bioRxiv preprint doi: https://doi.org/10.1101/2022.05.02.490336; this version posted May 3, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Systems biology

# Deep multi-omic network fusion for marker discovery of Alzheimer's Disease

**Linhui Xie [1,†], Yash Raj [2,†], Pradeep Varathan [2], Bing He [2], Kwangsik Nho [3], Paul Salama [1], Andrew J. Saykin [3], and Jingwen Yan [2,3,*]**

[1]Department of Electrical and Computer Engineering, Indiana University Purdue University Indianapolis, Indianapolis, IN 46204, USA, [2]Department of BioHealth Informatics, Indiana University Purdue University Indianapolis, Indianapolis, IN 46204, USA, [3]Department of Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, IN 46204, USA

*To whom correspondence should be addressed.

†Contributed equally.

## Abstract

**Motivation:** Multi-omic data spanning from genotype, gene expression to protein expression have been increasingly explored, with attempt to better interpret genetic findings from genome wide association studies and to gain more insight of the disease mechanism. However, gene expression and protein expression are part of dynamic process changing in various ways as a cell ages. Expression data captured by existing technology is often noisy and only capture a screenshot of the dynamic process. Performance of models built on top of these expression data is undoubtedly compromised. To address this problem, we propose a new interpretable deep multi-omic network fusion model (MoFNet) for predictive modeling of Alzheimer's disease. In particular, the information flow from DNA to protein is leveraged as a prior multi-omic network to enhance the signal in gene and protein expression data so as to achieve better prediction power.

**Results:** The proposed model MoFNet significantly outperformed all other state-of-art classifiers when evaluated using genotype, gene expression and protein expression data from the ROS/MAP cohort. Instead of individual markers, MoFNet yielded 3 major multi-omic subnetworks related to innate immune system, clearance of unwanted cells or misfolded proteins, and neurotransmitter release respectively.

**Availability:** The source code is available through GitHub (https://github.com/yashraj59/MoFNet). Multi-omic data used in this analysis is from the ROS/MAP project and is available upon application through the AMP-AD knowledge portal (https://adknowledgeportal.synapse.org).

**Contact:** jingyan@iupui.edu

## 1 Introduction

Large-scale genome-wide association studies (GWASs) have been widely applied to mine the associations between genomic variants and various human diseases including Alzheimer's disease (AD) (MacArthur *et al.*, 2017). Findings from these studies have significantly advanced our understanding of genetic factors underlying AD. However, the functional mechanism through which those candidate genetic variants exert effect to the downstream gene expression and protein expression remains largely unknown.

To better interpret these genetic findings from GWAS and to gain more insight of the disease mechanism, multi-omic data spanning from genotype, gene expression to protein expression have been increasingly explored (Hasin *et al.*, 2017; Huang *et al.*, 2017). Unlike other studies focusing on single -omics types, integrative -omics analysis holds potential to reveal the genetic markers evidenced from multiple aspects. Findings from integrative -omics studies are expected to have better interpretability and to provide more insights of functional mechanisms underlying AD.

Early multi-omic studies examined multi-omic data in an isolated fashion or in a sequential way. For the first type, each -omics data type were analyzed individually and then they seek overlapped genetic markers across them. For the second type, findings from upstream layers (e.g., genotype) were used as seed to narrow down the search space in the downstream layers (e.g., gene expression) (Nativio *et al.*, 2020). In both scenarios, -omics data types are examined one after another and connections between -omics layers are minimally considered. To

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.02.490336; this version posted May 3, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.



address this problem, recent multi-omic studies started to model multi-omic data together with the rich biological networks, capturing the functional interactions between genetic variations, genes and proteins (Xie *et al.*, 2020, 2021). Instead of individual multi-omic markers not guaranteed to interact, these new approaches help yield subnetworks of functionally connected multi-omic markers and provide a better foundation for generating new hypothesis of AD regarding the underlying molecular mechanism.

However, gene expression and protein expression are part of dynamic processes changing in various ways as a cell ages (Slavov, 2020; Cookson *et al.*, 2005). Expression data captured by existing technology are often noisy and only capture a screenshot of that dynamic process. Performance of models built on top of these expression data will undoubtedly be compromised. To address this problem, we propose a new interpretable graph neural network model to enhance the signals in the gene and protein expression data, where prior multi-omic network will be embedded. More specifically, given the information flow from DNA to protein, we hypothesize that gene expression is to some extent influenced by their upstream functional interactors (e.g., genetic variations or expression quantitative trait loci) and protein expression is partly determined by the expression of genes related to them. For each gene and protein, their functional interactors (either genetic variations or genes) should bear some information that can be leveraged to enhance its signals. Further, we hypothesize that not all functional interactors are equal contributors during this process. Another goal of the proposed model is to learn a subset of functional interactors for each gene and protein, with whom their expression data can be further enhanced to give more prediction power. A combination of selected interactors for each gene and protein will then form a multi-omic subnetwork that contributes to the final prediction performance. We evaluated our model together with five other state-of-art classification models using the genotype, gene expression and protein expression data from the ROS/MAP cohort. The proposed graph neural network model showed a significant improvement over all other models and yielded 3 major subnetworks related to innate immune system, clearance of unwanted cells or misfolded proteins, and neurotransmitter release respectively.

