DeepDrug2: A Germline-focused Graph Neural Network Framework for Alzheimer's Drug Repurposing Validated by Electronic Health Records

Victor OK Li[1*], Yang Han[1], Jacqueline CK Lam[1*], Jocelyn Downey
Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong

[1]Equal contributions.
[*]Corresponding and senior authors: Victor OK Li and Jacqueline CK Lam.
Email: vli@eee.hku.hk and jcklam@eee.hku.hk

Keywords: Alzheimer's disease, neurodegeneration, expert-led AI drug repurposing, germline mutation, graph neural network, directed biomedical graph, drug validation, electronic health records.

Abstract

Alzheimer's disease (AD) is a complex neurodegenerative disorder with limited therapeutic options. The original DeepDrug framework by Li et al. (2025) relied on somatic mutation data and emphasized long genes to guide AD drug repurposing. However, emerging evidence suggests that germline genetic variants play a more central role in AD pathogenesis. In response, we develop DeepDrug2, an enhanced AI-driven framework for AD drug repurposing centered on germline mutations and validated using real-world electronic health records (EHRs). DeepDrug2 introduces four major innovations. First, it proposes a different hypothesis prioritizing germline over somatic mutations in influencing AD risk. Second, it updates the signed directed heterogeneous biomedical graph by removing somatic mutations, long genes, and expert-led genes from the previous version, and incorporating new genes identified in recent genome-wide association study (GWAS) findings. Third, it generates a new list of drug candidates by encoding this updated graph into a new embedding space via a graph neural network (GNN) and calculating drug-gene scores. Fourth, it performs real-world clinical validation using EHR data from over 500,000 individuals (including more than 4,000 AD cases) in the UK Biobank, evaluating associations between drug usage and AD onset while controlling for demographic and comorbidity factors. DeepDrug2 has identified several promising drug candidates. Among the top 15 candidates with sufficient medication

records to support statistically powered analysis, Amlodipine (a calcium channel blocker), Indapamide (a thiazide-like diuretic), and Atorvastatin (a statin) were significantly associated with reduced AD risk ($p < 0.05$). These findings highlight the role of germline mutations in guiding AD drug repurposing and emphasize the value of integrating real-world clinical data into AI-driven drug discovery. To further validate these candidates, future work will involve experimental studies using mouse and zebrafish models of AD. DeepDrug2 offers a promising strategy to support future clinical studies and expand therapeutic options for AD.

1. Introduction

Alzheimer's disease (AD) significantly deteriorates human health and quality of life, affecting approximately 50 million people globally. Despite decades of research, no effective disease-modifying treatment or preventive therapy for AD has been established. Drug development remains a lengthy and expensive process, requiring an estimated USD $2.6 billion to bring a single drug candidate to regulatory approval [1]. To date, only six drugs (Donepezil, Rivastigmine, Galantamine, Memantine, Aducanumab, and Lecanemab) and one drug combination comprising two of these six drugs (Donepezil and Memantine) have been approved by the US Food and Drug Administration (FDA) for AD treatment, with most therapies aimed at symptom management rather than modifying the underlying disease process [2].

Given these challenges, drug repurposing, i.e., identifying new therapeutic uses for existing approved drugs, has emerged as a promising strategy to accelerate drug discovery, offering a faster, lower-risk path to AD treatments [3, 4]. Computational methods have increasingly driven drug repurposing efforts through biomedical data. Early AD drug repurposing strategies primarily utilized statistical analyses of molecular data. A widely used technique is transcriptomic signature matching, which involves deriving differential gene expression signatures from AD versus control samples and identifying drugs whose perturbation profiles inversely correlate with the disease signature [5]. However, signature matching methods are limited by the quality and availability of gene expression data and often overlook disease mechanisms that operate beyond transcriptomic changes, particularly those embedded in molecular interaction networks. In contrast, network-based approaches emerged to model the complex molecular landscape of AD. These methods often integrate AD-associated genes, protein-protein interaction (PPI) networks, and drug-target data to assess the network proximity of drugs to disease modules [6]. However, traditional network

proximity methods do not fully exploit the high-dimensional and non-linear relationships in biomedical networks.

Deep neural networks, particularly graph neural networks (GNNs), have recently been applied to model the complexity of biomedical systems for drug repurposing [7]. GNNs learn node representations (i.e., vector embeddings) by iteratively aggregating information from their neighbors, enabling the capture of both local and global structures within biomedical graphs. Biomedical entities, such as drugs, targets, pathways, and diseases, can be represented as nodes, with GNNs used to infer new edges, such as potential drug-disease associations, based on the learned representations. Several studies have tailored GNN models specifically for AD drug repurposing. NETTAG integrated genome-wide association study (GWAS) data and multi-omics evidence to predict AD risk genes and druggable targets through a GNN model designed to account for the sparse and modular topology of PPI networks [8]. A graph autoencoder model embedded nodes from an AD knowledge graph composed of drugs, genes, pathways, and gene ontology terms, facilitating the identification of AD drug candidates based on multi-level efficacy evidence, ranging from transcriptomic signatures to clinical trials [9]. DeepDrug presented the state-of-the-art advancements in AI-driven AD drug repurposing, incorporating expert-led domain-specific knowledge, such as long genes and somatic mutations detected in both blood and brain tissues [10-12], into a signed directed heterogeneous biomedical graph, enabling a GNN model to capture fine-grained drug-gene relationships and prioritize AD drug candidates validated through AD-related literature [13, 14].

However, despite its innovations in AI-driven AD drug repurposing, DeepDrug emphasized somatic mutations, which recent genetic and clinical research suggests are less central to sporadic (i.e., late onset) AD pathogenesis [15-18]. Large-scale GWAS studies have identified numerous germline variants significantly associated with AD risk, implicating core pathways involved in neuroinflammation, aging, and immune dysfunction [19-21]. These findings suggest that AD drug repurposing should prioritize germline-driven mechanisms, which better reflect the nature of AD pathophysiology, rather than relying predominantly on somatic mutation data. Furthermore, DeepDrug lacked validation using real-world clinical data, limiting its ability to assess whether the identified drug candidates can be translated into meaningful outcomes in patient populations.

