A random-walk-based learning framework to uncover novel gene candidates for Alzheimer's disease therapy

Alena Orlenko*[1], Binglan Li[1], Neda Khanjani[1], Mythreye Venkatesan[1], Li Shen[2,3], Marylyn D. Ritchie[2,3,4], Zhiping Paul Wang[1], Tayo Obafemi-Ajayi[5], Jason H. Moore[1]

[1]*Department of Computational Biomedicine*
Cedars-Sinai Medical Center, Los Angeles, CA, USA

[2]*Division of Informatics, Department of Biostatistics, Epidemiology, and Informatics*
[3]*Institute for Biomedical Informatics*
[4]*Department of Genetics*
Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[5]*Engineering Program*
Missouri State University, Springfield, Missouri, USA

Identifying repurposable therapeutic targets for Alzheimer's disease (AD) remains challenging due to various clinical and biological factors. This study aimed to identify candidate genes for AD therapy. We hypothesize that gene and disease-specific network properties—learnable from these large-scale biomedical knowledge graphs—can inform implicit gene-AD connections and prioritize repurposable AD drug targets. To evaluate the hypothesis, we focused on druggable genes curated from Drug-Gene Interaction Database and Alzheimer's Knowledge Base (AlzKB). We applied scalable random walk methods to Hetionet to learn unbiased gene and disease embeddings, representative of their topological and semantic network properties. The embeddings were then used to compute gene-AD similarity and derive network-based scores for each gene. To validate the scores, using Alzheimer's Disease Sequencing Project (ADSP) data, we constructed AD classifier models with Tree-based pipeline optimizer 2 (TPOT2), an automated machine learning framework. Models were optimized for performance, model complexity, and high aggregate network-based scores. Network-based scores successfully prioritized diverse feature sets—many not previously associated with AD—that are enriched in biologically meaningful body parts such as brain, and pathways including neuronal signaling, potassium channels, and creatine metabolism. The results suggested that knowledge graphs and network-informed embeddings can capture both known and novel insights into AD mechanisms. Additionally, integrating network-based scores with feature-set-guided TPOT2 offers a scalable and biologically interpretable framework for AD drug repurposing and discovery.

Keywords: genomics; Alzheimer's disease; random walks; knowledge graph; network analysis

*This work was supported by National Institutes of Health (USA) grants U01 AG066833, P30 AG094848 and R01 LM010098.

1. Introduction

Therapeutic intervention for Alzheimer's disease (AD) remains a significant challenge, largely due to pronounced clinical and etiological heterogeneity as well as a limited understanding of the genetic mechanisms underlying its pathogenesis.[1–3] Over 80 genetic loci have been associated with AD, including well-known risk factors such as common gene polymorphisms in *APOE* as well as rare mutations in *APP*, *PSEN1*, and *PSEN2*.[4] Despite advances in AD genotypic characterization, these variations explain only a limited fraction of the overall disease risk and offer few actionable insights for therapeutic development.

To address this gap, computational drug discovery methods provide a cost-efficient, systematic approach to identifying novel therapeutic targets.[5,6] These include methods such as computational molecular docking, pathway or network mapping, retrospective clinical analysis of electronic health records, and genetic associations. In addition, the utility of knowledge graphs (e.g., Hetionet,[7] PrimeKG,[8] and AlzKB[9]) can facilitate the drug discovery process by providing a framework to systematically integrate structured, relational information (e.g., gene–drug interactions) across diverse biomedical entities into disease-specific network contexts. For example, prior work identified AD therapy candidates using the AlzKB network mining of AD genes and their first neighbors, reflecting nonspecific gene-gene interactions.[10] Knowledge graphs also enable the inference of indirect relationships among domain-specific entities, such as predicting latent associations between drugs and diseases. Drug repurposing is sometimes modeled as a link prediction task in a knowledge graph to infer novel associations between drugs and diseases.[11]

The overall aim of this study is to identify novel, biologically relevant gene candidates for AD therapy by systematically leveraging insights derived from the knowledge graphs by using network learning models. We hypothesize that network properties learnable from these large-scale biomedical knowledge graphs can inform implicit gene-AD connections to prioritize repurposable, druggable gene targets. Note that druggability is based on an evidence-based interaction score derived from the Drug–Gene Interaction Database (DGIdb).[12] DGIdb is a knowledge repository built from expert curation and text mining of over 40 curated biological resources, with a specific focus on confirmed and potential drug–gene interactions.

A variety of computational approaches have been developed to exploit the topological and semantic structures of knowledge graphs for the link prediction task, including graph neural networks (GNNs),[13] random walk-based techniques (e.g., node2vec,[14] edge2vec[15]), and rotation-based embedding models (e.g., RotatE[16]). Collectively, these methods are capable of learning continuous vector representations that capture the complex relational patterns embedded within the knowledge graph. In this study, we applied random-walk-based approaches on Hetionet to learn gene and disease embeddings that could potentially encode useful topological and semantic properties. The walk-based approaches are scalable and can preserve both local connectivity and global network structural features. Given that Hetionet comprises relational information among heterogeneous biomedical entities beyond AD, extracting gene and disease embeddings from the network should mitigate biases towards AD and overfitting in downstream analysis. The learned embeddings are subsequently used to compute the similarity between a druggable gene node and the disease (AD) node. These similarities are applied to derive network-based scores for each gene, under the hypothesis that such scores approximate the likelihood of a gene as an AD drug target by integrating network topology and biological relevance.

To evaluate the efficacy of network-based learning approaches in uncovering novel drug targets, we constructed AD predictive models informed by network-derived scores using an automated machine learning framework, Tree-based Pipeline Optimizer 2 (TPOT2). TPOT2 incorporates both multi-objective optimization and a feature set selector, enabling the model selection process to be guided by both predictive performance and the relevance of predefined feature groups. By exploring two feature sets derived from pathways and drug–gene interactions, we identified a pareto front of optimal machine learning pipelines that produced AD-relevant solutions that were high-performing and minimally complex. Enrichment analysis of the top-performing models revealed functionally diverse gene sets, highlighting promising candidates for future investigation.