## 2 Material and Methods

### 2.1 Data

Multi-omic data sets used in this study were obtained from the Religious Orders Study (ROS) and Memory and Aging Project (MAP) cohorts (A Bennett *et al.*, 2012). Both cohorts study the risk factors for cognitive decline and incident AD dementia, and have been providing valuable multi-omic data resource to the research community. Via synapse AMP-AD portal (Hodes and Buckholtz, 2016), we downloaded imputed genotype, RNA-Seq gene expression, protein expression and diagnosis information for ~600 participants. All the gene expression and protein expression were collected from prefrontal cortex region from postmortem brains of participating subjects. Participants missing one or more -omics data types were excluded and finally 133 participants with full set of three -omics data types were included for the subsequent predictive modeling (77 cognitive normal (CN) and 56 AD patients). Shown in Table 1 is the detailed demographic information of included participants. We observed that the female/male ratio were relatively higher in AD group than in CN group, with an average of 3 years older. This is consistent with existing findings that age and gender are two prominent risk factors for AD. For another intrinsic feature, no significant difference was observed across diagnosis groups for education years.

Table 1. **Participant demographic information.**

| **Diagnosis** | **CN** | **AD** |
|---|---|---|
| Subject Number | 77 | 56 |
| ROS/MAP | 47/30 | 28/28 |
| Male/Female | 35/42 | 22/34 |
| Education(mean± std.) | 16.7 ± 3.2 | 16.8 ± 3.7 |
| Age(mean± std.) | 83.0 ± 4.5 | 86.3 ± 3.5 |

### 2.2 Multi-omic Data Preprocessing

GWAS genotype data preparation, such data were pre-processed through the Affymetrix GeneChip 6.0 platform (De Jager *et al.*, 2012). The minor allele frequency (MAF) of 1% were used to quality control GWAS data samples. Then, using HapMap 3 genotype data and multi-dimensional scaling analysis, specific participants were chosen by clustering with two other group populations (Horgusluoglu-Moloch *et al.*, 2017). With the 1000 Genomes Project as the reference panel, un-genotyped SNPs were imputed using the Markov Chain framework for genotype imputation and haplotyping (MaCH) (Nho *et al.*, 2013).

RNA-Seq gene expression data was first collected from prefrontal cortex tissue in the postmortem brains of participants from ROS/MAP cohort. The RNA-Seq data were reprocessed in parallel with RNA-Seq datasets from the Accelerating Medicines Partnership for Alzheimer's Disease (AMP-AD) (Hodes and Buckholtz, 2016). For subsequent analysis, the data with expected better quality was employed to be aligned reads in bam files that were then converted to fastq using the Picard SamToFastq tool. STAR was used to re-align fastq files to the reference genome, with twopassMode set to Basic. The effects of relevant factors such as age, gender, education, etc. were removed through normalization and adjustment.

Protein expression data was collected from exactly the same tissue samples as RNA-seq gene expression data. The samples were quantified with selected reaction monitoring (SRM) technique, and then prepared according to standard methodology for liquid chromatography-selected reaction monitoring (LC-SRM) analysis (Andreev *et al.*, 2012; Petyuk *et al.*, 2010). All of the data was manually checked to confirm that the peak assignments and boundaries were valid. A specific ratio to spiked sythetic peptides containing stable heavy isotopes was quantified from the abundant endogenous peptides. The "light/heavy" ratios were log2-transformed and shifted to guarantee a value zero for the median log2-ratio. The log2-ratios for each sample were changed during normalization to ensure that the median was set to zero. Similarly, the effects of age, gender, education, etc. were removed for peptide abundance data. In total, there are genotype of 6,115,610 single nucleotide polymorphism (SNPs) and expression data of 15,582 genes and 186 peptides.

### 2.3 SNP and Gene Filtering

Given the large number of SNPs and genes, it is infeasible to directly model all genetic features and evaluate their overall predictive power. Given that protein expression is the data bottleneck with only 186 peptides (mapped to 126 unique genes), we took a bottom-up approach to narrow down the total number of -omics features by using these peptides as seeds and selecting only a subset of SNPs, genes functionally related to them. In the proteomic layer, abundance level of 186 peptides were measured in the ROS/MAP cohort. These peptides were mapped to 126 unique genes (gene set A), which were found to interact with 954 genes (gene set B) in the functional interaction network obtained from the REACTOME database (Fabregat *et al.*, 2018). Among these 1080 (126 + 954) genes, 743 without missing RNA-seq data were included to represent the transcriptomic level in our model. In the genomic level, we identified SNPs located on the upstream

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.02.490336; this version posted May 3, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.



