. The heatmap in Figure 3 illustrates the interactions between SNP-protein pairs. Notably, the pair (rs429358, APOE) demonstrates the highest correlation, consistent with established findings regarding the strong association between Apolipoprotein E (APOE) and rs429358. In contrast, the competing methods yield numerous irrelevant signals that could potentially introduce misleading information in subsequent analyses. These results highlight the superior performance of MA-DP \times GE in identifying credible AD-affected biomarkers.

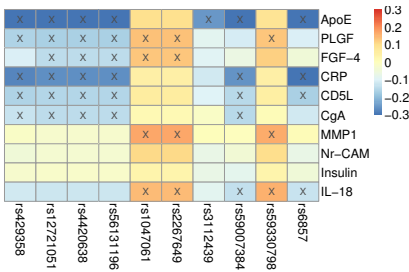


Fig. 3: Heatmap of pairwise correlation between top ten SNPs and proteomic analytes, where symbol “x” indicates the significance level ($p < 0.05$).

Interactions effects: In addition to main effects, our method also uncovers significant genotype-environment interactions. We present the top ten interactions as follows. The interactions identified by the slope are (rs439401, education), (rs59859410, education), (rs8106922, smoking), (rs157582, gender), (rs157581, stroke). The intercept derived interactions include (rs8106922, age), (rs1160984, motor), (rs73052335, stroke), (rs157582, stroke), (rs760136, blood pressure). Firstly, we observe rs439401-education is associated with the slope. Notably, rs439401 has been previously reported AD risk, and education is a important factor that could delay the onset of AD [1]. These interactions can provide promising evidence for precise diagnosis, because the co-occurrence of these abnormalities may help clinicians be confident to diagnose at-risk individuals. Meanwhile, rs1160984-motor is related to the rate of disease change, this might be attributed to that exercise-induced elevation of brain-derived neurotrophic factor is crucial for neuronal growth [11]. Further investigations into other interactions may reveal novel GE interactions associated with AD risk and pro-

<i>id</i>	SVR	NA	DTB	\mathcal{L}_{FIL}	CCC \uparrow	RMSE \downarrow
(a)	\times	\times	\times	\times	0.17 ± 0.04	2.23 ± 0.12
(b)	\checkmark	\times	\times	\times	0.19 ± 0.04	1.91 ± 0.10
(c)	\checkmark	\checkmark	\times	\times	0.20 ± 0.05	1.83 ± 0.08
(d)	\checkmark	\checkmark	\checkmark	\times	0.21 ± 0.03	1.74 ± 0.09
(e)	\checkmark	\checkmark	\checkmark	\checkmark	0.23 ± 0.03	1.67 ± 0.08

Table 1: Ablation studies on main modules of different design choices. ‘‘SVR’’ denotes sparse variable regularizer. ‘‘NA’’ is normal aging effect. ‘‘DTB’’ represents dynamic task balancing, ‘‘ \mathcal{L}_{FIL} ’’ means feature interaction learning. **Bold** indicates the best result, \uparrow means higher is better, and \downarrow lower is better.

gression. We also remove feature interaction learning module from MA-DP \times GE (Table 1). No surprisingly, the best performance is achieved when both the main and interaction effects of genotype and environment are considered, highlighting the necessity of incorporating both aspects. These findings demonstrate that GE interactions can yield interpretable biomarkers for AD progression prediction.

3.3 Ablation Study

We run ablation experiments to investigate the impact of main component on Correlation and Prediction tasks.

Effect of sparse variable regularizer module: We employ CCCs and RMSE as evaluation metrics to assess the performance. Table 1 presents the testing CCCs and RMSE scores for all ablation studies. Notably, the absence of SVR module results in inferior performance, underscoring the importance of incorporating SVR into predictive tasks. This suggests that the inclusion of irrelevant features may negatively impact the model’s performance.

Effect of the normal aging effect module: Considering RMSEs and CCCs, ablation results suggest that separating normal aging leads to a better fit for disease progression. These findings can be attributed to the potential hindrance caused by the normal aging effect on accurately modeling disease progression.

Effect of the dynamic task balancing module: Clearly, MA-DP \times GE achieves the highest average CCCs and smallest RMSE when incorporating the task balancing module, showcasing its superior overall performance. The observed improvement can be attributed to the suboptimal nature of treating all sub-objectives equally. A simple fusion strategy might prioritize easier sub-objectives while inadequately optimizing the more challenging ones.

Effect of the feature interaction learning module: Table 1 demonstrates the superior performance of our method over baselines. This highlights the effectiveness of feature interaction learning module. Additionally, our method excels in interpretability, emphasizing the significance of identifying interpretable GE interactions, which has the potential to explain the missing heritability of AD.

4 Conclusions

This study proposes the first interpretable framework for disease progression prediction, which can be applied to the diagnosis and prognosis of neurodegenerative diseases, namely MA-DP \times GE. Leveraging longitudinal imaging phenotypes, experimental results demonstrate the superior performance of our framework, accompanied by excellent built-in interpretability that supports the AD-specific progression trajectory. Moreover, our method enhances the prediction and identification of main and interaction effects, carrying significant implications. Furthermore, we extend the current task to a chromosome-wide setting and produce strong baseline results. However, focusing on non-Hispanic Caucasian individuals could lead to less generalisability and potential bias. In future endeavors, we are working on collecting additional ethnicity, and conducting repeated grouping experiments, to demonstrate general effectiveness under various biased conditions.

Acknowledgements. This work was supported in part by the STI2030-Major Projects (No. 2022ZD0213700), National Natural Science Foundation of China (No. 61973255, 62136004, 61936007, 62373306), Innovation Foundation for Doctor Dissertation (No. CX2023062) and Fundamental Research Funds for the Central Universities at Northwestern Polytechnical University.