2. Methods

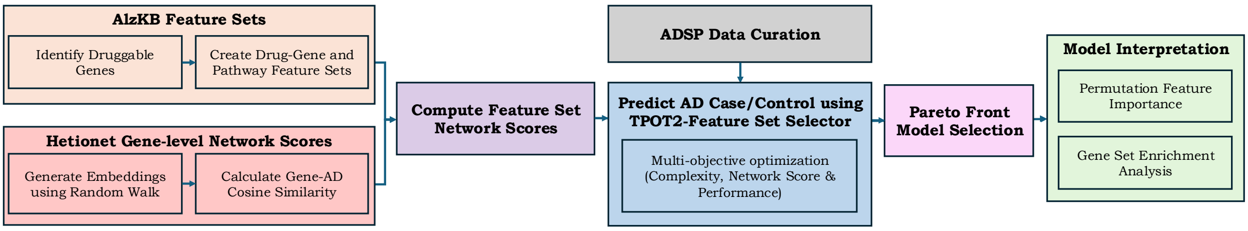

AlzKB Feature Sets
Identify Druggable Genes
Create Drug-Gene and Pathway Feature Sets
Hetionet Gene-level Network Scores
Generate Embeddings using Random Walk
Calculate Gene-AD Cosine Similarity
Compute Feature Set Network Scores
ADSP Data Curation
Predict AD Case/Control using TPOT2-Feature Set Selector
Multi-objective optimization (Complexity, Network Score & Performance)
Pareto Front Model Selection
Model Interpretation
Permutation Feature Importance
Gene Set Enrichment Analysis

Fig. 1. Flowchart of overall study design

2.1. *Selection of druggable genes and construction of feature sets*

To identify actionable AD therapeutic targets, we restricted our analysis to druggable protein-coding genes found in AlzKB,[9] as detailed in prior work.[10] This yielded 3,612 druggable genes for downstream analysis. We constructed two biologically informed feature sets (drug–gene interaction and pathway-based) from these genes based on specific edge relationships in AlzKB.

Drug-gene sets, defined as groups of genes associated with a given drug, were obtained by extracting drug nodes from the AlzKB knowledge graph that are directly connected to druggable genes, denoting interactions. Each set consists of gene nodes that are linked to a specific drug node via defined relationships including general interactions (*chemicalbindsgene*) and specific mechanisms of drug action (*chemicaldecreasesexpression* and *chemicalincreasesepression*). Likewise, pathway-based feature sets, defined as groups of genes associated with a given pathway, are generated by mapping druggable genes to biological pathways from AlzKB (*geneinpathway*). These pathway capture functional relationships relevant to clinical applications, particularly for AD pathology.

2.2. *Learning embeddings via random walk on Hetionet*

To learn the gene and disease embeddings from Hetionet, we applied two types of random-walk-based methods: node2vec[14] and edge2vec.[15] (Note that other types of random walk methods, such as DREAMwalk,[17] are not explored in this work as they require a diverse set of homogeneous network datasets, rather than a single heterogeneous knowledge graph.) Node2vec captures the sequence of nodes that a random walk traverses in the graph, while edge2vec includes the edge type information as well. The Hetionet network was downloaded as TSV files from the repository (https://github.com/hetio/hetionet), read into Python as a NetworkX object, and pruned to include only connected nodes. Node2vec was performed using using the $Node2Vec$ function from PyTorch Geometric while edge2vec was done based on scripts obtained from the edge2vec GitHub repository (https://github.com/RoyZhengGao/edge2vec). The extracted walk sequences were input into Word2vec to learn the continuous feature representations (embeddings). The hyperparameters used to learn the embeddings were optimized through grid search, guided by the prediction (XGBoost) performance on established gene–AD associations. The details of the optimized parameter sets are listed in the supplementary file [a] Finally, the similarity between each gene node and the derived AD embeddings was computed using the cosine similarity metric. These gene-level similarity scores ($S(G)$) were subsequently used to derive the feature set network scores.

2.3. *Computation of feature set network score*

The network scores for each feature set (per drug for the drug-gene sets while per pathway for the pathway sets) were computed as an aggregate of similarity-based scores for all genes within the set. There are four variations of network scores based on multiple levels of constraints on the raw similarity scores derived from the random walk step (see Section 2.2) from both individual gene and gene set levels. At the individual gene level, we imposed a constraint to penalize genes for proximity to known AD genes by discretizing $S(G)$ ($S_d(G)$) into five ordinal categories using quantile-defined thresholds (<50%, 50–75%, 75–85%, 85–95%, and >95%), corresponding to the scores −2, −1, 0, +1, and +2, respectively. An AD penalty score, $P(G)$, is mapped to each gene. $P(G)$ denotes the penalty for a gene (G) being a known AD gene (−2), a first-degree neighbor of AD gene (−1), or none of the above (+1). Thus $S*(G) = S_d(G) + P(G)$. At the feature set level, we imposed a constraint to exclude AD genes, i.e., all AD genes present in the feature set are excluded prior to computing the network score as an aggregate of the remaining genes. This yielded the following four setups for network scores. Let NS_i denote the network score for a feature gene set i (FS_i).

- Setup 1 (S1): NS_i is simply the average of all the raw similarity scores ($S(G)$), obtained from the random walk step, of the genes in set i. $NS_i^1 = avg(S(G))\ \forall G \in FS_i$. This served as the baseline experiment, as no constraints were imposed.
- Setup 2 (S2) adds a biological constraint by introducing penalties for proximity to known AD genes. $NS_i^2 = \sum_G S_d(G) + P(G)\ \forall G \in FS_i$.

[a] Supplementary information: `https://github.com/EpistasisLab/ADSP_Network_Score/`

- Setup 3 (S3) excludes AD genes from network score computation. $NS_i^3 = avg(S(G)\ \forall G \in FS_{i-\text{AD genes}}$, where $FS_{i-\text{AD genes}}$ denotes feature set i without the AD genes.
- Setup 4 (S4) assigns the strictest constraint of both AD penalty and excluding AD genes. $NS_i^4 = \sum_G S_d(G) + P(G)\ \forall G \in FS_{i-\text{AD genes}}$.

Note that the four types of aggregated network score are computed per feature set (for all drug-gene and pathway sets) per random walk type (node2vec vs. edge2vec).

2.4. *Multi-objective optimization with feature sets selector*

To assess the efficacy of using network scores with the gene feature sets, we construct AD case/control prediction models using TPOT2 with the feature selector set (FSS) module.[18,19] We configured the FSS using our predefined gene (drug-gene vs. pathway) feature sets (see Section 2.1). TPOT2 uses multi-objective optimization via the NSGA-II evolutionary algorithm[20] to balance performance (AUC) and model complexity (i.e., number of learned parameters). This generates a Pareto front, a set of solutions that are non-dominated by each other. We employed a custom objective function designed to jointly optimize three criteria: maximizing predictive performance, maximizing the feature set network score, and minimizing the complexity score [b]. This approach ensures that the selected solutions are simple, biologically relevant, and exhibit optimal model performance.

Optimal solutions were selected from the Pareto front set of solutions by applying median thresholds to each objective function component. Specifically, a solution was required to exceed the median AUC test set and median network score while remaining below the median model complexity. If no solution satisfied all three criteria, the first two thresholds (AUC and network score) were applied, and the solution with the lowest complexity among the remaining candidates was selected. Each selected solution was subsequently analyzed using permutation feature importance, knowledge graph-based network analysis, and functional gene set enrichment analysis.