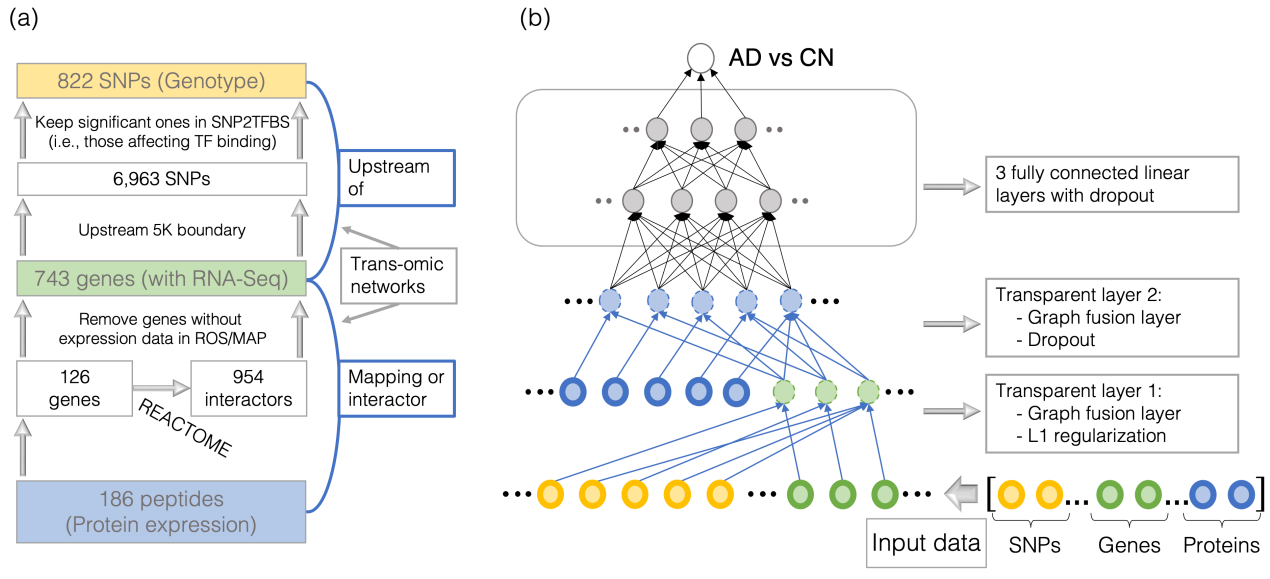


**Fig. 1.** SNP/Gene filtering and the architecture of MoFNet. (a) Workflow of SNP and gene filtering using peptide as seed. (b) Architecture of the proposed MoFNet model, with the prior trans-omic network in (a) embedded in the first two transparent layers.

of these 743 genes within the boundary of 5K. To ensure the functional connection of selected SNPs and their downstream genes, we included only SNPs significantly associated with the transcription factor-binding activity, based on the single nucleotide polymorphisms to transcription factor binding sites (SNP2TFBS) database (Kumar *et al.*, 2017). Taken together, we have 822 SNPs, 743 genes and 186 peptides for the subsequent modeling. The functional relationships used to filter the genes and SNPs form a trans-omic functional interaction network, which will be embedded into the deep neural network to guide the search of functionally connected features related to AD (Fig. 1 (a)).

## 2.4 Prediction Outcome

Extracted SNP genotype, gene expression and protein expression data were used to classify AD patients from cognitive normals (CN). For all the participants included in this study, their final clinical diagnosis at the point of brain tissue collection was used to indicate their disease status. In this case, the diagnosis time remains consistent with the expression data collection time. Since the mild cognitive impairment (MCI) participants in the ROS/MAP cohort were defined as non-symptomatic group, it makes more sense to group them with cognitive normal subjects but that will lead to extremely imbalanced dataset. Therefore, MCI subjects were excluded from the analysis.

## 2.5 Architecture of MoFNet

The proposed deep multi-omic network fusion model (MoFNet) is a graph neural network with two transparent layers structured based on the functional connectivity between SNPs, genes and proteins (i.e., trans-omic network shown in Fig. 1(a)). Shown in Fig. 1(b) is the detailed architecture of the proposed MoFNet. For the first transparent layer, we have 1565 input nodes, corresponding to 822 SNPs and 743 genes respectively, and 743 output nodes, corresponding to 743 enhanced genes. Links in the first transparent layer were added if one SNP (as an input node) is located upstream of one gene (as an output node). In addition, we added links between same genes (i.e., gene A node in the input and gene A node in the output). In this case, output gene nodes will have enhanced expression data by integrating information from its upstream SNPs and its original measured expression level. We assume that not all SNPs are equally informative and helpful in enhancing the signal of the downstream gene. Therefore, L1 regularization was applied for the first transparent layer such that links between non-informative SNPs and their downstream genes will mostly get zero weight. For the second transparent layer, we have 929 input nodes, corresponding to 743 enhanced genes and 186 peptides, and 186 output nodes, corresponding to 186 enhanced peptides. Links were added if one gene (as an input node) functionally interacts with the gene corresponding to the peptide (as an output node) as indicated in the REACTOME database. We also added links between the same peptide (i.e., peptide A node in the input and peptide A node in the output). Taken together, the enhanced peptide nodes, as the output of the second transparent layer, will capture the information from its corresponding genes, their interactors and upstream SNPs. After that, we have 3 fully connected layers to classify the AD patients from cognitive normal subjects. We used dropout and early termination to avoid overfitting.

1. Input $X_1$ is the concatenation of the gene expression matrix $G^{n\times g}$ ( $n$ samples by $g$ genes), and SNP genotype matrix $S^{n\times s}$ ( $n$ samples by $s$ SNPs ). $X_1^{n\times(g+s)} = [G, S]$ where $[\cdot]$ stands for row concatenation.
2. The output from the first transparent layer $Z_{11}$ has the dimension as the number of genes $g$. Links in this layer indicate the prior functional connections between SNPs and genes, and between same genes. Functional connections between SNPs and genes were encoded in an adjacency matrix $A_1^{s\times g}$. $A(i,j) = 1$ indicates SNP $i$ is located upstream of gene $j$ and likely to affect the transcription factor binding activity; $A(i,j) = 0$ otherwise. We also added self-connections to genes by adding another adjacency matrix $A_2^{g\times g}$, which is an identity matrix with $A_2(i,i) = 1$. Taken these two adjacency matrices together, the first transparent layer is a 'Biological DropConnect' layer

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.02.490336; this version posted May 3, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.