In this study, we present DeepDrug2, an enhanced AI-driven framework for AD drug repurposing centered on germline mutations and validated using real-world electronic health

records (EHRs). DeepDrug2 builds upon the strengths of the original DeepDrug but introduces four innovations (see Figure 1). First, it proposes a different hypothesis prioritizing germline over somatic mutations in influencing AD risk. Second, it updates the signed directed heterogeneous biomedical graph by removing somatic mutations, long genes, and expert-led genes from the previous version, and incorporating new genes identified in recent GWAS findings. Third, it generates a new list of drug candidates by encoding this updated graph into a new embedding space via a GNN model. Fourth, it performs real-world clinical validation using EHR data from over 500,000 individuals (including more than 4,000 AD cases) in the UK Biobank, evaluating associations between drug usage and AD onset while controlling for demographic and comorbidity factors.

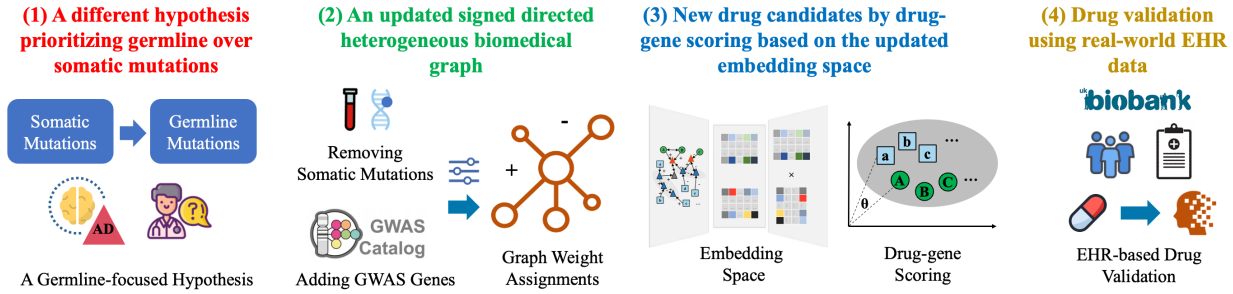

Figure 1. Novelties of DeepDrug2 for AD Drug Repurposing

DeepDrug2 has identified several promising AD drug candidates. Among the top 15 candidates with sufficient medication records to support statistically powered analysis, Amlodipine (a calcium channel blocker), Indapamide (a thiazide-like diuretic), and Atorvastatin (a statin) were significantly associated with reduced AD risk ($p < 0.05$). These findings highlight the role of germline mutations in guiding AD drug repurposing and emphasize the value of integrating real-world clinical data into AI-driven drug discovery. To further validate these candidates, future work will involve experimental studies using mouse and zebrafish models of AD. DeepDrug2 offers a promising strategy to support future clinical studies and expand therapeutic options for AD.

2. Data and Methodology

Our proposed DeepDrug2 methodology has five steps. First, we collected relevant biomedical graph and clinical data (Section 2.1). Second, we constructed a heterogeneous directed biomedical

graph encoding AD-related genes, proteins, drug targets, and drugs, capturing the core network properties of AD pathophysiology (Section 2.2). Third, we used a graph neural network to learn low-dimensional embeddings of drugs and genes, preserving the complex relationships inherent in the biomedical graph (Section 2.3). Fourth, we prioritized drug candidates based on the drug-gene scores calculated using the GNN-based embeddings (Section 2.4). Finally, we validated the selected drug candidates using EHRs to assess real-world clinical outcomes (Section 2.5). The details are described as follows.

2.1 Data Collection

We collected biomedical graph data on four node types, including genes, proteins, drug targets, and drugs, and three types of directed edges, including protein-protein interactions, drug-target interactions, and drug-drug interactions. Gene data were derived from a recent AD-focused GWAS study [20]. Drug and drug-drug interaction data were obtained from DrugBank [22]. Drug target information was aggregated from DrugBank [22], DrugCentral [23], ChEMBL [24], and BindingDB [25]. Protein-protein interaction and pathway data were sourced from the STRING database [26].

In addition to molecular and pharmacological data, we collected real-world clinical data from the UK Biobank to support post-hoc validation of drug candidates identified by DeepDrug2. The UK Biobank is a large-scale biomedical database containing health information from approximately 500,000 participants aged 40-69 years at recruitment [27]. For this study, we analyzed data from 502,163 individuals, including 4,451 participants diagnosed with AD. Medication usage data were extracted and mapped to the standardized drug names used in the biomedical graph, with brand name variations harmonized using DrugBank records [22]. AD diagnosis was based on ICD-9 and ICD-10 codes in hospital admission records. Clinical covariates, including age, gender, ethnicity, and key comorbidities (hypertension, diabetes, and coronary artery disease (CAD)) were extracted to control for potential confounding factors in statistical analysis [8].

2.2 Heterogeneous Signed Directed Biomedical Graph Construction

After data collection, we constructed a heterogeneous, signed, and directed biomedical graph to represent biologically meaningful interactions relevant to AD. The graph consisted of four distinct

node types, including genes, proteins, drug targets, and drugs, and four types of directed edges, including gene-protein, protein-protein, drug-target, and drug-drug interactions. To improve computational efficiency and focus the analysis on biologically relevant structures, isolated nodes were removed from the final graph.

Following the methodology established in the DeepDrug framework [14], graph construction started from identifying key genes implicated in AD and approved drugs relevant to AD, guided by expert-led domain-specific knowledge. Drug targets bound to the selected drugs were then connected to the proteins corresponding to the identified genes through PPIs. The identified genes also underwent pathway enrichment analysis to determine their roles in AD-associated pathways in the PPI network. Finally, expert-guided knowledge was integrated into the biomedical graph by determining edge directions (e.g., whether one protein acts on another protein) and signs (e.g., whether a drug activates or inhibits its target) and assigning weights to nodes and edges (e.g., gene nodes weighted by their importance to AD based on statistical significance from GWAS analysis).

The original biomedical graph developed in DeepDrug [14] was further updated to reflect the updated hypothesis of AD pathophysiology, emphasizing germline over somatic mutations in influencing AD risk. The three major changes are listed as follows:

1. Removal of somatic mutation data: All somatic mutation features, long genes, and previously expert-curated gene sets focusing on somatic alterations were excluded based on recent GWAS evidence highlighting the predominance of germline mutations in contributing to AD risk [20].
2. Integration of germline-based AD genes: We incorporated 107 AD-associated genes identified from the recent GWAS findings [20]. Each gene node was weighted by its normalized -log(p) value to reflect the strength of GWAS association.
3. Updated pathway enrichment analysis: Using the STRING pathway analysis tool [26], we identified 11 statistically significant AD-related pathways (FDR < 0.05) from the KEGG database [28] based on the 107 AD-associated genes. For PPI edges involving proteins in these pathways, edge weights were further increased based on normalized -log(FDR) values to emphasize biologically relevant pathways in the graph.