2.5. *Model interpretation*

We examined the biological entities in the AlzKB that are associated with genes within a specific drug-gene or pathway set. For each gene in a given feature set, we retrieved the connected AlzKB entities, including body parts (anatomical structures) and biological pathways.

Enrichment analyses are also conducted on each drug or pathway set using hypergeometric tests to evaluate the over-representation of particular biological entities and context. For each biological entity (body parts or pathways), p-values from hypergeometric tests were corrected for multiple testing using the Benjamini-Hochberg method.

2.6. *AD data sample and curation*

This study utilized genetic data obtained from the Alzheimer's Disease Sequencing Project (ADSP), R4 v11 2023 release.[21] Data preprocessing and population stratification adjustments

[b] Supplementary information: `https://github.com/EpistasisLab/ADSP_Network_Score/`

followed the methodology in Orlenko et al.[10] Quality control was performed to exclude duplicate samples, singletons, rare variants, low-call-rate variants (missing at >1%), and poorly genotyped samples (missing at >5%). Only common variants (MAF >1%) were selected, resulting in 9,520,653 variants and 34,971 samples.

A novel propensity score matching (PSM) method was used to correct for population stratification in the ADSP dataset.[10] Principal component analysis (PCA) was performed on a subset of independent loci (MAF > 2%, Hardy-Weinberg equilibrium p > 1e-7, linkage disequilibrium $R^2 < 0.1$) to derive eight principal components. To balance Alzheimer's disease cases and controls, PSM was applied using logistic regression and k-nearest neighbor algorithms, resulting in a matched dataset of 22,560 samples. A quarter of the matched dataset was reserved for GWAS to generate gene-level risk scores (GRS). The remainder was evenly divided into training, validation, and test sets (5,640 samples each) using a multi-objective optimization method that preserved both case–control matching and the distribution of 30 significant SNPs identified from the Lancet 2023 review study[4] and additional filtering.

To compute GRS, GWAS was performed on the held-out subset using the SAIGE package,[22] which corrects for relatedness and imbalanced case-control ratios. Post-GWAS filtering involved clumping via PLINK2[23] with settings (`--clump p1`=1e-4, `p2`=1, `r2`=0.1), retaining only SNPs that met the threshold of $p \leq 0.05$. The GWAS variants were annotated using the Variant Effect Predictor),[24] and those located in the gene region or within 500 kilobases upstream were mapped to corresponding genes. Gene-level risk scores were computed using the PLINK2 `--score` function by multiplying individual genotype (0, 1, or 2) with GWAS-derived beta coefficients for each variant. These weighted values were summed per gene and normalized by the number of variants in the gene to account for gene size differences.

3. Results and Analysis

3.1. *Hetionet learned similarity scores*

The distributions of the cosine similarities between the AD disease node and the druggable gene nodes differed substantially between edge2vec and node2vec (Fig. 2). Node2vec produced a unimodal cosine similarity distribution centered around 0.08, with values ranging from –0.24 to 0.36. In contrast, edge2vec yielded a left-shifted distribution with a higher median value of 0.56 and a range from 0.10 to 0.81. These results suggest that the two embedding methods capture distinct aspects of the network structure relative to AD node proximity.

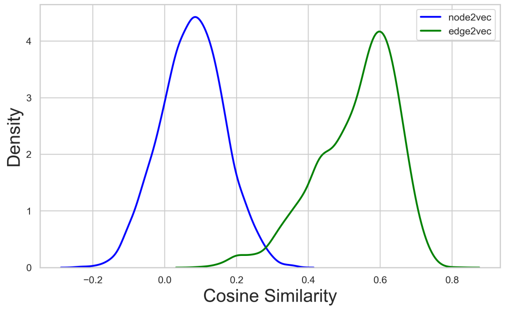
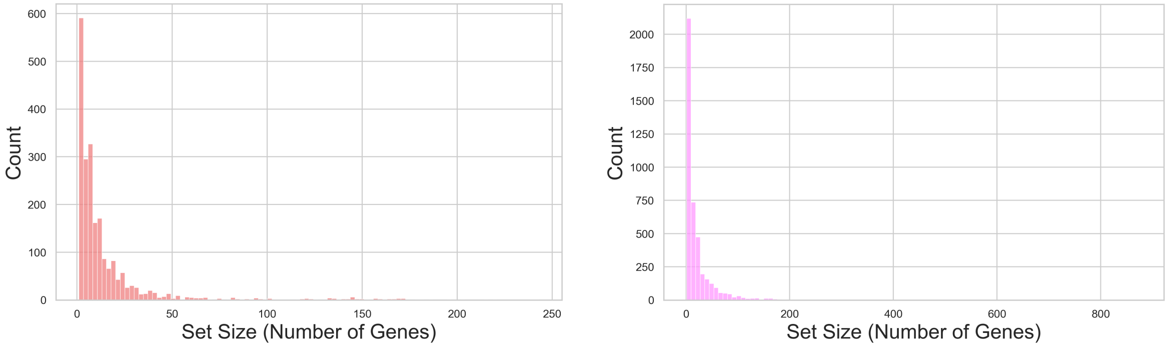


Fig. 2. Distribution of Hetionet gene similarity scores: edge2vec vs. node2vec.

3.2. *Multi-objective optimization with feature set selector identifies biologically informed AD predictive models*

The druggable gene space yielded 2,169 drug–gene feature sets and 4,233 pathway sets. The distribution of drug-gene set sizes was highly skewed, with a small number of drugs associated with large gene sets and most drugs linked to fewer than 10 genes (Fig.3(a)) The largest set, corresponding to drug DB01254 (Dasatinib), comprised 243 genes, while the median set size was 7 genes. Pathway gene sets displayed a broader range with a median size of 9 genes and the largest set, Signal Transduction, containing 878 genes. The distribution was also skewed, indicating a prevalence of smaller pathways (Fig.3(b)).



(a) Drug-gene set

(b) Pathway sets

Fig. 3. Distribution of the gene feature sets' size

The distribution of the Pareto front solutions produced by TPOT2 with FSS varied greatly across all experiments (Fig.4, Table 1). The drug-gene feature set analysis based on node2vec yielded the largest number of Pareto front solutions. Its S2 experimental setup produced 98 models and 46 unique feature sets, while S3 had 97 with 54 distinct feature sets (Table 1). For edge2vec, the drug-gene feature set analysis produced fewer solutions and less unique feature sets: 55/30 (S1), 68/26 (S2), 56/39 (S3), and 43/26 (S4), but still substantially more solutions than pathway feature set analyses based on either walk methods, suggesting that biological constraints shaping drug-gene feature sets provide more flexibility in exploration of the tradeoffs among the three objectives. Drug-gene feature set analyses based on either node2vec or edge2vec prioritized feature sets containing a greater number of genes compared to pathway analyses. Notably, drug-gene feature set based on node2vec led to the best-performing models, specifically S1 (64.2%), S2 (63.9%), S3 (64.9%), and S4 (64.2%).