Table 2. Performance comparison of MoFNet and other competing classification methods across five test data sets (mean ± std).

| | Accuracy | F1 score | Precision | Sensitivity | Specificity | AUC |
|---|---|---|---|---|---|---|
| MoFNet(Proposed) | 0.75 ± 0.097 | 0.68 ± 0.115 | 0.74 ± 0.167 | 0.65 ± 0.159 | 0.82 ± 0.119 | 0.73 ± 0.099 |
| Logistic(Modularity) | 0.64 ± 0.084 | 0.54 ± 0.097 | 0.59 ± 0.132 | 0.50 ± 0.079 | 0.74 ± 0.101 | 0.67 ± 0.113 |
| Logistic(ElasticNet ) | 0.62 ± 0.051 | 0.51 ± 0.083 | 0.56 ± 0.062 | 0.48 ± 0.117 | 0.73 ± 0.073 | 0.61 ± 0.052 |
| Logistic(GraphNet) | 0.62 ± 0.104 | 0.46 ± 0.143 | 0.59 ± 0.181 | 0.38 ± 0.124 | 0.80 ± 0.093 | 0.59 ± 0.117 |
| Logistic(Lasso) | 0.61 ± 0.070 | 0.60 ± 0.061 | 0.53 ± 0.065 | 0.70 ± 0.069 | 0.55 ± 0.100 | 0.62 ± 0.066 |
| Random Forest | 0.67 ± 0.076 | 0.37 ± 0.166 | 0.87 ± 0.163 | 0.25 ± 0.147 | 0.73 ± 0.073 | 0.59 ± 0.091 |

(Nguyen *et al.*, 2021; Wan *et al.*, 2013). Therefore, weight matrix of this layer $W_1$ has a sparse structure with a dimension of $(s+g) \times g$. Output of this layer $Z_1 = f\left(X_1\left(W_1 \odot \left[A_1^T, A_2^T\right]^T\right) + b_1\right)$ where $\odot$ is the Hadamard product, and $(\cdot)^T$ is the matrix transpose operator.

3. The second transparent layer resembles the structure of the second part of the prior trans-omic network, i.e., the functional connections between genes and proteins. The input of this layer is the concatenation of the protein expression (e.g. peptides) data $X_2^{n\times p}$ ( $n$ samples by $p$ peptides) and output of the first transparent layer $Z_{11}$, i.e., $Z_1 = [X_2, Z_{11}]$. The output of the second transparent layer $Z_2$ has a dimension of the number of peptides. Weight of this layer $W_2$ has a dimension of $(g+p) \times p$. The adjacency matrix $A_3^{g\times p}$ indicates the functional connections between genes and proteins, where $A_3(i,j) = 1$ if gene $i$ encodes protein $j$ itself or the functional interactor of protein $j$; $A_3(i,j) = 0$, otherwise. Similarly, we added self-connections between peptides with an identity adjacency matrix $A_4^{p\times p}$ where $A_4(i,i) = 1$. The output of the second transparent layer is $Z_2 = f\left(Z_1\left(W_2 \odot \left[A_3^T, A_4^T\right]^T\right) + b_2\right)$.
4. Three fully connected hidden layers $Z_l$ index by $l \in \{3 \ldots L-1\}$ were used together with a sigmoid function in the last layer. $Z_L = \sigma\left(Z_{L-1}W_L + b_L\right)$.
5. Finally, we use binary cross-entropy loss to quantify the classification error: $L(y,\hat{y}) = -1/n \sum_{i=1}^{n} y_i \log\left(\hat{y}_i\right) + (1-y_i)\log\left(1-\hat{y}_i\right)$.

### 2.6 Parameter Tuning

Given the relatively small sample size, we adopted 5 fold cross validation with grid search to tune the parameters.Training-testing split was set as 8:2 for the entire dataset. The AD:CN ratio was maintained the same across all training and testing set in each fold. To ensure the fair comparison of competing methods, all methods including the proposed MoFNet went through exactly the same parameter tuning process with 5-fold cross validation. All the training and test partitions used across methods were kept exactly the same. In addition, early termination and dropout layers were added in MoFNet to avoid overfitting. Finally, the best performance of MoFNet was obtained with L1 regularization rate as 0.005, learning rate as 0.001, dropout rate as 0.5, and weight decay as 0.0008. The dimension of last 3 fully connected layers are (186, 96), (96,16) and (16,1) respectively.

### 2.7 Interpretability

The interpretability of the proposed model is achieved in two folds. First, with L1 regularization, weight learned for the links in the first two transparent layers has many zeros. Mapping those weights to the prior trans-omic network will help prune the prior network and lead to subnetworks indicating important information flow from DNA to protein that contributes to the prediction of AD. Secondly, integrated gradient was applied to prioritize node importance for the outcome (Sundararajan *et al.*, 2017). The gradient of model prediction for each multi-omic feature indicates how the prediction outcome responds to the changes of the multi-omic features. This importance score will provide the potential explanations of which part of the pruned trans-omic subnetwork contribute the most to the disease outcome.