The resulting graph captures the hierarchical flow from AD-associated genes through proteins and drug targets to candidate drugs, enabling advanced graph representation learning for further analysis. Figure 2 displays a simplified visualization of the updated DeepDrug2 biomedical graph,

highlighting interactions among the top 50 nodes from each node category and showing the largest connected component in the graph.

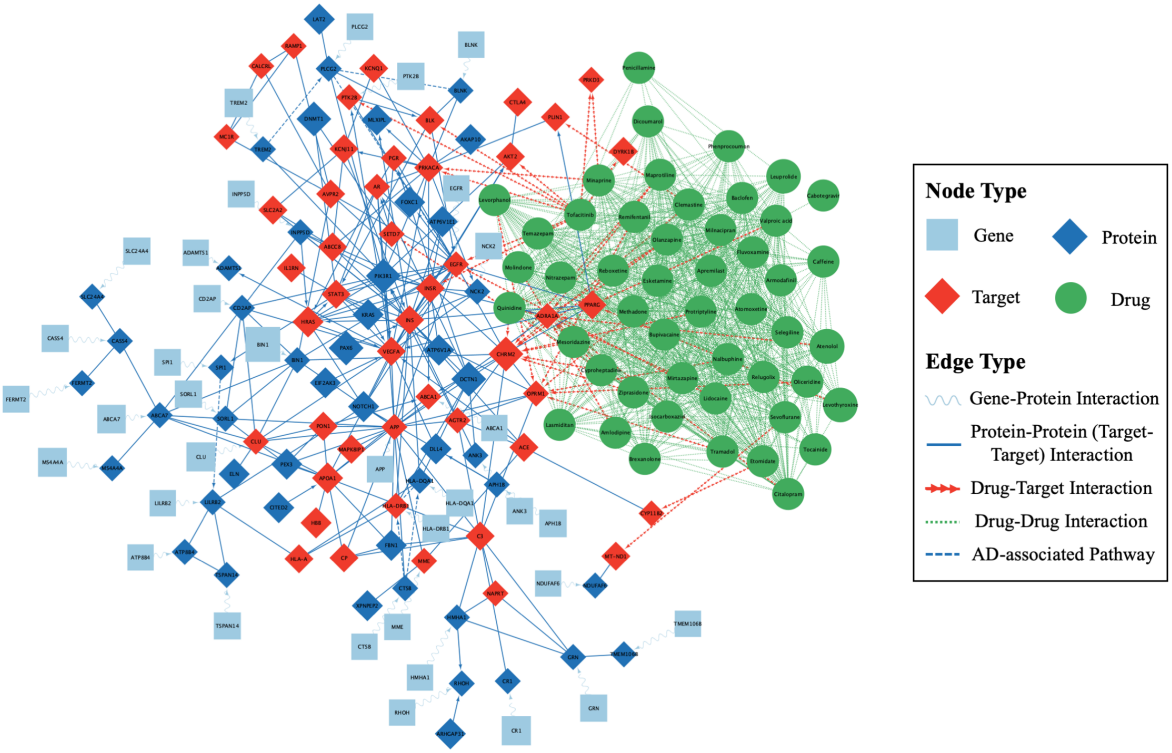

Figure 2. DeepDrug2 Heterogeneous Signed Directed Biomedical Graph (Simplified Version)

2.3 Graph Neural Network Model Development

After the biomedical graph construction, we used the signed directed GNN model proposed in the DeepDrug framework [14] to embed biomedical nodes, including genes and drugs, into a unified low-dimensional embedding space. The GNN framework is based on a graph autoencoder (GAE) architecture, which learns node embeddings by encoding both node features and the signed directed adjacency structure of the biomedical graph. Specifically, the encoder uses signed and directed graph convolutional network (GCN) layers to generate positive and negative embeddings for each node, which are then combined into source and target embeddings. The decoder reconstructs the network by optimizing an edge probability matrix derived from these embeddings. Conceptually, the GNN framework is described as follows:

$$\boldsymbol{Z_p}, \boldsymbol{Z_n} = \text{SignedGCN}\left(\boldsymbol{X}, \boldsymbol{A_p}, \boldsymbol{A_n}\right) \tag{1}$$

$$\boldsymbol{Z_{sp}}, \boldsymbol{Z_{tp}} = \text{DirectedGCN}(\boldsymbol{Z_p}, \boldsymbol{A_p}) \tag{2}$$

$$\boldsymbol{Z_{sn}}, \boldsymbol{Z_{tn}} = \text{DirectedGCN}(\boldsymbol{Z_n}, \boldsymbol{A_n}) \tag{3}$$

$$\boldsymbol{Z_s} = \text{Concatenate}(\boldsymbol{Z_{sp}}, \boldsymbol{Z_{sn}}) \tag{4}$$

$$\boldsymbol{Z_t} = \text{Concatenate}(\boldsymbol{Z_{tp}}, \boldsymbol{Z_{tn}}) \tag{5}$$

$$\boldsymbol{A'} = \sigma(\boldsymbol{Z_s}\boldsymbol{Z_t}^T) \tag{6}$$

where

SignedGCN is a signed GCN with the ReLU activation function,

DirectedGCN is a directed GCN with the ReLU activation function,

$\boldsymbol{Z}$ represents low-dimensional node embeddings,

σ is the sigmoid activation function,

and $\boldsymbol{A'}$ is the edge probability matrix of all node pairs.

The GNN model was trained using a link prediction task, framing the problem as binary classification to estimate the likelihood of an edge between any two nodes, and enabling drug candidates to be prioritized based on their proximity to AD-associated genes. To enhance biological relevance, pathway-guided regularization was applied. This included two auxiliary classification losses based on node embeddings: (1) node type classification (cross-entropy loss), ensuring each node maps to one of the four predefined types (gene, protein, drug target, drug) and (2) pathway membership classification (binary cross-entropy loss), allowing nodes to belong to zero or more pathway, preserving the network structure and the functional convergence/divergence of AD-related pathways in the embedding space.