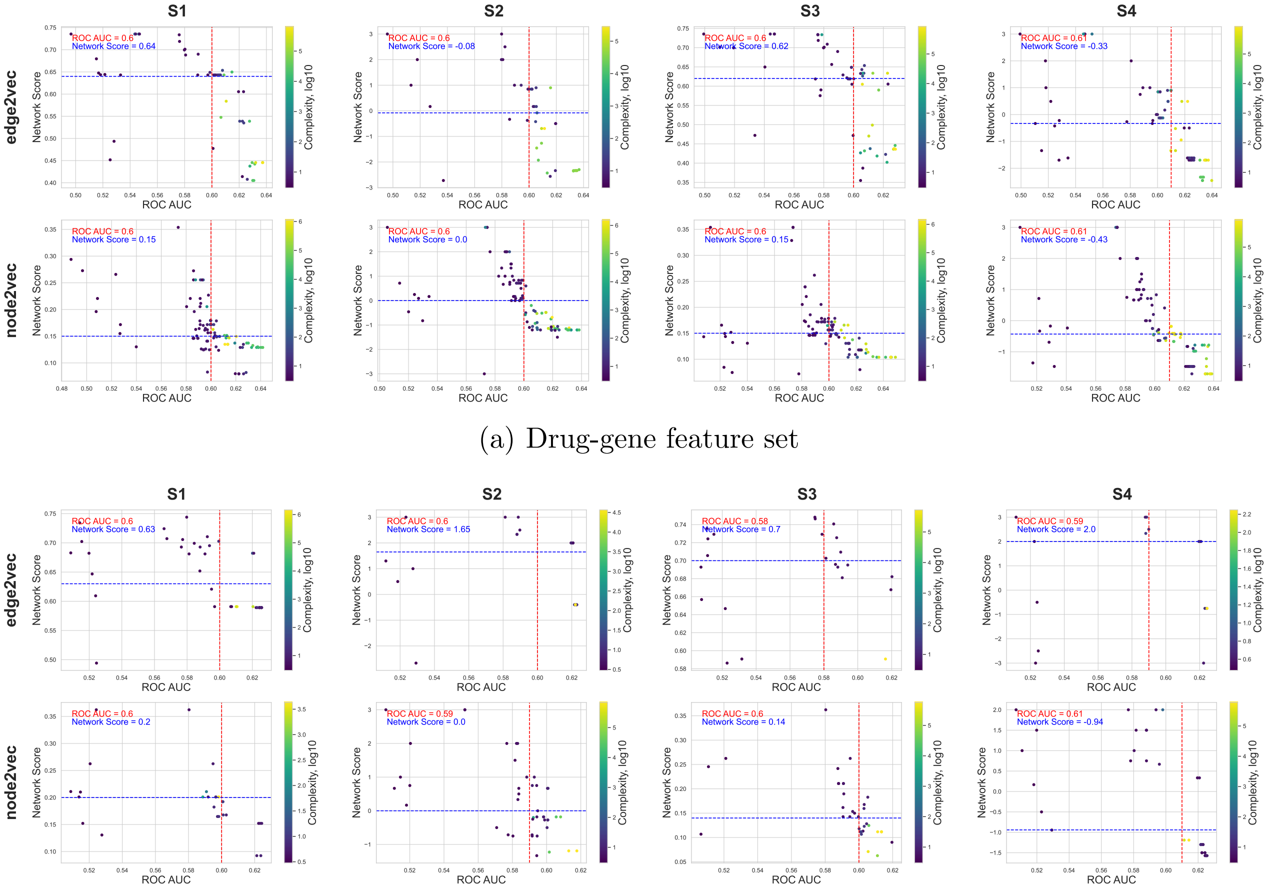
Pathway feature set analyses produced less complex models, likely due to smaller feature set sizes, while maintaining performance comparable to the more complex solutions. Specifically, in edge2vec S2 and node2vec S1 experiments, we observed the least complex median models. Network scores associated with Pareto front solutions from edge2vec-based pathway set analyses were notably higher than those from the corresponding drug-gene feature set setups (S2, S3, and S4); however, this trend was not observed in the node2vec experiments. Note that network scores are only directly comparable between S1 and S3, and between S2 and S4, due to differences in score computation methods across setups (see Section 2.3).

Table 1. Summary of Pareto front models for different walk methods (edge2vec and node2vec) on drug-gene and pathway feature sets across four experimental setups of network scores calculations.

Setup	**# models**	**# unique feature sets [range size]**	**Complexity**	**Network score**	**Test ROC AUC (%)**
Walk method: edge2vec - Drug-gene feature set analysis					
S1	55	30 [1, 221]	2.7 [1.1, 13.4]	0.6 [0.4, 0.7]	60.4 [49.7, 63.7]
S2	68	26 [1, 172]	2.7 [1.1, 13.6]	-0.3 [-2.5, 3.0]	60.8 [49.9, 64.0]
S3	56	39 [1, 198]	2.4 [1.1, 13.6]	0.6 [0.4, 0.7]	59.9 [49.9, 62.8]
S4	43	26 [1, 166]	2.6 [1.1, 12.5]	-0.1 [-2.7, 3.0]	60.5 [49.6, 63.7]
Walk method: node2vec - Drug-gene feature set analysis					
S1	95	43 [1, 180]	2.1 [1.1, 13.9]	0.2 [0.1, 0.4]	59.9 [48.7, 64.2]
S2	98	46 [1,174]	2.4 [1.1,13.6]	-0.4 [-1.7,3.0]	60.5 [50.9,63.9]
S3	97	54 [1, 217]	2.2 [1.1, 14.2]	0.1 [0.1, 0.4]	60.1 [50.8, 64.9]
S4	96	52 [1, 217]	2.2 [1.1, 14.3]	0.0 [-3.0, 3.0]	59.9 [50.5, 63.8]
Walk method: edge2vec - Pathway feature set analysis					
S1	40	22 [1, 21]	1.6 [1.1, 14.2]	0.6 [0.5, 0.7]	59.6 [50.9, 62.6]
S2	16	10 [2, 26]	1.4 [1.1, 5.2]	2.0 [-3.0, 3.0]	58.9 [51.2, 62.4]
S3	23	21 [1, 21]	1.4 [1.1, 13.2]	0.7 [0.6, 0.7]	57.9 [50.7, 62.0]
S4	20	12 [1, 10]	1.7 [1.1, 10.5]	1.6 [-2.7, 3.0]	60.5 [51.2, 62.3]
Walk method: node2vec - Pathway feature set analysis					
S1	28	14 [1, 14]	1.7 [1.1, 8.4]	0.2 [0.1, 0.4]	59.6 [50.9, 62.4]]
S2	35	15 [1, 21]	1.6 [1.1, 13.3]	-0.9 [-1.6, 2.0]	61.4 [50.7, 62.6]
S3	36	26 [1, 31]	1.9 [1.1, 13.2]	0.1 [0.1, 0.4]	60.0 [50.6, 62.0]
S4	43	25 [1, 28]	1.6 [1.1, 13.4]	0.0 [-1.3, 3.0]	59.2 [50.6, 61.8]

