**Identification of therapeutic targets for Alzheimer’s disease treatment using a deep learning and network analysis based computational framework**

**Daniel Zhang**

**Abstract**

Alzheimer's disease (AD) is the most common type of dementia and 6th leading cause of death in the United States. Although extensive studies have been conducted to understand the root causes of AD, current treatments are unable to stop disease progression. Due to the complexity of AD, identification of effective therapeutic targets requires systematic studies of underlying pathogenic molecular pathways and their interrelationship. The availability of large-scale multi-omics datasets provides opportunities to use computational approaches for such studies. Here, I developed a computational framework using a semi-supervised deep learning approach to predict additional AD-associated genes based on 44 known AD genes, followed by protein-protein interaction (PPI) network-based analysis to identify functionally important AD genes (e.g., master regulators) as potential targets for AD treatment. The deep learning classifier predicted 1,272 AD genes that are functionally enriched in core AD pathogenic molecular pathways. Subsequent PPI network analysis revealed an interrelated network of the AD-associated molecular pathways and their master regulators, some of which such as APP and FYN are well investigated AD treatment targets. Interestingly, two of the master regulators GNB1 and GNAI1 (encoding G protein complex subunits) were barely studied previously in AD context, although they play important roles in neurodevelopment supported by mouse and human genetic evidence. In addition, Kininogen-1 was identified as a novel potential coregulator of APP in amyloid beta related AD pathology. Finally, the computational framework presented in this study provides a valuable tool that can also be used for similar studies in other diseases.

**Introduction**

Alzheimer’s disease (AD) is the most common type of dementia. Approximately 5.8 million people in the U.S. currently have AD. The number of Americans with Alzheimer’s is projected to nearly triple to 14 million by 2060 [ref]. AD is ultimately fatal and is the 6th leading cause of death in the U.S. killing more people than breast cancer and prostate cancer combined [ref, Matthews, K. A.,Alzheimer’s & Dementia].

Although extensive studies have been conducted to understand the root causes of AD, no current treatment is able to stop the disease progression. The presence of amyloid beta (Aβ) plaques, neurofibrillary tangles (NFT), neuronal death and synaptic loss, blood clots and inflammation in the brain have been described as pathological mechanisms of AD. Current therapeutic efforts have largely focused on targeting and reducing the accumulation of Aβ and NFT, but so far have yet to produce desirable outcomes, underscoring the need for novel research approaches. Considering the complexity of AD, systematically understanding the interrelationship of underlying pathogenic molecular pathways, rather than focusing on each individual functional pathway (Aβ plaques, NFT, etc.), is more likely to unbiasedly identify important AD genes as effective therapeutic targets. However, systematic study of AD pathways and their interactions remains challenging due to limited known AD genes, which have been discovered mainly through genome-wide association studies (GWAS) [ref].

Studies have demonstrated that genes associated with the same disease possess similar biological fingerprints including gene expression (Ala et al., 2008), function (Ideker and Sharan, 2008) and physical protein-protein interaction (Brunner et al., 2004; Goh et al., 2007). Thus, with large-scale multi-omics data from patient cohorts and various gene annotation databases becoming available, machine learning has been applied to discover genes that underlie the pathology of diseases including AD [ref]. However, standard unsupervised machine learning is not able to utilize known knowledge (e.g. known disease-associated genes), which often results in suboptimal prediction performance. Moreover, standard supervised binary machine learning, which requires sufficient and well-defined both positively labeled data (e.g. genes associated with the disease) and negatively labeled data (e.g. genes not associated with the disease) as training datasets, is also not suitable for this study, because training datasets for disease gene discovery often consist of a small number of known disease-associated genes (labeled data) and a large number of unknown genes (unlabeled data), of which a small portion may be unidentified disease genes. Therefore, semi-supervised machine learning methods such as baggingPU (ensemble-based) [ref] and ladder\_net (deep learning-based) [ref] have been developed to use unlabeled as well as labeled data to improve classification performance, and demonstrated to outperform conventional unsupervised and supervised machine learning approaches on this type of data [ref]. Remarkably, semi-supervised deep learning approaches have been demonstrated to be able to learn the underlying relationships across a wide array of samples, presenting powerful tools to identify disease genes from diverse and large-scale datasets from patient cohorts.