2.4 Drug Candidate Scoring and Selection

Following model training and validation, we used the trained GNN to generate embedding vectors for all drug and gene nodes in the biomedical graph. To prioritize drug candidates, we computed drug-gene scores based on the proximity of each drug to AD-associated genes in the embedding space. A higher score indicated a drug was more closely related to AD genes. Incorporating the direction and sign of biomedical interactions, we quantified drug-gene relationships based on the anti-correlation of embeddings, prioritizing drug inhibitors close to AD-risk genes and drug activators close to AD-protective genes:

$$\text{Score}_{d,g} = \text{cosine}\left(Z_{s,d}, -Z_{s,g}\right) \tag{7}$$

where

d represents one drug,

g represents one gene,

and Z_s is the node's embedding vector accounting for the drug-target direction or gene-protein direction.

Next, we calculated the score of a single drug candidate based on the average drug-gene scores across all AD-related genes:

$$\text{Score}_d = \frac{1}{N}\sum_{g} \text{Score}_{d,g} \tag{8}$$

where

d represents one drug,

g represents one gene,

and N is the total number AD-associated genes.

Further, we selected the top K single drug candidates based on their drug-gene scores. These candidates were subsequently validated using real-world EHR data.

2.5 Drug Validation Using Electronic Health Records

To assess the real-world effectiveness of the drug candidates identified by DeepDrug2, we performed post-hoc validation using EHR data from the UK Biobank. Specifically, we employed a logistic regression model to evaluate the association between candidate drug usage and the risk of AD. In this model, AD diagnosis served as the binary outcome variable, while the primary exposure variable was binary drug usage (1 = exposed; 0 = unexposed), indicating whether an individual had a documented medication history of using the drug. To account for potential confounding effects, the model included six demographic and clinical covariates, including age, gender, ethnicity, and comorbidities known to be associated with AD risk, including hypertension,

diabetes, and CAD [8]. Each candidate drug was evaluated independently. To ensure statistical robustness, we included only those drugs for which at least 70 individuals were classified as exposed, following the rule of thumb that regression analysis requires a minimum of ten observations per independent variable. Conceptually, the statistical model is described as follows:

$$\begin{aligned}\text{Logit}(P(\text{AD})) = \beta_0 + \beta_1 \cdot \text{Drug Usage} + \beta_2 \cdot \text{Age} + \beta_3 \cdot \text{Gender} + \beta_4 \cdot \text{Ethnicity} \\ + \beta_5 \cdot \text{Hypertension} + \beta_6 \cdot \text{Diabetes} + \beta_7 \cdot \text{CAD} + \varepsilon \end{aligned} \quad (9)$$

where

$P(\text{AD})$ is the probability of AD diagnosis,

β_0 is the intercept,

$\beta_1, \ldots, \beta_7$ are the regression coefficients for each variable,

and ε is the error term.

The coefficient β_1 represents the estimated effect of drug exposure on AD risk, controlling for age, gender, ethnicity, and comorbidities. Specifically, β_1 quantifies the log odds change in the probability of an AD diagnosis associated with exposure to a specific drug, while holding all other covariates constant. To facilitate interpretation, β_1 was exponentiated to obtain the odds ratio (OR): $\text{OR} = e^{\beta_1}$. If OR > 1, drug exposure increases the risk of AD. If OR < 1, drug exposure decreases the risk of AD. If OR = 1, drug exposure has no effect on AD risk.

3. Results

3.1 Experimental Setup

We used a random 80/10/10 split of all existing edges on the graph as the training set, validation set, and test set. The training set was used for the GNN model training. The validation set was used for hyperparameter selection and early stopping during training. After model training, the test set was used for performance evaluation. The optimized node embeddings were used to classify the existence of edges on the test set, using the Area under the ROC Curve (AUC) and Average Precision (AP) as the model performance evaluation metrics, which are two commonly used evaluation metrics for link prediction [29]. The AUC and AP metrics measure how well the model can distinguish existing edges and non-existent edges on the biomedical graph. AUC ranges from

0 to 1, with an AUC of 1 indicating perfect performance. AP is the area under the precision-recall curve, summarizing the trade-off between precision and recall at various thresholds into a single value. AP ranges from 0 to 1, with an AP of 1 indicating perfect performance.

The hyperparameter settings are listed as follows. The dimension of node feature vectors was 260, including four dimensions from the expert-guided node weight vector (representing gene, protein, drug, and target nodes) and 256 dimensions derived from the corresponding biological structure (gene sequence, protein sequence, or drug compound). Each GCN in the encoder had two layers, and the hidden unit size was 128. The embedding dimension size was selected empirically and set to 32. The number of training epochs was 100. The Adam optimizer was used to learn the model parameters, with a learning rate of 0.01. For the directed GCNs, two hyperparameters α and β were used ([0.4, 0.6], [0.5, 0.5], or [0.6, 0.4]) to control the relative importance of source and target information. The weight of the pathway-guided regularization term was 0.1, 0.5, or 1.0.

3.2 Performance Evaluation

As shown in Figure 3, the signed directed GNN model demonstrates exceptional performance, achieving 98% classification accuracy, with an AUC score of 0.98 and an AP score of 0.98. These high metrics underscore the model's ability to effectively capture and represent the complex biomedical graph structure, accurately modeling the relationships between drug and gene nodes. This performance indicates that the model is not only robust in its ability to classify drug-gene interactions but also highly proficient in preserving the intricate dependencies and interactions inherent to AD-related molecular networks.