3.2.1. *Drug-gene and pathway feature sets optimal solutions*

For each experimental setup, no solutions satisfied all three criteria. Hence, we selected solutions that exceeded the median thresholds for both performance and network scores. The results of permutation feature importance (PFI) analysis on the selected solutions are presented in Fig.5. For the drug-gene feature sets results (Fig.5(a)), specifically for edge2vec, the optimal solution converged on the same drug–gene set for 3 of the 4 experimental setups: DB06637 (Dalfampridine). Dalfampridine is a potassium channel blocker consisting of 40 genes predominantly voltage-gated potassium channel genes (*KCN*) and one solute carrier transporter (*SLC22A2*) (Fig.5(a)). All three models reported an AUC of 61% but displayed variability in the PFI rankings as well as a small mean decrease in accuracy. This implies that, while no gene had a strong main effect, there could exist some interactions among the genes. For the S3 experiment, the optimal solution was the DB00128 (Aspartic acid) feature set. This drug is a non-essential amino acid commonly used in amino acid supplementation therapies. The feature set included 8 genes with diverse functional affiliations. The DB00128 model had an AUC of 61%, but its PFI analysis did not reveal any strong contributions from individual genes.



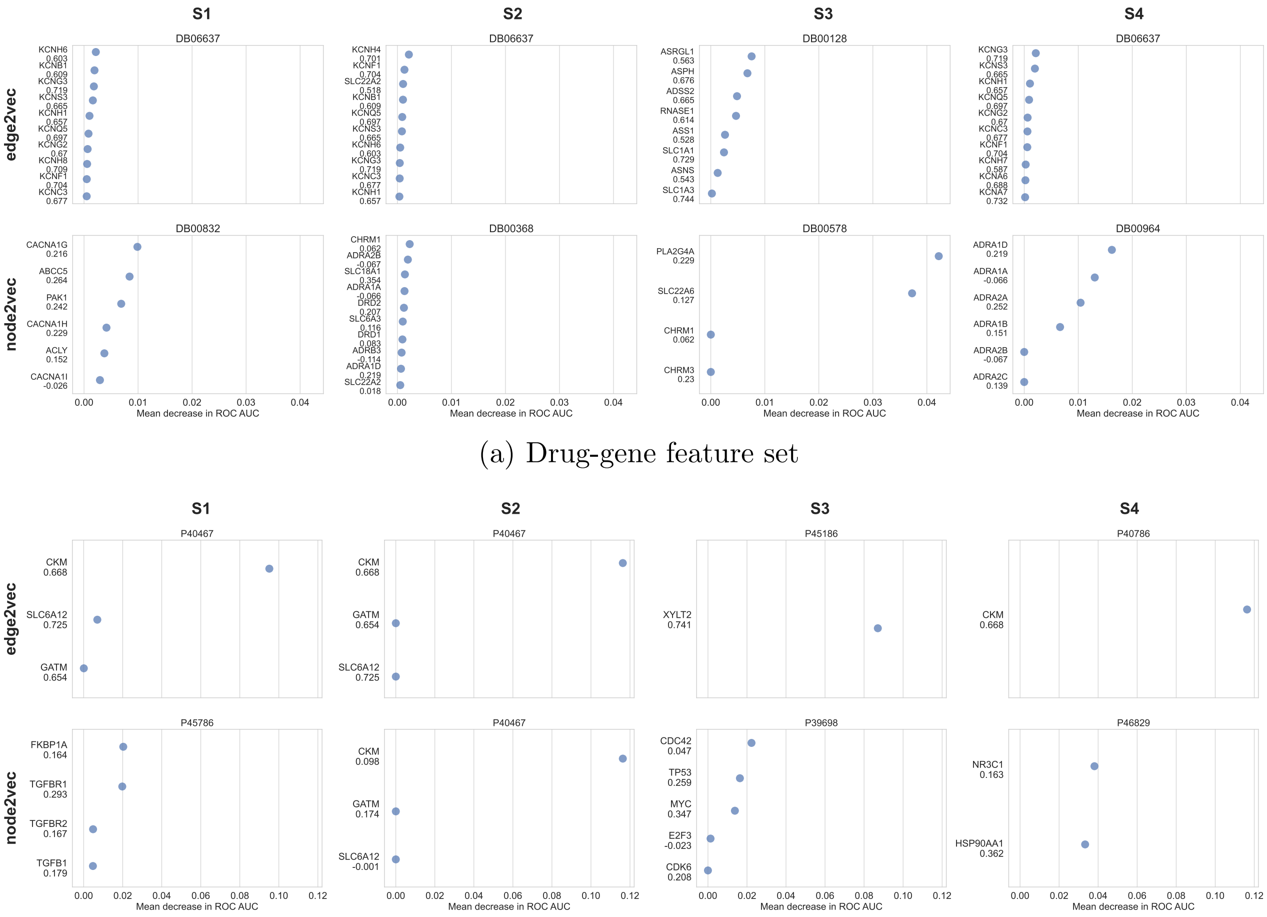
(a) Drug-gene feature set

(b) Pathway feature set

Fig. 4. TPOT2 Pareto front solutions for drug-gene and pathway feature sets experiments

For node2vec, all four experimental setups identified an optimal solution. The best models were DB00832 (Phensuximide), DB00368 (Norepinephrine), DB00578 (Carbenicillin), and DB00964 (Apraclonidine) for S1, S2, S3, and S4, respectively. Their predictive performances were comparable. The drug-gene feature set sizes ranged from 6 to 22, with diverse functionality. The drugs also varied in their known functions. Phensuximide is an anticonvulsant medication used to treat absence seizures. Norepinephrine is a blood pressure medication, while Carbenicillin is an antibiotic drug. Apraclonidine is an adrenergic agonist used to treat raised intraocular pressure.