Disclosure of Interests. No competing interests.

References

1. Aslam, M.M., Fan, K.H., Lawrence, E., Bedison, M.A., Snitz, B.E., DeKosky, S.T., Lopez, O.L., Feingold, E., Kamboh, M.I.: Genome-wide analysis identifies novel loci influencing plasma apolipoprotein e concentration and alzheimer’s disease risk. *Molecular Psychiatry* **28**(10), 4451–4462 (2023)
2. Bettencourt, C., Skene, N., Bandres-Ciga, S., Anderson, E., Winchester, L.M., Foote, I.F., Schwartzenuber, J., Botia, J.A., Nalls, M., Singleton, A., et al.: Artificial intelligence for dementia genetics and omics. *Alzheimer’s & Dementia* **19**(12), 5905–5921 (2023)
3. Chen, Z., Qing, J., Zhou, J.H.: Cinematic mindscapes: High-quality video reconstruction from brain activity. *Advances in Neural Information Processing Systems* **36** (2024)
4. Du, J., Rolls, E.T., Gong, W., Cao, M., Vatansever, D., Zhang, J., Kang, J., Cheng, W., Feng, J.: Association between parental age, brain structure, and behavioral and cognitive problems in children. *Molecular Psychiatry* **27**(2), 967–975 (2022)
5. Du, L., Zhang, J., Liu, F., Wang, H., Guo, L., Han, J., Initiative, A.D.N., et al.: Identifying associations among genomic, proteomic and imaging biomarkers via adaptive sparse multi-view canonical correlation analysis. *Medical Image Analysis* **70**, 102003 (2021)
6. Du, L., Zhang, J., Zhao, Y., Shang, M., Guo, L., Han, J.: inmtscca: An integrated multi-task sparse canonical correlation analysis for multi-omic brain imaging genetics. *Genomics, Proteomics & Bioinformatics* **21**(2), 396–413 (2023)

7. Du, L., Zhao, Y., Zhang, J., Shang, M., Zhang, J., Han, J.: Identification of genetic risk factors based on disease progression derived from longitudinal brain imaging phenotypes. *IEEE Transactions on Medical Imaging* (2023)
8. Guo, Y., You, J., Zhang, Y., Liu, W.S., Huang, Y.Y., Zhang, Y.R., Zhang, W., Dong, Q., Feng, J.F., Cheng, W., et al.: Plasma proteomic profiles predict future dementia in healthy adults. *Nature Aging* **4**(2), 247–260 (2024)
9. Liu, F., Xu, J., Guo, L., Qin, W., Liang, M., Schumann, G., Yu, C.: Environmental neuroscience linking exposome to brain structure and function underlying cognition and behavior. *Molecular Psychiatry* **28**(1), 17–27 (2023)
10. Lu, L., Wang, H., Yao, X., Risacher, S., Saykin, A., Shen, L.: Predicting progressions of cognitive outcomes via high-order multi-modal multi-task feature learning. *International Symposium on Biomedical Imaging* pp. 545–548 (2018)
11. Migliore, L., Coppedè, F.: Gene–environment interactions in alzheimer disease: the emerging role of epigenetics. *Nature Reviews Neurology* **18**(11), 643–660 (2022)
12. Serrano-Pozo, A., Das, S., Hyman, B.T.: Apoe and alzheimer’s disease: advances in genetics, pathophysiology, and therapeutic approaches. *The Lancet Neurology* **20**(1), 68–80 (2021)
13. Shen, L., Thompson, P.M.: Brain imaging genomics: integrated analysis and machine learning. *Proceedings of the IEEE* **108**(1), 125–162 (2019)
14. Sims, R., Hill, M., Williams, J.: The multiplex model of the genetics of alzheimer’s disease. *Nature neuroscience* **23**(3), 311–322 (2020)
15. Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S.L., Saykin, A.J., Shen, L., Initiative, A.D.N.: Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the adni cohort. *Bioinformatics* **28**(2), 229–237 (2012)
16. Wang, H., Nie, F., Huang, H., Risacher, S., Ding, C., Saykin, A.J., Shen, L.: Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. In: *International Conference on Computer Vision*. pp. 557–562 (2011)
17. Wang, H., Nie, F., Huang, H., Yan, J., Kim, S., Risacher, S., Saykin, A., Shen, L.: High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer’s disease progression prediction. *Advances in neural information processing systems* **25** (2012)
18. Wang, Y., Zhang, J., Li, Y., Qi, S., Zhang, F., Ball, L.J., Duan, H.: Preventing prefrontal dysfunction by tdcs modulates stress-induced creativity impairment in women: an fnirs study. *Cerebral Cortex* **33**(20), 10528–10545 (2023)
19. Wu, H., Eckhardt, C.M., Baccarelli, A.A.: Molecular mechanisms of environmental exposures and human disease. *Nature Reviews Genetics* **24**(5), 332–344 (2023)
20. Zhang, J., Ma, Z., Yang, Y., Guo, L., Du, L., Initiative, A.D.N.: Modeling genotype–protein interaction and correlation for alzheimer’s disease: a multi-omics imaging genetics study. *Briefings in Bioinformatics* **25**(2), bbae038 (2024)
21. Zhang, J., Shang, M., Xie, Q., Zhang, M., Xi, D., Guo, L., Han, J., Du, L.: A sparse multi-task contrastive and discriminative learning method with feature selection for brain imaging genetics. *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* pp. 660–665 (2022)
22. Zhang, J., Wang, H., Zhao, Y., Guo, L., Du, L., Initiative, A.D.N.: Identification of multimodal brain imaging association via a parameter decomposition based sparse multi-view canonical correlation analysis method. *BMC bioinformatics* **23**(Suppl 3), 128 (2022)