The association of a gene to a disease cannot necessarily be inferred as a causal relationship. Therefore, experimentally validating the causality of candidate genes to a disease remains crucial to develop successful treatments. However, it can be both costly and time-consuming if a large number of candidates are presented. Various computational approaches have been utilized to prioritize candidate genes for more efficient drug target validation. Among these computational methods, protein-protein interaction (PPI) network-based ranking of the importance of disease genes has been broadly demonstrated as an effective approach, because the biological functional importance of a gene can be estimated by the extent of its connectivity/interactivity to other genes, as well as its influence of information flow to the PPI networks formed by genes with direct or indirect interaction with it, both of which can be assessed by network-based centrality analysis [ref].

In this study, I built a computational framework using semi-supervised deep learning to predict new AD-associated genes, followed by network-based ranking of all predicted and known AD genes for their importance in AD development to systematically understand the molecular pathways underlying AD pathogenesis, as an effort to identify potential novel targets for AD treatment.

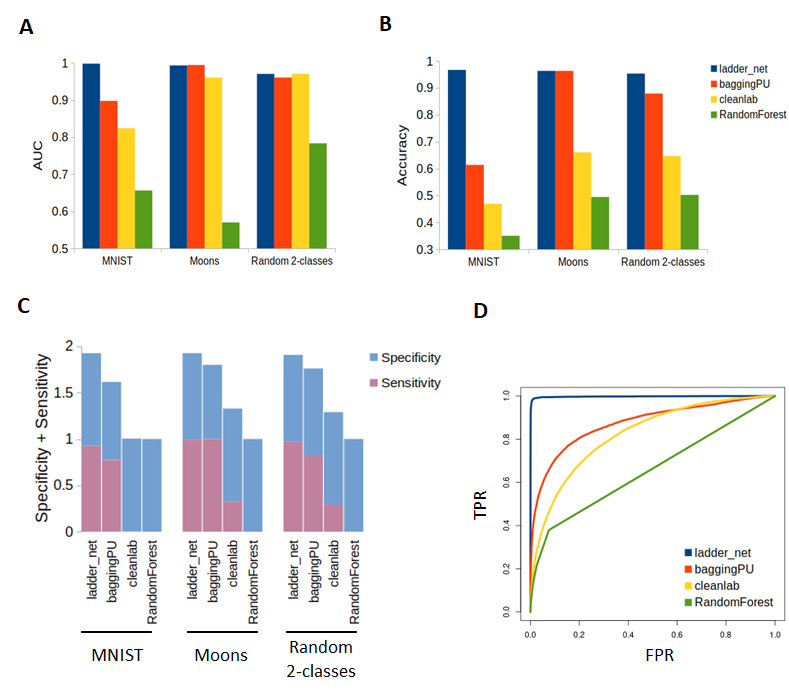
**Results**

**Comparison of semi-supervised machine learning methods**

In order to select an effective classifier for AD gene identification, three semi-supervised machine learning methods were benchmarked against RandomForest, a popular standard binary classifier. The three tested methods represent three different commonly used approaches: baggingPU ̶ ̶ an ensemble approach, cleanlab ̶ ̶ learning with noisy labels, and ladder\_net ̶ ̶ a neural network-based approach [ref].

Three benchmark datasets including a commonly used synthetic dataset for evaluation of binary classification (Moons), a simulated random 2-class dataset (Random 2-classes), and a real-life handwritten digit dataset (MNIST), were used to evaluate the performance of the aforementioned three classifiers. Each dataset was randomly split into a training set (80 % of samples) and a test set (20% of samples).

All except for 100 of the positive samples were masked as unlabeled in the training dataset to mimic the real datasets for disease gene discovery. After training, the prediction performance of the trained model on the corresponding testing dataset was comprehensively evaluated using standard measurements for model performance including AUC (Tr), accuracy, specificity and sensitivity. Unsurprisingly, all three semi-supervised learning classifiers outperform the standard binary classifier RandomForest. Among the three semi-supervised learning methods, the neural network-based classifier ladder\_net has overall best scores of AUC, accuracy, specificity and sensitivity (Figure 1). Therefore, ladder\_net was selected as the disease gene classifier for this study.