For edge2vec-based pathway feature sets (Fig. 5(b)), both S1 and S2 (top row) identified models with the Creatine Metabolism pathway as the optimal solution (AUC of 62%). This pathway consisted of three druggable genes: *CKM* (Creatine Kinase, Muscle), *GATM* (Glycine Amidinotransferase), and *SLC6A12* (Solute Carrier Family 6 Member). Notably, *CKM* exhibited a strong main effect in both models, with PFI coefficients of 0.09 and 0.12 mean decrease in AUC, respectively. In experiment S3, the Glycosaminoglycan Biosynthesis pathway was identified as the optimal solution (AUC=59%). This pathway has only one druggable gene, *XYLT2*. Best solution for S4 was Creatine-Phosphate Biosynthesis pathway (AUC=62%) and only included *CKM* gene (which was also key for S1 and S2 models).



(a) Drug-gene feature set

(b) Pathway feature set

Fig. 5. PFI for drug-gene and pathway feature sets optimal solution. Individual gene network scores are shown below each gene name.

For node2vec-based pathway set results, four distinct pathways were obtained per setup. Similar to edge2vec results, the Creatine Metabolism pathway was also identified as the optimal solution for S2 (AUC=62%). This model also indicates a strong individual contribution from *CKM* gene. For S1, the best solution was TGF-beta Receptor 1 (*TGFBR1*) Loss-of-Function pathway (AUC=60%). It includes 4 genes, primarily TGF-beta receptors. S3 best solution was the Metastatic Brain Tumor Process with a five-gene set: *E2F3*, *MYC*, *CDC42*, *CDK6*, and *TP53*, and AUC=60%. S4 yielded an optimal solution of Prednisolone Metabolism Pathway (AUC=60%) containing two genes, *HSP90AA1* and *NR3C1*.

3.3. *Functional analysis of drug-gene and pathway feature set models*

The gene set enrichment analysis results on the optimal set of solutions for the drug-gene feature sets are presented in Table 2. Multiple drug–gene sets were significantly enriched in genes associated with specific body parts and biological pathways. For edge2vec, the Dalfampridine gene set was enriched for brain-associated genes as well as pathways related to the neuronal system, voltage-gated and potassium channels (Table 2). The NADH-associated gene sets from S3 and S4 were enriched for blood- and heart-associated genes, along with key metabolic

Table 2. Enrichment analysis for drug-gene feature sets

Setup	Drug name	Size	AUC	Comp.	Body parts	Pathways
Enrichment analysis for edge2vec						
S1	Dalfampridine	40	60.8	1.63	Brain (27)*	Neuronal System (40), Voltage gated Potassium channels (39), Potassium Channels (39)*
S1	Amifampridine	42	61.5	4.24		Neuronal System (40), Voltage gated Potassium channels (39), Potassium Channels (39)*
S2	Guanidine	49	60.8	1.72		Neuronal System (43), Voltage gated Potassium channels (39), Potassium Channels (39)*
S2	Dalfampridine	40	61.0	1.46	Brain (27)*	Neuronal System (40), Voltage gated Potassium channels (39), Potassium Channels (39)*
S3	Pyridoxal phosphate	18	60.6	1.21		Metabolic pathways (17)*
S3	Aspartic acid	8	60.5	1.04		Alanine, aspartate and glutamate metabolism (4)*
S3	NADH	87	60.6	1.95	Blood (67), Heart (63)*	Metabolic pathways (81)*
S3	Guanidine	49	60.7	1.70		Neuronal System (43), Voltage gated Potassium channels (39), Potassium Channels (39)*
S3	Dalfampridine	40	60.7	1.43	Brain (27)*	Neuronal System (40), Voltage gated Potassium channels (39), Potassium Channels (39) *
S4	Pyridoxal phosphate	18	60.5	1.32		Metabolic pathways (17)*
S4	NADH	87	60.6	1.95	Blood (67), Heart (63)*	Metabolic pathways (81)*
S4	Dalfampridine	40	60.5	1.26	Brain (27)*	Neuronal System (40), Voltage gated Potassium channels (39), Potassium Channels (39)*
Enrichment analysis for node2vec						
S1	Vindesine	22	59.9	1.11		p53 signaling pathway (5)*
S1	Carbinoxamine	12	60.2	5.75		GnRH signaling pathway (4)*
S1	Phensuximide	6	60.0	0.95		Axon guidance (4)*
S1	Thioproperazine	7	60.4	1.00		
S1	Flubendazole	15	60.0	1.00	Endometrium (11)*	Imipramine Action Pathway (6)*
S2	Clotrimazole	67	61.2	4.67		Signal Transduction (35)*
S2	Clozapine	62	61.8	5.92	Brain (38)*	
S2	Norepinephrine	22	60.7	0.98		Monoamine GPCRs (12)*
S2	Haloprogin	33	61.1	5.77		Signal Transduction (16)*
S2	Dopamine	32	61.0	5.77		Signal Transduction (15)*
S3	Flunisolide	32	60.3	0.85	Adrenal cortex (19)*	Nuclear Receptors Meta-Pathway (8)*
S3	Vindesine	22	60.2	1.40		p53 signaling pathway (5)*
S3	Nitrofural	10	60.4	5.50		Neuroactive ligand-receptor interaction (38)*
S3	Carbenicillin	4	60.3	0.70		Signaling by GPCR (3)*
S3	Cefixime	18	60.9	5.75		Proton/oligonucleotide cotransporters (2)*
S3	Clobetasol propionate	35	60.5	1.19	Adrenal cortex (19)*	Nuclear Receptors Meta-Pathway (8)*
S3	Moclobemide	18	61.1	5.50		Integrated Pancreatic Cancer Pathway (4)*
S3	Thioproperazine	7	60.4	1.00		
S3	Flubendazole	15	60.2	0.78	Endometrium (11)*	Imipramine Action Pathway (6)*
S3	Brigatinib	14	60.2	1.20		Pathways in cancer (6)*
S4	Pseudoephedrine	11	59.9	1.04		Signal Transduction (7)*
S4	Apraclonidine	6	59.9	0.85		GPCR downstream signaling (6)*
S4	Lisdexamfetamine	5	59.9	0.90		SLC-mediated transmembrane transport (4)*
S4	Dextroamphetamine	10	60.1	1.11		Imipramine Action Pathway (5)*

* *FDR correction < 0.05 for body parts and pathways.*

pathways. Similarly, the Guanidine models (S2 and S3) and the Amifampridine model (S1) showed enrichment for neuronal system components and ion channel activity. For the node2vec results, the Clozapine model from S2 showed significant enrichment for brain-related genes. Several models, across all network score types, were enriched for signaling pathways. This included Vindesine (S1, S3), Carbinoxamine (S1), Clotrimazole, Haloprogin, and Dopamine (S2), Carbenicillin (S3), and both Pseudoephedrine and Apraclonidine (S4).