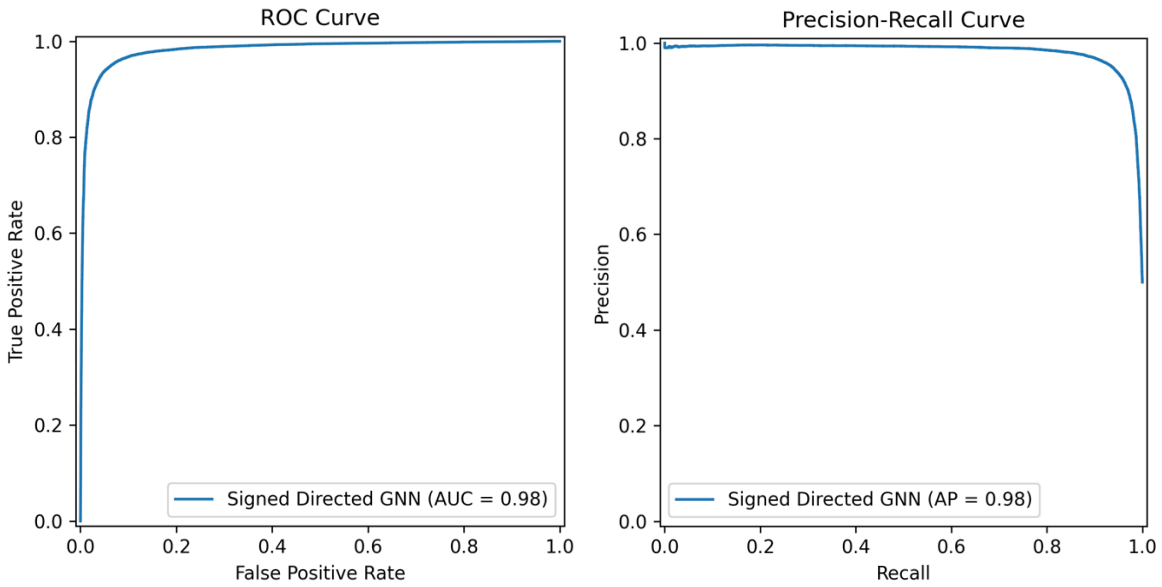

Figure 3. ROC and Precision-recall Curves of the GNN Model

3.3 Top Drug Candidates

We generated a list of top repurposed approved drugs based on the DeepDrug2 framework. Table 1 shows the top 50 drug candidates. Compared to the top 15 drug candidates identified by the original DeepDrug framework [14], DeepDrug2 retains 10 of the original 15 top candidates, indicating significant but not complete overlap. These shared candidates include Tofacitinib, Raloxifene, Empagliflozin, Doxercalciferol, Palbociclib, Febuxostat, Olanzapine, Baricitinib, Miconazole, and Valproic Acid. Notably, Tofacitinib, Baricitinib, Empagliflozin, and Doxercalciferol were also part of the lead five-drug combination identified by the original DeepDrug framework [14].

DeepDrug2 preserves core candidates identified by DeepDrug, showing continuity in prioritizing JAK inhibitors (e.g., Tofacitinib and Baricitinib) and metabolic regulators (e.g., Empagliflozin and Doxercalciferol). However, it substantially broadens the candidate space by introducing drugs from more diverse therapeutic classes (e.g., antidepressants and statins). The absence of several DeepDrug candidates reflects DeepDrug2's updated biomedical graph, which removes somatic mutation data and incorporates germline mutations and pathways, leading to a more comprehensive drug prioritization strategy.

Table 1. Top Repurposed Drug Candidates for AD Identified by DeepDrug2

Ranking	Score	Drug	Approved Treatment	Description
1	0.3386	Tofacitinib	Rheumatic conditions	Janus kinase (JAK) inhibitor
2	0.3252	Raloxifene	Osteoporosis	Estrogen receptor modulator
3	0.2983	Triethylenetetramine	Wilson's disease	Copper chelating agent
4	0.2977	Empagliflozin	Diabetes	SGLT2 inhibitor
5	0.2876	Doxercalciferol	Secondary hyperparathyroidism	Synthetic vitamin D2
6	0.2860	Dihydrotachysterol	Rickets, osteomalacia, hypocalcemia	Synthetic vitamin D
7	0.2856	Dexlansoprazole	Gastroesophageal reflux disease	Proton pump inhibitor (PPI)
8	0.2854	Olopatadine	Allergic conjunctivitis and rhinitis	Histamine H1 antagonist
9	0.2814	Amlodipine	Hypertension	Calcium channel blocker
10	0.2807	Olanzapine	Schizophrenia Bipolar I disorder	Thienobenzodiazepine
11	0.2772	Samidorphan	In combination with antipsychotics	Opioid-system modulator
12	0.2757	Cabotegravir	HIV-1 infection	HIV-1 integrase inhibitor
13	0.2739	Sodium bicarbonate	Heartburn, acid indigestion, upset	pH buffering agent

			stomach, metabolic acidosis	
14	0.2737	Mianserin	Depression and anxiety	Tetracyclic antidepressant
15	0.2717	Febuxostat	Chronic hyperuricemia	Xanthine oxidase inhibitor
16	0.2713	Palbociclib	Breast cancer	Endocrine-based chemotherapeutic agent
17	0.2712	Ziprasidone	Schizophrenia, bipolar mania, and agitation	Atypical antipsychotic
18	0.2711	Levallorphan	Complete or partial reversal of narcotic depression	Opioid antagonist
19	0.2706	Miconazole	Fungal infections	Azole antifungal
20	0.2698	Indapamide	Hypertension	Thiazide diuretic
21	0.2686	Risperidone	Mental health disorders	Antipsychotic medication
22	0.2684	Dapagliflozin	Type 2 diabetes mellitus	Sodium-glucose cotransporter 2 inhibitor
23	0.2681	Cannabidiol	Seizures	Cannabinoid
24	0.2677	Paricalcitol	Hyperparathyroidism	Vitamin D
25	0.2670	Vericiguat	Heart failure-related hospitalization and cardiovascular death	Soluble guanylate cyclase stimulator
26	0.2670	Pantoprazole	Gastroesophageal reflux disease	Proton pump inhibitor

27	0.2666	Baricitinib	Rheumatic conditions	Janus kinase (JAK) inhibitor
28	0.2659	Chlorthalidone	Hypertension or edema	Diuretic
29	0.2658	Amitriptyline	Depression	Tricyclic antidepressant
30	0.2652	Ferric cation	Iron deficiency anemia	Iron supplement
31	0.2649	Methadone	Severe pain	Opioid analgesic
32	0.2647	Saquinavir	HIV-1 with advanced immunodeficiency	HIV protease inhibitor
33	0.2646	Quinidine	Atrial fibrillation and flutter, ventricular arrhythmias	D-isomer of quinine
34	0.2644	Lumateperone	Schizophrenia	Antipsychotic
35	0.2644	Diltiazem	Hypertension	Calcium channel blocker
36	0.2641	Duloxetine	Anxiety disorder, neuropathic pain, osteoarthritis, and stress incontinence	Serotonin norepinephrine reuptake inhibitor
37	0.2641	Danazol	Endometriosis	Synthetic steroid and pituitary gonadotropin inhibitor
38	0.2637	Topiramate	Seizures	Anticonvulsant
39	0.2636	Cyproheptadine	Allergic symptoms	Combined serotonin and histamine antagonist