**Figure 1.** Comparison of methods for semi-supervised machine learning methods. Three semi-supervised machine learning methods ladder\_net, baggingPU and cleanlab were compared with a standard binary classifier RandomForest using two synthetic datasets Moons and Random 2-classes and a real life dataset MNIST. **A.** AUC (Area Under the Receiver Operating Characteristic Curve) values Prediction performance. **B.** Accuracy of prediction performance. **C.** Specificity and sensitivity of prediction performance. **D.** The Receiver Operating Characteristic Curve of prediction performance of the four methods for benchmarking on the MNIST dataset. FPR: false positive rate. TPR: true positive rate.

**Prediction of AD-associated genes using semi-supervised deep learning**

To build an AD gene prediction model, a large-scale multi-omics dataset was constructed by merging three public datasets, including two gene relationship annotation datasets extracted from Gene Ontology (GO) and PathDIP (signaling pathways) databases [ref], and a curated genome-wide gene expression dataset, which contains 3,401 AD patient samples from brain tissues and 11,428 brain and non-brain tissue samples from healthy people and patients with other common age-related diseases (e.g., Parkinson's disease, Atherosclerosis and Osteoporosis). The data from patients with other age-related diseases were used for better sensitivity and specificity of AD gene identification (Table 2).

Of the genes in the constructed dataset, 44 high confidence AD-associated genes were labeled as known AD genes. These 44 genes were identified from large scale genome-wide association studies (GWAS) and/or collated from Human Phenotype Ontology (HPO) [ref], a database that contains phenotypic information of human genes (Table 1).

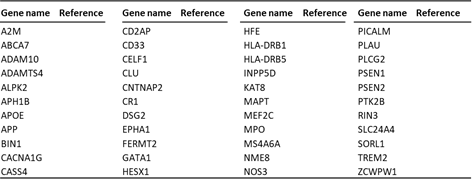
The input data were normalized, standardized and merged as being described in Materials and Methods session, and then were split into training (75% of samples) and testing (25% of samples) datasets to train and evaluate the performance of ladder\_net. After training, the prediction performance of ladder\_net was tested on the testing dataset, and then AUC, accuracy and recall were calculated for performance evaluation. The results (0.9029, 0.8909 and 0.8942 for AUC, accuracy and recall, respectively) demonstrated reasonably good prediction performance for the trained ladder\_net (Figure 2A). The values of AUC and accuracy should be considered as approximate estimation, since only the positive labels (known AD genes) are true in the dataset.

Using the trained ladder\_net model, the probability of each gene being associated with AD was predicted. The distribution of predicted probabilities shows that the known AD genes are highly enriched in the high probabilities end (> 0.75), while the majority of unknown/unlabeled genes are at the low probability end, except for a small group of unknown genes, which present at high probability end (Figure 2B). This justified the use of 0.75, instead of the conventional 0.5, as the threshold for AD gene classification to minimize false positives. In total, 1,272 unknown genes passed the threshold and were considered as potential AD genes. Only a few known AD genes (5, or ~11% of all known AD genes) have prediction probabilities lower than 0.75, which could be due to misclassification or false positives in the annotated AD genes.

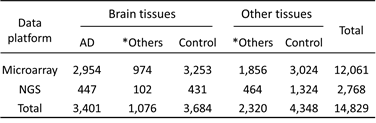
Functional enrichment analysis of the 1,316 genes (1,272 predicted potential AD genes and 44 known AD genes) was performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.8 [ref], and the result indicates that these candidate genes are enriched in the core biology pathways implicated in AD pathogenesis, such as Amyloid precursor protein processing, innate immune response, signaling pathways (Rap1, Ras, Notch, Wnt and PI3K-Akt), regulation of cell division, synapse and blood clotting (Figure 2C). Interestingly, immune and inflammation is the most enriched functional category, suggesting the central role of immune response in AD, which is consistent with the results of recently published large-scale GWAS and functional studies [ref].