Pathway gene sets for both edge2ve and node2vec were not enriched for any body partassociated genes. Most of the pathways feature sets were enriched with genes associated with the same pathway or a similar process (see supplementary file [c]). Creatinine metabolism and related pathways, which included the *CKM* gene, were presented across multiple setups in both edge2vec and node2vec results.

4. Discussion and Conclusion

This study presents a random-walk-based learning framework that identifies potential novel gene candidates for Alzheimer's disease therapeutics. Beginning with the heterogeneous knowledge graph Hetionet, we learned embeddings for druggable gene nodes and the AD node using two random walk methods, edge2vec and node2vec. The embeddings learned from edge2vec and node2vec captured different structural aspects of the network. Gene–AD similarities were quantified using the cosine similarity metric. The cosine similarity between the gene and AD embeddings provided a relative ranking of genes and served as the basis for computing the network scores to prioritize genes relevant to AD. We introduced four variants of the network score calculations that reflected varying levels of constraints (AD gene proximity penalty and/or exclusion of AD genes from sets). This resulted in four experimental setups (S1, S2, S3, and S4), with S1 being the least constrained one.

To validate the utility of these learned network scores in identifying novel drug targets, we constructed AD predictive models using ADSP genetic data. We employed TPOT2 with three objectives (predictive performance, complexity, and network scores) and FSS to explore the optimal solutions for both drug-gene interaction and pathway-gene feature sets.

TPOT2 identified multiple non-dominated (Pareto front) models for each experimental setup (Table1, Fig.4). Overall, experiments on drug-gene feature sets yielded more Pareto-optimal solutions and tended to produce more complex models, reflecting the use of larger gene sets. These experiments also included the higher-performing models (based on the AUC metric) across all setups. In contrast, experiments on pathway feature sets produced simpler models, likely as a result of smaller gene sets, yet achieved comparable predictive performance.

The optimal solutions across experiments demonstrated the diversity of feature sets enabled by network score gene differentiation (Fig.5). Optimal solution convergence was observed predominantly in drug-gene feature set edge2vec analyses for Dalfampridine, a potassium channel blocker. This feature set was primarily composed of voltage-gated potassium channel genes, which play critical roles in regulating diverse physiological and pathophysiological processes. These channels have been previously proposed as therapeutic targets for several conditions, including atrial fibrillation, epilepsy, neuropathic pain, and potentially neuropsychiatric disorders.[25] For node2vec-based drug-gene feature set analyses, the optimal solutions varied across setups. Notably, the S1 experiment identified the Phensuximide drug-gene feature set, consisting of six genes, including voltage-gated calcium channels. These channels play a critical role in nervous system function at both cellular and network levels and have been proposed as therapeutic targets for conditions such as Parkinson's disease, drug addiction, pain, anxiety, and epilepsy.[26] Pathway feature set analyses all found *CKM* (Creatine Kinase M-type) as the

[c] Supplementary information: `https://github.com/EpistasisLab/ADSP_Network_Score/`

most informative contributor to the optimal model for the creatine metabolism pathway. Notably, this gene alone was sufficient to predict AD with a 62% AUC, as reported in the Pareto front for the S4 edge2vec pathway feature set analyses. *CKM* isoenzymes are central to energy transduction in tissues with high and fluctuating energy demands, including skeletal muscle, heart, and brain. Although *CKM* has not been previously reported as a direct therapeutic target, it has been suggested as a biomarker of therapeutic effects in similar neurodegenerative Parkinson's disease.[27]

Gene set enrichment analysis revealed that several near-optimal solutions were significantly enriched for genes related to neurological functions. In the edge2vec drug-gene feature set analysis, the Dalfampridine gene set, selected across all experimental setups, was enriched for brain-associated genes and pathways related to the neuronal system and potassium channel activity (Table 2). Other edge2vec drug-gene feature sets (Amifampridine, Guanidine) and node2vec drug-gene feature sets (Clozapine and Lisdexamfetamine) were also linked to neuronal processes.

Both random-walk methods had comparable performance in terms of AUC (see Table 1), although node2vec[14] was computationally more efficient. However, edge2vec[15] retains edge types (relationships) between biomedical entities, such as drugs that upregulate versus downregulate gene expression levels, which provides an additional advantage in potentially capturing biologically meaningful results. This is reflected in the larger spectrum of results identified by node2vec compared to edge2vec which was fewer but more consistent (see Table 2). A limitation of this study is that we did not explore graph-based methods beyond node2vec and edge2vec due to computational and scalability constraints. Future work could examine other approaches capable of learning meaningful embeddings for nodes in a biomedical knowledge graph, such as GNNs[28,29] and representation learning models (e.g., RotateE,[30] TransE[31]). In addition, we did not examine alternative features commonly used in drug target discovery, such as molecular structure embeddings for drugs, which are more suitable for GNNs or representation learning methods.

Overall, the results demonstrate that the proposed random-walk-based learning method, integrated with an informed machine learning approach powered by TPOT, successfully identified functionally diverse gene sets. The achieved predictive performance is comparable to that of established AD risk factors,[10] underscoring its potential for the discovery of novel genetic candidates. This approach is generalizable to other diseases where structured knowledge graphs are available to guide drug repurposing efforts. Ultimately, we aim to develop a system that automatically learns meaningful semantic representations from heterogeneous knowledge graphs to profile AD patients and prioritize druggable gene candidates. Future work will explore alternative embedding methods and integrate additional biologically relevant features.

Acknowledgements

Data for this study were prepared, archived, and distributed by the National Institute on Aging Alzheimer's Disease Data Storage Site (NIAGADS) at the University of Pennsylvania (U24-AG041689), funded by the National Institute on Aging.