## 3 Result

The performance of the proposed MoFNet model was evaluated using the genotype, gene expression and protein expression data collected from prefrontal cortex tissue of postmortem brain samples in the ROS/MAP cohort. We compared its performance with random forest and four other state-of-the-art logistic regression based classification models, using modularity, elastic net, GraphNet and Lasso as penalty terms, respectively (Newman, 2006; Zou and Hastie, 2005; Grosenick *et al.*, 2013; Tibshirani, 1996). These sparse logistic regression models were selected because they can perform classification and feature selection at the same time. Classic classification models, such as support vector machine (SVM) and k nearest neighbor (KNN), can not provide selection of features, and therefore are not included for comparison. The modularity constrained logistic regression was implemented in MATLABw (Xie *et al.*, 2021). GraphNet weasre implemented using R package (Chen *et al.*, 2015). Elastic constrained logistic regression, traditional logistic regression with lasso penalty, and random forest were implemented using python scikit-learn package (Pedregosa *et al.*, 2011). To provide an unbiased comparison of performance, partition of subjects in all training and testing set was kept identical for all methods. We performed grid search and 5-fold cross validation for all methods, and hyper-parameters that yielded best prediction performance across 5 folds were selected as optimal parameters.

### 3.1 Performance Comparison

Shown in Table. 2 is the average performance (and standard deviation) of the proposed MoFNet and other competing methods across five test set. Due to slight imbalance of case and control numbers in our data, we reported not only accuracy and AUC, but also F1 score, precision, specificity and sensitivity metrics to give a comprehensive comparison of performance from multiple perspectives. In particular, F1 score combines precision and recall into a single metric, and has been used as a major criteria for evaluation of model performance when dealing with imbalanced data sets. We observed that the proposed MoFNet largely outperforms other state-of-the-art classification models, with highest average accuracy, specificity, AUC, and F1-score, indicating its capability in handling the imbalanced dataset. Compared with penalized sparse logistic regression models, our model has a significant improvement on precision, the availability to correctly identify AD patients. Random forest scored the second-best accuracy. However, its prediction is highly biased toward the class with larger sample size and therefore ended up with worst F1 score.

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.02.490336; this version posted May 3, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.



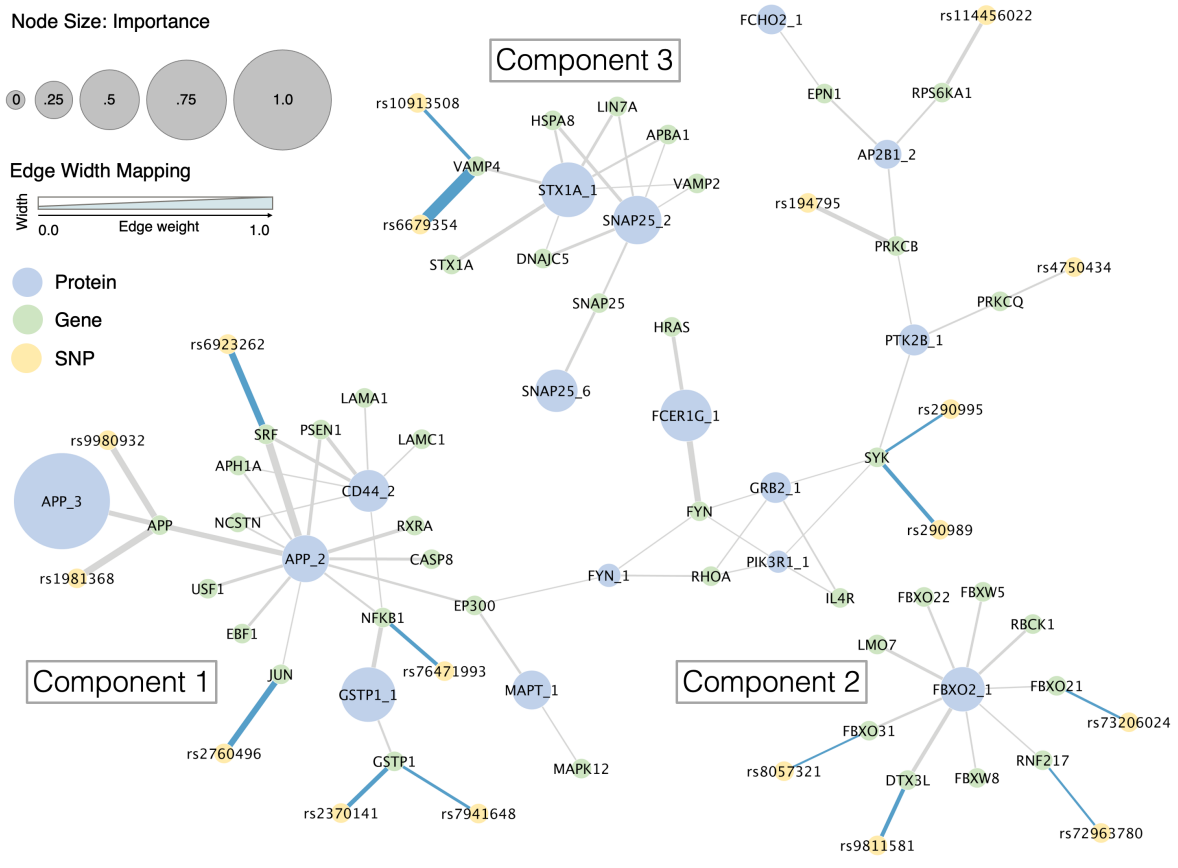


**Fig. 2.** Three major connected components (i.e., subnetwork) identified from pruning the prior trans-omic network using the weight and node importance score derived from MoFNet. Node size is proportional to the node importance score. Edge width is proportional to the weight obtained on the trained biological network. Blue edges indicates already known SNP-gene pairs in which a SNP is associated with the expression of its connecting gene in the frontal cortex tissue, where our expression data was collected. Numbers shown at the end of protein names indicate different peptides corresponding to that protein.