Taken together, these results suggest that the predicted AD genes are not randomly selected, and they likely present the overall molecular mechanisms underlying Alzheimer’s disease. However, followup experiments and analyses are required to validate and understand the importance of these genes in AD pathogenesis.

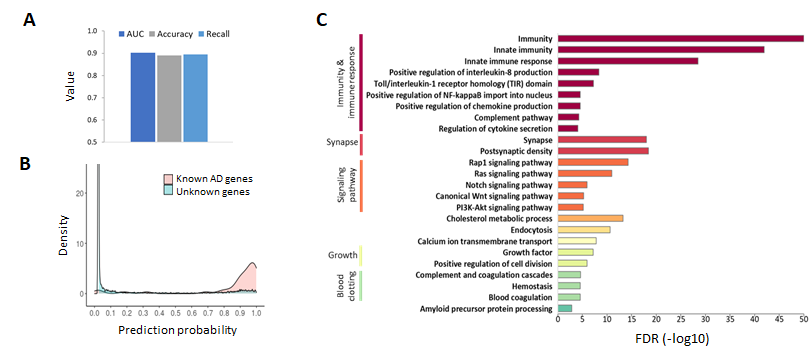
**Table 1.** Curated known Alzheimer’s disease genes



**Table 2**. Summary of collated human gene expression data from public databases



\* Other age related diseases: Dementia, Parkinsonism, Atherosclerosis, Stroke, Osteoarthritis, Osteoporosis, Chronic obstructive pulmonary disease and Type 2 diabetes



**Figure 2.** Identification of AD genes. **A.** Performance of ladder\_net on the dataset for AD gene prediction. Since only the positive labels (known AD genes) are true in the dataset, the values of AUC and Accuracy should be considered as approximate estimation **B.** The distribution of prediction probability of genes being associated with AD. **C.** Functional enrichment of predicted and known AD genes. DAVID (Database for Annotation, Visualization and Integrated Discovery) v6.8 was used for the enrichment analysis. FDR: false discovery rate.

**Prioritization of AD-associated genes using PPI network-based computational approach**

Using computational approaches to prioritize disease genes has been demonstrated to be valuable as a fast and low-cost approach for preliminary functional validation of genes of interest [ref]. Among such computational methods, the PPI network-based approach is able to uncover the causal relationships between genes and disease pathogenesis, and therefore has been broadly used for ranking the importance of disease genes [ref]. I used PPI database STRING to build an AD gene interaction network. In total, 1,120 out of the 1,316 AD genes (1,272 predicted and 44 known AD genes) have protein-protein interaction data in the STRING database. Within the network of the 1,120 genes, any two genes with at least one directional interaction with highest confidence (confidence score >= 900) are connected by an edge to form an AD gene interaction network. In total, 500 genes are selected in this network for further analysis.

The importance of a node in a network is often measured by centrality analysis [ref]. Ashtiani M. et al. has reported that the Latora closeness centrality method gives the best overall result after comprehensively benchmarking current available centrality measures using various types of data [ref]. Thus, Latora closeness was used to calculate the centrality scores of the 500 AD genes included in the AD gene interaction network. Subsequently, these genes were ranked for their importance in AD according to their centrality scores ̶ the higher the centrality score of a gene, the more important the gene in theAD gene interaction network (Table 3).

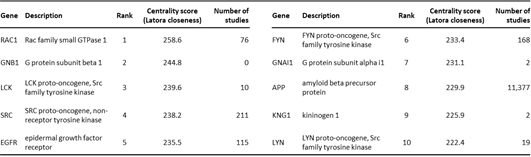
As a further validation of the predicted and top ranked AD genes, a PubMed search using advanced filters (geneName[Title/Abstract]) AND (Alzheimer[Title/Abstract]) was performed to examine how well the top ranked genes have been studied in an AD context. The result (Table 3) indicates that APP (amyloid beta precursor protein)—the protein basis of AD pathology hallmark amyloid plaques and the target of FDA newly approved AD drug Aduhelm ̶ ̶ unsurprisingly, has been studied and reported 11,377 times. Some other top ranked genes, such as SRC family kinases (SRC, FYN, LCK and LYN), EGFR (epidermal growth factor receptor) and RAC1 (Rac family small GTPase 1), have generally good study coverage and have been investigated as potential AD treatment targets in many preclinical and clinical studies [ref]. Table 3 also indicates that three of the top ranked genes—GNB1 (G protein subunit beta 1) and GNAI1 (G protein subunit alpha i1), encoding two of the three G protein complex subunits, and KNG1 (kininogen 1), involved in blood coagulation and immune response—were barely studied, suggesting a potential novelty in this list of top ranked genes.