40	0.2635	Paroxetine	Major depressive disorder	Selective serotonin reuptake inhibitor
41	0.2635	Digoxin	Mild to moderate heart failure	Cardiac glycoside
42	0.2634	Bupropion	Major depressive disorder	Norepinephrine and dopamine reuptake inhibitor
43	0.2633	Cefoperazone	Bacterial infections	Cephalosporin antibiotic
44	0.2631	Clemastine	Allergic rhinitis	Antihistamine
45	0.2630	Pravastatin	Cardiovascular events (myocardial infarction and stroke)	HMG-CoA reductase inhibitor
46	0.2629	Valproic acid	Seizures	Anticonvulsant
47	0.2629	Fluoxetine	Major depressive disorder	Selective serotonin reuptake inhibitor
48	0.2629	Imipramine	Depression	Tricyclic antidepressant
49	0.2627	Atorvastatin	Lower lipid levels	HMG-CoA reductase inhibitor
50	0.2627	Celecoxib	Osteoarthritis, rheumatoid arthritis, acute pain, menstrual symptoms	Selective cyclooxygenase-2 (COX-2) inhibitor

3.4 Drug Validation Results

Among the top 50 drug candidates identified by DeepDrug2 (Table 1), 15 had sufficient medication records in the UK Biobank to enable statistical validation using EHR data. We assessed the association between exposure to each drug and the risk of AD using logistic regression, adjusting for age, gender, ethnicity, and key comorbidities. The results are summarized in Table 2.

Table 2. Top Repurposed Drug Candidates Validated by EHR Data

Ranking[1]	**Drug**[2]	**Odds Ratio (OR)**	***p*-value**[3]
2	Raloxifene	0.71	0.4962
9	Amlodipine	0.87	0.0182 *
10	Olanzapine	1.62	0.2491
20	Indapamide	0.41	0.0006 ***
21	Risperidone	2.16	0.1328
26	Pantoprazole	1.49	0.0787
35	Diltiazem	1.05	0.8351
36	Duloxetine	0.69	0.5163
38	Topiramate	1.13	0.8640
40	Paroxetine	2.12	0.0000 ***
45	Pravastatin	0.91	0.5258
47	Fluoxetine	1.63	0.0001 ***
48	Imipramine	1.05	0.9583
49	Atorvastatin	0.81	0.0007 ***
50	Celecoxib	0.77	0.5604

Notes:

1. Ranking among the top 50 drug candidates identified by DeepDrug2.
2. Only 15 drugs with sufficient medication records in the UK Biobank were selected.
3. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$

Several candidate drugs demonstrated statistically significant associations with AD risk. Notably, Indapamide (OR = 0.41, $p = 0.0006$) and Atorvastatin (OR = 0.81, $p = 0.0007$) were significantly associated with a reduced risk of AD, supporting their potential as AD drugs. Amlodipine, a calcium channel blocker, also showed a modest but statistically significant protective effect (OR = 0.87, $p = 0.0182$). These findings align with previous studies suggesting the neuroprotective effects of certain antihypertensive and lipid-lowering agents in reducing AD risk [30-32].

Conversely, several drugs were associated with increased AD risk. Most notably, Paroxetine (OR = 2.12, $p < 0.0001$) and Fluoxetine (OR = 1.63, $p = 0.0001$) exhibited strong associations with higher AD risk. Both are selective serotonin reuptake inhibitors (SSRIs). While SSRIs have been explored for their potential neurocognitive effects [33], these results suggest the need for further investigation into long-term SSRI exposure and its impact on cognitive decline.

Other drugs, including Olanzapine, Risperidone, and Pantoprazole, showed elevated odds ratios but did not reach statistical significance, potentially due to limited sample size or confounding by indication. It is also worth noting that some drugs, such as Raloxifene, Duloxetine, and Celecoxib, trended toward protective associations but lacked statistical power to confirm significance.

In addition to the 15 drugs analyzed, many of the top-ranking candidates identified by DeepDrug2 could not be evaluated using the UK Biobank data due to insufficient exposure records. This limitation reflects the relatively low prescription rates or recent approval status of some drugs in the general population. As a result, the current validation may underrepresent the full therapeutic potential of the DeepDrug2 candidates. Future validation efforts leveraging larger or more diverse real-world datasets will be critical to systematically evaluate these additional candidates.

4. Discussion

The DeepDrug2 framework demonstrates a robust and innovative approach to drug repurposing for AD, leveraging a signed directed GNN to model complex molecular interactions within a biomedical knowledge graph. The model achieved strong classification performance (AUC = 0.98, AP = 0.98), underscoring its ability to capture nuanced relational patterns between drug and gene nodes in the AD molecular network. These results affirm the model's capacity to represent both the directionality and polarity of interactions, providing a more biologically grounded and informative platform for prioritizing drug candidates.

Compared to the original DeepDrug framework, DeepDrug2 preserves much of the foundational insight while improving and expanding the candidate space. Ten of the top 15 candidates from DeepDrug were retained in DeepDrug2's top 50 list, demonstrating continuity in highlighting JAK inhibitors (e.g., Tofacitinib and Baricitinib) and metabolic regulators (e.g., Empagliflozin and Doxercalciferol). This overlap validates the consistency of the core methodology while also highlighting how updates to the underlying graph, such as the removal of

somatic mutations and the incorporation of germline genetics, can meaningfully shift the prioritization landscape. Importantly, DeepDrug2 introduces greater diversity in therapeutic classes, surfacing antidepressants, antiepileptics, statins, and calcium channel blockers, thereby broadening the horizon of potential AD therapeutics.

The EHR-based drug validation using UK Biobank data provides further support for the clinical relevance of DeepDrug2 predictions. Of the 50 top-ranked candidates, 15 had sufficient medication records to permit statistical evaluation. Among these, three drugs, Indapamide, Atorvastatin, and Amlodipine, were significantly associated with reduced AD risk after adjusting for confounding variables. These results echo prior epidemiological studies suggesting protective effects of antihypertensive and lipid-lowering therapies on cognitive health, reinforcing the potential repurposing value of these widely used agents. Conversely, some candidate drugs were associated with an increased risk of AD. Paroxetine and Fluoxetine, both SSRIs, showed significant positive associations with AD diagnosis. While SSRIs have been studied for their neuropsychiatric benefits in aging populations, these findings call attention to potential long-term risks and underscore the need for more nuanced investigations into their role in neurodegeneration.