## 3.2 Multi-omic Subnetwork Extraction

In addition to the classification performance, it is also of great importance to identify a subset of multi-omic features that contribute to the final prediction and to learn how they are functionally connected in prior knowledge. For all competing methods, only modularity constrained lasso and elastic net constrained logistic regression models yielded a subset of SNPs, genes and proteins with some known functional connections in the prior network. Multi-omic features identified by other competing methods mostly scattered around the prior trans-omic network without much known connections.

For MoFNet, we mapped the weight of first two transparent layers from each fold to the prior trans-omic network and then pruned the network by removing links with coefficient as or close to zero. We carefully selected the cut-off thresholds to filter out the links and the nodes with low importance score. Various cutoff values were tested and final cutoff was selected by observing the number of nodes in the largest connected component, which converges around the cutoff threshold and suddenly increase significantly after that. This helped set both thresholds at $10^{(-7)}$. The combined two-stage filtering (i.e., network pruning) procedures are as follows: 1) removing the links with weight smaller than edge cut-off threshold; 2) further removing the links connecting any node with importance score less than node cut-off threshold. As such, we obtained 5 pruned trans-omic networks from 5 folds and links that appear in >=3 folds were kept in the final network. The final pruned network has total number of 81 multi-omic features, including 18 proteins, 45 genes and 18 SNPs.

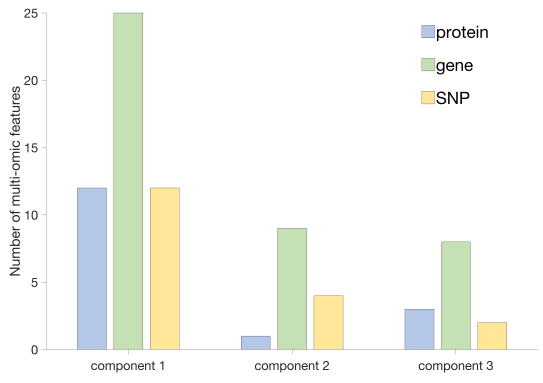


**Fig. 3.** Count of proteins, genes and SNPs in 3 major connected component in the multi-omic network pruned by weights derived from MoFNet.

Among those, 2 proteins and 3 genes were found to be individual nodes without any connection in the prior network. The rest of the 76 multi-omic features formed 3 major connected components (i.e., subnetworks), as shown in Fig. 2. Here, node size is proportional to the average node importance score across 5 folds, calculated through integrated gradient. The larger the node size, the more contribution it makes to the final prediction. Similarly, edge width is proportional to the average edge

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.02.490336; this version posted May 3, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.



weight of first two transparent layers across 5 folds. The thicker the edge is, the more information integration occurs from SNPs to genes or from genes to proteins. Edges highlighted as blue are SNP-genes pairs in which SNPs are known to influence the expression of the connecting gene in the frontal cortex, where the gene expression and protein expression data were collected. As expected, nodes with top importance scores (i.e., large nodes) are mostly proteins. As the output of the second transparent layer, enhanced protein nodes integrated information from SNPs, genes and raw protein expression measurements, and therefore bear much more predictive power.

Shown in Fig. 3 is the number of SNPs, genes and proteins in all three connected components. The largest connected component has 12 SNPs, 25 genes and 12 proteins, including the amyloid precursor protein *APP* and its corresponding gene. In particular, peptide *APP* 3 has the largest importance score indicating its major contribution to the prediction performance of MoFNet. The thick edges between peptide *APP* 3 and gene *APP*, and between gene *APP* and SNPs rs9980932, rs1981368 suggest an important information flow from DNA to RNA and then protein that may be responsible for the development of AD. Some other trans-omic paths from SNPs to genes and proteins with high weights and high node importance scores are (rs6923262, *SRF*, *APP* 2), (rs6679354, *VAMP4*, *STX1A* 1), and (rs9811581, *DTX3L*, *FBXO2* 1). All these trans-omic paths warrant further investigation in terms of their specific roles in AD.

## 3.3 eQTL Analysis

We further investigated the function of all SNPs in 3 major components on the downstream transcriptome level. We examined all 18 SNPs in the BRAINEAC database and found that all of them are significant expression quantitative trait locus (eQTL) in the frontal cortex region. That means, variations in these SNPs are associated with gene expression levels in frontal cortex tissue. Then, we took a step further and examined whether those SNPs are eQTLs of their connecting genes in these 3 connected components. Among all 18 SNP-gene pairs as shown in Fig. 2, 13 of them have been identified to have significant associations in the frontal cortex tissue in the BRAINEAC database (edges highlighted in blue in Fig. 2). While each gene has multiple transcripts, those 13 SNPs were found to be significantly associated with at least one transcript of its connecting genes. Due to the space limit, we only listed the strongest association between each SNP and gene in Table. 3.