In order to further understand the interactions among the top 10 ranked genes, additional examination and visualization of the PPI network were performed using Cytoscape [ref]. Figure 3A shows that the network consists of several connected subnetworks, in which the top ranked genes act as hubs/master regulators. Functional enrichment analysis of each subnetwork reveals that these subnetworks are associated with functions implicated in AD pathogenesis [ref]. For example, RAC1 interacts with a group of genes enriched in actin cytoskeleton organization (e.g., FGD4, RAC2, RHOU, WASF2, WASF3, FGD2 and PAFAH1B1) and neurogenesis (e.g., NGFR, CYFIP1, SLIT2, SRGAP2, PAFAH1B1 and ROBO1), which have been linked to the AD pathology of neurodegeneration [ref]. As expected, the interaction of APP with genes involved in amyloid deposit (e.g., PSEN1 and ADAM10) is also observed (Figure 3A).

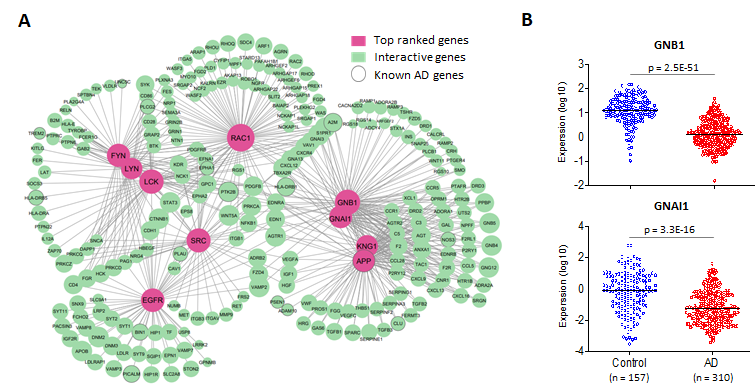
Surprisingly, KNG1 interacts with many genes that also interact with APP (Figure 3A), which are involved in blood clotting (e.g., SERPINE1, FERMT3, FGG, VWF, PROS1, HRG and THBS1) and inflammatory response (e.g., CCR1, CXCL9, CXCL13, C3, C5, GAL, CCL5 and CCR5), the two AD pathologies reportedly associated with amyloid plaques [ref]. This observation suggests that combination treatments targeting KNG1 and APP may be a better approach than targeting APP alone.

Intriguingly, the G protein genes GNB1 and GNAI1 serve as the hubs of a subnetwork interacting with two other subnetworks (RAC1 and APP-KNG1 as the hubs) associated with the core AD pathologies, implying that GNB1 and GNAI1 could play a role in AD pathogenesis (Figure 3A). Although little is known about their functions in AD context, Okae H. et al. reported that Gnb1 knockout causes neural tube defects and impaired neural progenitor cell proliferation in mouse brains [ref], and in humans, variants in GNB1 and GNAI1 have been associated with a severe neurodevelopmental disorder [ref]. In addition, G-protein signaling pathways are functionally involved in neurodevelopment, and GNB1 and GNAI1 are important regulators of G-protein signaling pathways [ref], as also evidenced by their role in regulating a group of G-protein signaling pathway genes (e.g., GNG12, ADRA2A, SMO, P2RY1, GNB5, AGTR2 and DRD3) in the subnetworks (Figure 3A). [ref]. Furthermore, the mRNA expression of GNB1 and GNAI1 is significantly downregulated in AD patient brains (p<0.0001) compared to healthy controls’ (Figure 3B), which is in line with the previously reported significant downregulation of GNB1 protein in AD patient prefrontal cortex, the brain part involved in memory [ref]. Collectively, these results suggest that GNB1 and GNAI1 can potentially be promising novel AD treatment targets.