It is also important to note that many top-ranked DeepDrug2 candidates were excluded from the EHR validation due to limited real-world exposure data. For example, drugs like Tofacitinib and Empagliflozin, despite high prioritization scores, are relatively new or prescribed for narrow indications, resulting in insufficient representation in the UK Biobank cohort. This highlights an inherent limitation of retrospective validation in existing datasets and signals the need for complementary validation strategies using more diverse, contemporary, or longitudinally rich health records.

Overall, DeepDrug2 presents a significant advancement in AI-driven drug repurposing for AD. Its focus on germline mutations and expanded candidate space provide a strong foundation for future translational studies. The findings from clinical validation of DeepDrug2's predictions underscore the importance of integrating real-world data into computational drug repurposing pipelines. Future work will benefit from broader EHR data integration and prospective validation to better understand the causal effects of these candidate drugs on AD progression. Translating these insights into clinical interventions will be a critical step toward realizing the therapeutic potential of DeepDrug2 in AD prevention and treatment.

5. Conclusion

DeepDrug2 advances AD drug repurposing by integrating a signed directed GNN with an updated biomedical graph that shifts the foundational hypothesis: from emphasizing somatic mutations to focusing on germline genetic risks and pathways. This conceptual shift reflects a more current understanding of AD pathophysiology and enables a biologically grounded, generalizable approach to AD drug candidate prioritization. DeepDrug2 demonstrates strong predictive performance and uncovers a broader, more diverse array of therapeutic candidates. Real-world validation supports the clinical relevance of several top drug candidates, reinforcing their translational potential. Future work will involve experimental studies in mouse and zebrafish models of AD to provide critical preclinical evidence for advancing these candidates toward therapeutic development.

Acknowledgments

ChatGPT-4o was used to improve the language of the manuscript and to assist in the writing of the manuscript (only covering literature review and discussion of results). The authors reviewed the content generated and took full responsibility for the content of the manuscript.

This research was supported in part by the US National Academy of Medicine (NAM) Healthy Longevity Catalyst Award 2021 and "An Artificial Intelligence (AI)-driven Causal Approach for Early Diagnosis and Treatment of Late Onset Alzheimer's Disease" under Seed Fund for Collaborative Research, The University of Hong Kong, June 2023.

Data Availability

The drug and drug-drug interaction data were obtained from DrugBank (https://go.drugbank.com). The drug target data were obtained from DrugBank (https://go.drugbank.com), DrugCentral (https://drugcentral.org), ChEMBL (https://www.ebi.ac.uk/chembl/), and BindingDB (https://www.bindingdb.org/). The protein-protein interaction and pathway data were obtained from STRING (https://string-db.org). The preprocessed biomedical network datasets, including genes and their weights (Table S1), drugs and their weights (Table S2), proteins/targets and their weights (Table S3), and edges and their weights (Tables S4-7), are available in Supplementary Data. The UK Biobank data used in this study can be accessed by qualified researchers through

formal application at https://www.ukbiobank.ac.uk, subject to approval and compliance with ethical guidelines.

References

[1] J. D. Ravula, R. Nirogi, and M. D. Janodia, "Review on 505 (b)(2) drug products approved by USFDA from 2010 to 2020 emphasizing intellectual property and regulatory considerations for reformulations and new combinations," *Journal of Pharmaceutical Sciences,* vol. 112, no. 8, pp. 2146-2175, 2023.

[2] A. Perna, K. S. Montine, L. R. White, T. J. Montine, and B. A. Cholerton, "Paradigm shift: multiple potential pathways to neurodegenerative dementia," *Neurotherapeutics,* vol. 20, no. 6, pp. 1641-1652, 2023.

[3] C. Ballard, D. Aarsland, J. Cummings, J. O'Brien, R. Mills, J. L. Molinuevo, T. Fladby, G. Williams, P. Doherty, and A. Corbett, "Drug repositioning and repurposing for Alzheimer disease," *Nature Reviews Neurology,* vol. 16, no. 12, pp. 661-673, 2020.

[4] J. L. Cummings, Y. Zhou, A. Van Stone, D. Cammann, R. Tonegawa-Kuji, J. Fonseca, and F. Cheng, "Drug repurposing for Alzheimer's disease and other neurodegenerative disorders," *Nature Communications,* vol. 16, no. 1, p. 1755, 2025.

[5] R. Shukla, N. D. Henkel, K. Alganem, A.-r. Hamoud, J. Reigle, R. S. Alnafisah, H. M. Eby, A. S. Imami, J. F. Creeden, and S. A. Miruzzi, "Signature-based approaches for informed drug repurposing: targeting CNS disorders," *Neuropsychopharmacology,* vol. 46, no. 1, pp. 116-130, 2021.

[6] Y. Zhou, J. Fang, L. M. Bekris, Y. H. Kim, A. A. Pieper, J. B. Leverenz, J. Cummings, and F. Cheng, "AlzGPS: a genome-wide positioning systems platform to catalyze multi-omics for Alzheimer's drug discovery," *Alzheimer's Research & Therapy,* vol. 13, pp. 1-13, 2021.

[7] X. Pan, X. Lin, D. Cao, X. Zeng, P. S. Yu, L. He, R. Nussinov, and F. Cheng, "Deep learning for drug repurposing: Methods, databases, and applications," *Wiley Interdisciplinary Reviews: Computational Molecular Science,* vol. 12, no. 4, p. e1597, 2022.

[8] J. Xu, C. Mao, Y. Hou, Y. Luo, J. L. Binder, Y. Zhou, L. M. Bekris, J. Shin, M. Hu, and F. Wang, "Interpretable deep learning translation of GWAS and multi-omics findings to identify pathobiology and drug repurposing in Alzheimer's disease," *Cell Reports,* vol. 41, no. 9, 2022.

[9] K.-L. Hsieh, G. Plascencia-Villa, K.-H. Lin, G. Perry, X. Jiang, and Y. Kim, "Synthesize heterogeneous biological knowledge via representation learning for Alzheimer's disease drug repurposing," *iScience,* vol. 26, no. 1, 2023.

[10] J. Downey, J. C. K. Lam, V. O. K. Li, and I. Gozes, "Somatic mutations and Alzheimer's disease," *Journal of Alzheimer's Disease,* vol. 90, no. 2, pp. 475-493, 2022.

[11] S. Sragovich, M. Gershovits, J. C. K. Lam, V. O. K. Li, and I. Gozes, "Putative blood somatic mutations in post-traumatic stress disorder-symptomatic soldiers: High impact of cytoskeletal and inflammatory proteins," *Journal of Alzheimer's Disease,* vol. 79, no. 4, pp. 1723-1734, 2021.