Table 3. Significant eQTL association of our identified SNP-Gene pairs in the frontal cortex tissue based on the BRAINEAC database.

| Chromosome | Gene | SNP | Adjusted P-val |
|---|---|---|---|
| 11 | *GSTP1* | rs7941648 | $1.20\text{E}-11$ |
| 11 | *GSTP1* | rs2370141 | $1.20\text{E}-11$ |
| 1 | *VAMP4* | rs6679354 | $3.90\text{E}-03$ |
| 1 | *VAMP4* | rs10913508 | $8.50\text{E}-03$ |
| 16 | *FBXO31* | rs8057321 | $4.90\text{E}-03$ |
| 9 | *SYK* | rs290995 | $5.30\text{E}-03$ |
| 9 | *SYK* | rs290989 | $5.30\text{E}-03$ |
| 4 | *NFKB1* | rs76471993 | $7.10\text{E}-03$ |
| 1 | *JUN* | rs2760496 | $1.40\text{E}-02$ |
| 12 | *FBXO21* | rs73206024 | $1.80\text{E}-02$ |
| 3 | *DTX3L* | rs9811581 | $2.40\text{E}-02$ |
| 6 | *SRF* | rs6923262 | $4.20\text{E}-02$ |
| 6 | *RNF217* | rs72963780 | $4.70\text{E}-02$ |

## 3.4 Pathway Enrichment Analysis

While our identified multi-omic features naturally form 3 major connected components, it will be of great interest to examine the function of each component. For all the genes and proteins in each component, we performed enrichment analysis in REACTOME pathways using EnrichR (Kuleshov *et al.*, 2016; Fabregat *et al.*, 2018). For the largest component with hubs *APP* and *CD44*, 25 genes and 12 proteins were mapped to 33 unique genes. Top pathways enriched by these genes are mostly related to innate immune system, as shown in Table. 4. More specifically, two third of the genes in component 1 are related to signal transduction and 18 of them are related to immune system.

The second largest connected component is a subnetwork centered around protein *FBXO2*. It is found closely related to ubiquitination & proteasome degradation as part of antigen processing (adjusted p=6.456e-25 in pathway enrichment analysis) (Fabregat *et al.*, 2017). Neuronal death in Alzheimer's diseases has a strong connection with misfolded proteins that aggregate within the brain, e.g. amyloid and tau tangles. Ubiquitination & proteasome degradation is one of the two major pathways that help get rid of unwanted cells or misfolded proteins to prevent their accumulation and to maintain the health of a cell (Schmidt *et al.*, 2021).

The third component is a small subnetwork centered around *STX1A* and *SNAP25*. Seven out of 8 unique genes in this subnetworks is found to be related to neurotransmitter release cycle (adjusted p=1.78e-16). More specifically, most of them are involved in the GABA synthesis, release, reuptake and degradation (adjusted p=2.44e-14). GABA has been found to have significantly reduced levels in severe cases of AD (Solas *et al.*, 2015). Selective inhibition of astrocytic GABA synthesis or release has been suggested as a potential therapeutic strategy for treating memory impairment in AD (Jo *et al.*, 2014).

Table 4. Top 10 enriched pathways by genes and proteins in component 1, ranked by adjusted p value.

| Pathway | Overlap | Adjusted P |
|---|---|---|
| Developmental Biology | 17/786 | 2.99E-13 |
| Signal Transduction | 22/2465 | 9.70E-11 |
| Immune System | 18/1547 | 4.31E-10 |
| Innate Immune System | 14/807 | 1.17E-09 |
| Hemostasis | 12/552 | 3.12E-09 |
| Signalling by NGF | 11/450 | 5.88E-09 |
| Platelet activation, signaling and aggregation | 9/253 | 1.31E-08 |
| GPVI-mediated activation cascade | 6/53 | 1.46E-08 |
| Axon guidance | 11/515 | 1.64E-08 |
| C-type lectin receptors (CLRs) | 7/123 | 4.50E-08 |

# 4 Summary

We proposed a new deep graph fusion network to leverage the information flow from DNA to proteins such that gene expression and protein expression data can be denoised and enhanced with improved prediction power. Prior relationship between SNPs, genes and proteins was embedded into the network model as prior knowledge. Edge weight and node importance score learned from MoFNet further helped prune down the prior network where only subnetworks predictive of the disease outcome will be retained. MoFNet showed superior performance over other integrative -omics models in three ways: 1) it jointly models genotype, gene expression, protein expression and their prior functional relationships, 2) it yields subnetworks predictive of outcomes, instead of individual markers with less interpretability, and 3) it enhances gene expression and protein

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.02.490336; this version posted May 3, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.



expression data by leveraging the information flow from DNA to protein. Trans-omic paths from MoFNet findings suggested that AD may be partly the result of genetic variations due to their potential cascading effects on the downstream transcriptome and proteome levels. While none of the prior relationship were extracted in a tissue specific manner, eQTL analysis showed that MoFNet can accurately pick out those tissue specific relationships between SNPs and genes.

It is worth mentioning that the integrative analysis in this paper is only based on a small set of functionally connected multi-omic features, the number of which was further limited because of the bottleneck in protein expression data of the ROS/MAP cohort. Further, due to the incompleteness of multi-omic data, only a portion of participants from the ROS/MAP were involved in this study. Therefore, the classification performance can not reflect the true predictive power of these three types of multi-omic data, which is expected to be much higher if with more -omics features and samples. This study is also limited in that amnestic mild cognitive impairment group (MCI, a transition stage between CN and AD) is excluded from the analysis. How to include this group and investigating whether the predictive power of multi-omic features identified in this paper is stage specific warrants further efforts. In addition, like many current multi-omic models for joint analysis, MoFNet requires concatenation of multi-omic features, leading to exclusion of large chunk of samples. An improved version of MoFNet capable of handling the incomplete multi-omic data will also be of great value to the field.