**Table 3.** Top ranked AD genes



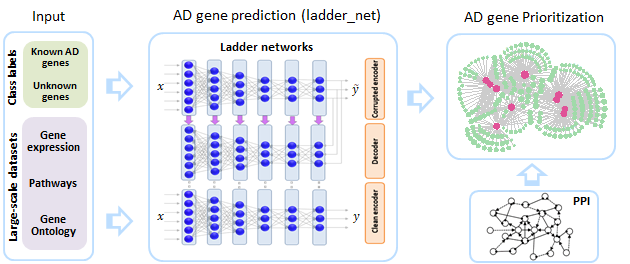
PubMed search performed on 9/6/2021



**Figure 3.** Prioritization of AD genes. **A.** The protein-protein interaction network of the top 10 ranked AD genes. Identified AD genes (in green) that have reliable interaction with the top 10 ranked AD genes (in red) are included in this network. Within the network, any two genes that have reliable interaction are connected by a gray line. Reliable interaction: directional and highest confidence (confidence score > 900) interaction. Node sizes reflect AD genes’ importance. The bigger, the more important. **B.** The expression of GNB1 and GNAI1 in AD patient and healthy control brain tissues (prefrontal cortex). Expression data of GNB1 and GNAI1 were extracted from a public dataset GSE33000 in NCBI (National Center for Biotechnology Information) GEO (Gene Expression Omnibus) database.

**Summary and discussion**

In this study, I presented a computational framework for the discovery of potential novel targets for AD treatment (Figure 4). This framework takes known AD genes and large-scale multi-omics datasets as input to predict new AD-associated genes using semi-supervised deep learning method ladder\_net, and then identify genes that are essential to AD pathogenesis as potential treatment targets using PPI network-based disease gene prioritization. The framework developed for this study is able to systematically identify AD associated genes and assess their importance by using genome-wide datasets as inputs, thus avoiding the bias that can potentially be introduced by traditional approaches focusing on individual genes or functional pathways.



**Figure 4.** A computational framework for therapeutic target discovery in Alzheimer’s disease treatment. The large-scale multi-omics datasets with genes labeled as known AD genes and unknown genes were input into the semi-supervised deep learning classifier ladder\_net to predict additional/new AD associated genes. Then, these predicted and known AD genes were prioritized using PPI network-based ranking to identify potential therapeutic targets for AD treatment.

This study systematically reveals molecular pathways and their interactions underlying AD pathogenesis, and presents a comprehensive molecular landscape of AD associated genes. The network analysis of AD disease genes suggests that more than one master regulators can contribute to a single pathology associated with AD. Moreover, different AD pathologies are connected at molecular levels. Collectively, these results suggest the importance of combination therapy for AD.

In this study, KNG1 was identified as a potentially important coregulator of APP in Aβ related AD pathologies, suggesting that KNG1 could be a new candidate for AD treatments in combination with APP. More interestingly, GNB1 and GNAI1, which were barely studied in AD context but reportedly play important roles in neurodevelopment, were identified as two of the top ranked AD genes, warranting further study and validation as potential new targets for AD treatments.

Last but not least, the framework presented in this study can be easily adapted for identifying therapeutic targets in other diseases by simply replacing the known AD genes and the AD-specific gene expression dataset with any disease of interest, as long as the input datasets are large and diverse enough to capture possible gene-gene and gene-disease associations to ensure that robust results can be generated.

**Materials and Methods**

**Datasets for comparison of prediction methods**

**MNIST**: a dataset of 70,000 handwritten digits of 28-by-28 pixel size. The raw pixel values are normalized to a range 0 to 1 and flattened to a vector of 784 dimensions. Link for downloading the dataset: <http://yann.lecun.com/exdb/mnist/>.

**Moons**: make\_moons from the scikit-learn Python machine learning module was used to simulate a 2-D Moons dataset with the customized values for parameters: n\_samples=75,000, noise=0.1 and random\_state=1. Default values for other parameters were used.