[12] V. O. K. Li, J. C. K. Lam, Y. Han, L. Y. L. Cheung, J. Downey, T. Kaistha, and I. Gozes, "Designing a protocol adopting an artificial intelligence (AI)–driven approach for early diagnosis of late-onset Alzheimer's disease," *Journal of Molecular Neuroscience,* vol. 71, no. 7, pp. 1329-1337, 2021.

[13] V. O. K. Li, Y. Han, T. Kaistha, Q. Zhang, J. Downey, I. Gozes, and J. C. K. Lam, "DeepDrug: An Expert-led Domain-specific AI-Driven Drug-Repurposing Mechanism for Selecting the Lead Combination of Drugs for Alzheimer's Disease," *medRxiv,* 2024.

[14] V. O. K. Li, Y. Han, T. Kaistha, Q. Zhang, J. Downey, I. Gozes, and J. C. K. Lam, "DeepDrug as an expert guided and AI driven drug repurposing methodology for selecting the lead combination of drugs for Alzheimer's disease," *Scientific Reports,* vol. 15, 2025, Art no. 2093.

[15] G. Nicolas, R. Acuña-Hidalgo, M. J. Keogh, O. Quenez, M. Steehouwer, S. Lelieveld, S. Rousseau, A. C. Richard, M. S. Oud, and F. Marguet, "Somatic variants in autosomal dominant genes are a rare cause of sporadic Alzheimer's disease," *Alzheimer's & Dementia,* vol. 14, no. 12, pp. 1632-1639, 2018.

[16] Z. Li, S. Min, N. Alliey-Rodriguez, G. Giase, L. Cheng, D. W. Craig, G. J. Faulkner, H. Asif, C. Liu, and E. S. Gershon, "Single-neuron whole genome sequencing identifies increased somatic mutation burden in Alzheimer's disease related genes," *Neurobiology of Aging,* vol. 123, pp. 222-232, 2023.

[17] V. O. K. Li, J. C. K. Lam, and Y. Han, "LMP-TX: An AI-driven Integrated Longitudinal Multi-modal Platform for Early Prognosis of Late Onset Alzheimer's Disease," *medRxiv,* 2024.

[18] V. O. K. Li, Y. Han, and J. C. K. Lam, "Unravelling Causal Genetic Biomarkers of Alzheimer's Disease via Neuron to Gene-token Backtracking in Neural Architecture: A Groundbreaking Reverse-Gene-Finder Approach," *arXiv:2502.03938,* 2025.

[19] I. E. Jansen, J. E. Savage, K. Watanabe, J. Bryois, D. M. Williams, S. Steinberg, J. Sealock, I. K. Karlsson, S. Hägg, and L. Athanasiu, "Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk," *Nature Genetics,* vol. 51, no. 3, pp. 404-413, 2019.

[20] C. Bellenguez, F. Küçükali, I. E. Jansen, L. Kleineidam, S. Moreno-Grau, N. Amin, A. C. Naj, R. Campos-Martin, B. Grenier-Boley, and V. Andrade, "New insights into the genetic etiology of Alzheimer's disease and related dementias," *Nature Genetics,* vol. 54, no. 4, pp. 412-436, 2022.

[21] D. P. Wightman, I. E. Jansen, J. E. Savage, A. A. Shadrin, S. Bahrami, D. Holland, A. Rongve, S. Børte, B. S. Winsvold, and O. K. Drange, "A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease," *Nature Genetics,* vol. 53, no. 9, pp. 1276-1282, 2021.

[22] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, and Z. Sayeeda, "DrugBank 5.0: a major update to the DrugBank database for 2018," *Nucleic Acids Research,* vol. 46, no. D1, pp. D1074-D1082, 2018.

[23] S. Avram, C. G. Bologa, J. Holmes, G. Bocci, T. B. Wilson, D.-T. Nguyen, R. Curpan, L. Halip, A. Bora, and J. J. Yang, "DrugCentral 2021 supports drug discovery and repositioning," *Nucleic Acids Research,* vol. 49, no. D1, pp. D1160-D1169, 2021.

[24] A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, and E. Cibrián-Uhalte, "The ChEMBL database in 2017," *Nucleic Acids Research,* vol. 45, no. D1, pp. D945-D954, 2017.

[25] M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, and J. Chong, "BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology," *Nucleic Acids Research,* vol. 44, no. D1, pp. D1045-D1053, 2016.

[26] D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay, T. Fang, and P. Bork, "The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets," *Nucleic Acids Research,* vol. 49, no. D1, pp. D605-D612, 2021.

[27] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, and J. O'Connell, "The UK Biobank resource with deep phenotyping and genomic data," *Nature,* vol. 562, no. 7726, pp. 203-209, 2018.

[28] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "KEGG: new perspectives on genomes, pathways, diseases and drugs," *Nucleic Acids Research,* vol. 45, no. D1, pp. D353-D361, 2017.

[29] M. Zhang and Y. Chen, "Link prediction based on graph neural networks," *Advances in Neural Information Processing Systems,* vol. 31, 2018.

[30] D. Pinakhina, A. Loboda, A. Sergushichev, and M. Artomov, "Gene, cell type, and drug prioritization analysis suggest genetic basis for the utility of diuretics in treating Alzheimer disease," *Human Genetics and Genomics Advances,* vol. 4, no. 3, 2023.

[31] E. Olmastroni, G. Molari, N. De Beni, O. Colpani, F. Galimberti, M. Gazzotti, A. Zambon, A. L. Catapano, and M. Casula, "Statin use and risk of dementia or Alzheimer's disease: a systematic review and meta-analysis of observational studies," *European Journal of Preventive Cardiology,* vol. 29, no. 5, pp. 804-814, 2022.

[32] K. Sandoval, "Amlodipine use in Mild Cognitive Impairment reduces the risk of incident dementia in National Alzheimer's Coordinating Center participants aged≥ 55 years from 2014-2019," *The Journal of Pharmacology and Experimental Therapeutics,* vol. 385, p. 428, 2023.

[33] H. Wang, S. Li, J. Zhang, W. Peng, T. Li, and J. Zhang, "Efficacy of selective serotonin reuptake inhibitors-related antidepressants in Alzheimer's disease: a meta-analysis," *European Journal of Medical Research,* vol. 29, no. 1, p. 438, 2024.