## Acknowledgements

The results published here are in whole or in part based on data obtained from the AMP-AD Knowledge Portal. Study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA (grants P30AG10161, R01AG15819, R01AG17917, R01AG30146, R01AG36836, U01AG32984, U01AG46152), the Illinois Department of Public Health and the Translational Genomics Research Institute.

## Funding

This research was supported by National Institutes of Health (grants R01 LM013463, R21 AG066135, R21 AG072101, R01 EB022574, R01 AG019771, P30 AG010133), and by National Science Foundation (grants CRII 1755836, CAREER 1942394).

## References

A Bennett, D. *et al.* (2012). Overview and findings from the religious orders study. *Current Alzheimer Research*, **9**(6), 628–645.

Andreev, V. P. *et al.* (2012). Label-free quantitative lc–ms proteomics of alzheimer's disease and normally aged human brains. *Journal of proteome research*, **11**(6), 3053–3067.

Chen, L. *et al.* (2015). glmgraph: an r package for variable selection and predictive modeling of structured genomic data. *Bioinformatics*, **31**(24), 3991–3993.

Cookson, S. *et al.* (2005). Monitoring dynamics of single-cell gene expression over multiple cell cycles. *Molecular Systems Biology*, **1**(1), 2005–0024.

De Jager, P. L. *et al.* (2012). A genome-wide scan for common variants affecting the rate of age-related cognitive decline. *Neurobiology of aging*, **33**(5), 1017–e1.

Fabregat, A. *et al.* (2017). Reactome pathway analysis: a high-performance in-memory approach. *BMC bioinformatics*, **18**(1), 1–9.

Fabregat, A. *et al.* (2018). The reactome pathway knowledgebase. *Nucleic acids research*, **46**(D1), D649–D655.

Grosenick, L. *et al.* (2013). Interpretable whole-brain prediction analysis with graphnet. *NeuroImage*, **72**, 304–321.

Hasin, Y. *et al.* (2017). Multi-omics approaches to disease. *Genome biology*, **18**(1), 1–15.

Hodes, R. J. and Buckholtz, N. (2016). Accelerating medicines partnership: Alzheimer's disease (amp-ad) knowledge portal aids alzheimer's drug discovery through open data sharing. *Expert opinion on therapeutic targets*, **20**(4), 389–391.

Horgusluoglu-Moloch, E. *et al.* (2017). Targeted neurogenesis pathway-based gene analysis identifies adora2a associated with hippocampal volume in mild cognitive impairment and alzheimer's disease. *Neurobiology of aging*, **60**, 92–103.

Huang, S. *et al.* (2017). More is better: recent progress in multi-omics data integration methods. *Frontiers in genetics*, **8**, 84.

Jo, S. *et al.* (2014). Gaba from reactive astrocytes impairs memory in mouse models of alzheimer's disease. *Nature medicine*, **20**(8), 886–896.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., *et al.* (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, **44**(W1), W90–W97.

Kumar, S. *et al.* (2017). Snp2tfbs–a database of regulatory snps affecting predicted transcription factor binding site affinity. *Nucleic acids research*, **45**(D1), D139–D144.

MacArthur, J. *et al.* (2017). The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic acids research*, **45**(D1), D896–D901.

Nativio, R. *et al.* (2020). An integrated multi-omics approach identifies epigenetic alterations associated with alzheimer's disease. *Nature genetics*, **52**(10), 1024–1035.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, **103**(23), 8577–8582.

Nguyen, N. D. *et al.* (2021). Varmole: a biologically drop-connect deep neural network model for prioritizing disease risk variants and genes. *Bioinformatics*, **37**(12), 1772–1775.

Nho, K. *et al.* (2013). Whole-exome sequencing and imaging genetics identify functional variants for rate of change in hippocampal volume in mild cognitive impairment. *Molecular psychiatry*, **18**(7), 781–787.

Pedregosa, F. *et al.* (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, **12**, 2825–2830.

Petyuk, V. A. *et al.* (2010). Mapping protein abundance patterns in the brain using voxelation combined with liquid chromatography and mass spectrometry. *Methods*, **50**(2), 77–84.

Schmidt, M. F. *et al.* (2021). Ubiquitin signalling in neurodegeneration: mechanisms and therapeutic opportunities. *Cell Death & Differentiation*, **28**(2), 570–590.

Slavov, N. (2020). Unpicking the proteome in single cells. *Science*, **367**(6477), 512–513.

Solas, M. *et al.* (2015). Treatment options in alzheimer s disease: the gaba story. *Current pharmaceutical design*, **21**(34), 4960–4971.

Sundararajan, M. *et al.* (2017). Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**(1), 267–288.

Wan, L. *et al.* (2013). Regularization of neural networks using dropconnect. In *International conference on machine learning*, pages 1058–1066. PMLR.

Xie, L. *et al.* (2020). Identification of functionally connected multi-omic biomarkers for alzheimer's disease using modularity-constrained lasso. *Plos one*, **15**(6), e0234748.

Xie, L. *et al.* (2021). Integrative-omics for discovery of network-level disease biomarkers: a case study in alzheimer's disease. *Briefings in Bioinformatics*, **22**(6), bbab121.

Zou, H. and 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.