**Random 2-classes**: make\_classification from the scikit-learn Python module was used to simulate a 2-classes dataset with the customized values for parameters: n\_samples=75,000, n\_features=10 and random\_state=1. Default values for other parameters were used.

**Machine learning methods for the benchmark**

**RandomForestClassifier**: Random Forest is a popular method for binary classification. The RandomForestClassifier from the scikit-learn Python module was used as a baseline performance for benchmarking binary classifiers. Default parameter values were used for benchmarking, except the following custom values: n\_estimators=1,000 and n\_jobs=-1.

**BaggingClassifierPU**: BaggingClassifier from the scikit-learn Python module is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions to form a final prediction. It encompasses several works from the literature [ref, check sklearn web]. BaggingClassifierPU is a more user-friendly adaptation version of BaggingClassifier (<https://github.com/roywright/pu_learning>). Default parameter values were used for the comparison, except the custom values: n\_estimators=1,000 and n\_jobs=-1.

**cleanlab**: cleanlab (<https://github.com/cleanlab/cleanlab>) is a Python package for machine learning with noisy labels using algorithms called confident learning [ref, *cite confident learning*]. The LearningWithNoisyLabels function with default parameter values from cleanlab was used for benchmarking.

**ladder\_net**: ladder\_net (https://github.com/divamgupta/ladder\_network\_keras) is a recent implementation in Keras of Ladder Networks, a neural networks-based model for semi-supervised learning [ref]. Ladder\_net was originally designed for multinomial classification. To adapt to binary classification, ladder\_net was modified by replacing the activation function of the last layer softmax with sigmoid, and the loss function categorical\_crossentropy with binary\_crossentropy.

**Large-scale datasets for AD gene prediction**

**Gene Ontology (GO)**: the complete dataset of human GO annotation file (goa\_human.gaf) was downloaded from GENEONTOLOGY website (<http://geneontology.org/docs/download-go-annotations/>).

**PathDIP**: Pathway-gene association information was collected from the PathDIP v.4 database (<http://ophid.utoronto.ca/pathDIP/>).

The downloaded Gene Ontology and PathDIP datasets were transformed to binary matrices, in which rows are genes and columns are GO terms or pathways, respectively.

**Gene Expression**: the gene expression data from AD patients, patients with other age-related diseases (Dementia, Parkinsonism, Atherosclerosis, Stroke, Osteoarthritis, Osteoporosis, Chronic obstructive pulmonary disease and Type 2 diabetes) and their corresponding healthy controls were collected from NCBI GEO database (<https://www.ncbi.nlm.nih.gov/gds/>) by searching ((diseaseName[MeSH Terms] OR diseaseName[All Fields]) AND ("gene expression"[MeSH Terms] OR gene expression[All Fields]) AND ("hominidae"[MeSH Terms] OR "Homo"[Organism] OR homo[All Fields])) AND "Homo sapiens"[porgn]. Eventually, 133 GEO datasets generated using microarray or next generation sequencing (NGS) platforms were collated for this study. Sequencing read counts data from NGS datasets were transformed to TPM (transcripts per kilobase million) first. Then NGS and microarray data were centered and scaled (0 to 1) by sample.

Finally, Gene Ontology, PathDIP and all Gene Expression data were merged by gene, and then genes with at least quarter of non-null values across all the samples were selected. This data process gives a 13,445 by 30,419 gene expression matrix as prediction input.

**Building Protein-protein interaction (PPI) networks and centrality analysis**

The PPI dataset (STRING 10.5 protein.actions restricted to Homo sapiens) was downloaded from the STRING database (<https://version-10-5.string-db.org/cgi/download.pl?species_text=Homo+sapiens>). The STRING PPIs that are directional and highest confident (confidence score >= 900) were selected, and then the selected PPIs were mapped to the AD gene set (1,272 predicted potential AD genes and 44 known AD genes). This process results in an AD gene interaction network containing 500 genes.

R package igraph [ref] was used to transform the AD gene network as an igraph object. Latora closeness centrality function from CINNA R package [ref] was used to calculate the centrality score of these AD genes using the igraph object as input.