References

1. J. Cummings, T. Morstorf and K. Zhong, Alzheimer's disease drug-development pipeline: few candidates, frequent failures. alzheimers res ther. 2014; 6 (4): 37.
2. S. Mukhopadhyay and D. Banerjee, A primer on the evolution of aducanumab: the first antibody approved for treatment of alzheimer's disease, *Journal of Alzheimer's Disease* **83**, 1537 (2021).
3. C. H. Van Dyck, C. J. Swanson, P. Aisen, R. J. Bateman, C. Chen, M. Gee, M. Kanekiyo, D. Li, L. Reyderman, S. Cohen *et al.*, Lecanemab in early alzheimer's disease, *New England Journal of Medicine* **388**, 9 (2023).
4. S. J. Andrews, A. E. Renton, B. Fulton-Howard, A. Podlesny-Drabiniok, E. Marcora and A. M. Goate, The complex genetic architecture of alzheimer's disease: novel insights and future directions, *EBioMedicine* **90** (2023).
5. S. Pushpakom, F. Iorio, P. A. Eyers, K. J. Escott, S. Hopper, A. Wells, A. Doig, T. Guilliams, J. Latimer, C. McNamee *et al.*, Drug repurposing: progress, challenges and recommendations, *Nature reviews Drug discovery* **18**, 41 (2019).
6. R. Gupta, D. Srivastava, M. Sahu, S. Tiwari, R. K. Ambasta and P. Kumar, Artificial intelligence to deep learning: machine intelligence approach for drug discovery, *Molecular diversity* **25**, 1315 (2021).
7. D. S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S. L. Chen, D. Hadley, A. Green, P. Khankhanian and S. E. Baranzini, Systematic integration of biomedical knowledge prioritizes drugs for repurposing, *Elife* **6**, p. e26726 (2017).
8. P. Chandak, K. Huang and M. Zitnik, Building a knowledge graph to enable precision medicine, *Scientific Data* **10**, p. 67 (2023).
9. J. D. Romano, V. Truong, R. Kumar, M. Venkatesan, B. E. Graham, Y. Hao, N. Matsumoto, X. Li, Z. Wang, M. D. Ritchie *et al.*, The Alzheimer's Knowledge Base: A Knowledge Graph for Alzheimer Disease Research, *Journal of Medical Internet Research* **26**, p. e46777 (2024).
10. A. Orlenko, M. Venkatesan, L. Shen, M. D. Ritchie, Z. P. Wang, T. Obafemi-Ajayi and J. H. Moore, Biologically enhanced machine learning model to uncover novel gene-drug targets for alzheimer's disease, in *Biocomputing 2025: Proceedings of the Pacific Symposium*, 2024.
11. D. Bang, S. Lim, S. Lee and S. Kim, Biomedical knowledge graph learning for drug repurposing by extending guilt-by-association to multiple layers, *Nature Communications* **14**, p. 3570 (June 2023).
12. M. Cannon, J. Stevenson, K. Stahl, R. Basu, A. Coffman, S. Kiwala, J. F. McMichael, K. Kuzma, D. Morrissey, K. Cotto *et al.*, DGIdb 5.0: rebuilding the drug–gene interaction database for precision medicine and drug discovery platforms, *Nucleic acids research* **52**, D1227 (2024).
13. L. Wu, P. Cui, J. Pei, L. Zhao and X. Guo, Graph neural networks: foundation, frontiers and applications, in *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 2022.
14. A. Grover and J. Leskovec, node2vec: Scalable feature learning for networks, in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016.
15. Z. Gao, G. Fu, C. Ouyang, S. Tsutsui, X. Liu, J. Yang, C. Gessner, B. Foote, D. Wild, Y. Ding *et al.*, edge2vec: Representation learning using edge semantics for biomedical knowledge discovery, *BMC bioinformatics* **20**, p. 306 (2019).
16. Z. Sun, Z.-H. Deng, J.-Y. Nie and J. Tang, Rotate: Knowledge graph embedding by relational rotation in complex space, *arXiv preprint arXiv:1902.10197* (2019).
17. D. Bang, S. Lim, S. Lee and S. Kim, Biomedical knowledge graph learning for drug repurposing by extending guilt-by-association to multiple layers, *Nature Communications* **14**, p. 3570 (June 2023).
18. P. Ribeiro, A. Saini, J. Moran, N. Matsumoto, H. Choi, M. Hernandez and J. H. Moore, TPOT2:

A New Graph-Based Implementation of the Tree-Based Pipeline Optimization Tool for Automated Machine Learning, in *Genetic Programming Theory and Practice XX*, (Springer, 2024) pp. 1–17.

19. T. T. Le, W. Fu and J. H. Moore, Scaling tree-based automated machine learning to biomedical big data with a feature set selector, *Bioinformatics* **36**, 250 (2020).
20. K. Deb, A. Pratap, S. Agarwal and T. Meyarivan, A fast and elitist multiobjective genetic algorithm: Nsga-ii, *IEEE transactions on evolutionary computation* **6**, 182 (2002).
21. Y. Y. Leung, W.-P. Lee, A. B. Kuzma, P. Gangadharan, H. I. Nicaretta, L. Qu, Y. Ren, L. B. Cantwell, O. Valladares, Y. Zhao *et al.*, Adsp whole genome sequencing (wgs) release 4 data update from genome center for alzheimer's disease, *Alzheimer's & Dementia* **19**, p. e077351 (2023).
22. W. Zhou, J. B. Nielsen, L. G. Fritsche, R. Dey, M. E. Gabrielsen, B. N. Wolford, J. LeFaive, P. VandeHaar, S. A. Gagliano, A. Gifford *et al.*, Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies, *Nature genetics* **50**, 1335 (2018).
23. C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell and J. J. Lee, Second-generation plink: rising to the challenge of larger and richer datasets, *Gigascience* **4**, s13742 (2015).
24. W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, A. Thormann, P. Flicek and F. Cunningham, The ensembl variant effect predictor, *Genome biology* **17**, p. 122 (2016).
25. H. Wulff, N. A. Castle and L. A. Pardo, Voltage-gated potassium channels as therapeutic targets, *Nature reviews Drug discovery* **8**, 982 (2009).
26. G. W. Zamponi, Targeting voltage-gated calcium channels in neurological and psychiatric diseases, *Nature reviews Drug discovery* **15**, 19 (2016).
27. Y. Gong, S. Qian, D. Chen, M. Ye, J. Wu and Y.-l. Wang, Serum blmh and ckm as potential biomarkers for predicting therapeutic effects of deep brain stimulation in parkinson's disease: a proteomics study, *Journal of Integrative Neuroscience* **22**, p. 163 (2023).
28. M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov and M. Welling, Modeling Relational Data with Graph Convolutional Networks, in *The Semantic Web*, eds. A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai and M. Alam (Springer International Publishing, Cham, 2018).
29. Z. Hu, Y. Dong, K. Wang and Y. Sun, Heterogeneous Graph Transformer, in *Proceedings of The Web Conference 2020*, WWW '20 (Association for Computing Machinery, New York, NY, USA, April 2020).
30. Z. Sun, Z.-H. Deng, J.-Y. Nie and J. Tang, RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space, in *International Conference on Learning Representations*, September 2018.
31. A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston and O. Yakhnenko, Translating embeddings for modeling multi-relational data, in *Advances in Neural Information Processing Systems*, eds. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Weinberger (Curran Associates, Inc., 2013).