OXFORD

Briefings in Bioinformatics, 2025, **26(1)**, bbae704
https://doi.org/10.1093/bib/bbae704
Problem Solving Protocol

Genome-wide association neural networks identify genes linked to family history of Alzheimer's disease

Upamanyu Ghose[1,2,*], William Sproviero[1], Laura Winchester[1], Najaf Amin[3], Taiyu Zhu[1], Danielle Newby[1,4], Brittany S. Ulm[2,3], Angeliki Papathanasiou[1], Liu Shi[1,5], Qiang Liu[1,6], Marco Fernandes[1,7], Cassandra Adams[2,8], Ashwag Albukhari[2,9], Majid Almansouri[2,10], Hani Choudhry[2,9], Cornelia van Duijn[2,3], Alejo Nevado-Holgado[1,2]

[1]Department of Psychiatry, University of Oxford, Oxford, United Kingdom
[2]King Abdulaziz University and the University of Oxford Centre for Artificial Intelligence in Precision Medicine (KO-CAIPM), Jeddah, Saudi Arabia
[3]Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom
[4]Centre for Statistics in Medicine, Nuffield Department of Orthopedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom
[5]Department of Translational Medicine, Nxera Pharma UK Limited, Cambridge, United Kingdom
[6]School of Engineering Mathematics and Technology University of Bristol, Ada Lovelace Building, Bristol, United Kingdom
[7]School of Medicine, University of St Andrews, St Andrews, United Kingdom
[8]Centre for Medicines Discovery, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom
[9]Biochemistry Department, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia
[10]Clinical Biochemistry Department, Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia

*Corresponding author: Department of Psychiatry, University of Oxford, Warneford Hospital, Oxford OX3 7JX, United Kingdom.
E-mail: upamanyu.ghose@psych.ox.ac.uk

Abstract

Augmenting traditional genome-wide association studies (GWAS) with advanced machine learning algorithms can allow the detection of novel signals in available cohorts. We introduce "genome-wide association neural networks (GWANN)" a novel approach that uses neural networks (NNs) to perform a gene-level association study with family history of Alzheimer's disease (AD). In UK Biobank, we defined cases ($n = 42\,110$) as those with AD or family history of AD and sampled an equal number of controls. The data was split into an 80:20 ratio of training and testing samples, and GWANN was trained on the former followed by identifying associated genes using its performance on the latter. Our method identified 18 genes to be associated with family history of AD. *APOE*, *BIN1*, *SORL1*, *ADAM10*, *APH1B*, and *SPI1* have been identified by previous AD GWAS. Among the 12 new genes, *PCDH9*, *NRG3*, *ROR1*, *LINGO2*, *SMYD3*, and *LRRC7* have been associated with neurofibrillary tangles or phosphorylated tau in previous studies. Furthermore, there is evidence for differential transcriptomic or proteomic expression between AD and healthy brains for 10 of the 12 new genes. A series of *post hoc* analyses resulted in a significantly enriched protein–protein interaction network (P-value $< 1 \times 10^{-16}$), and enrichment of relevant disease and biological pathways such as focal adhesion (P-value $= 1 \times 10^{-4}$), extracellular matrix organization (P-value $= 1 \times 10^{-4}$), Hippo signaling (P-value $= 7 \times 10^{-4}$), Alzheimer's disease (P-value $= 3 \times 10^{-4}$), and impaired cognition (P-value $= 4 \times 10^{-3}$). Applying NNs for GWAS illustrates their potential to complement existing algorithms and methods and enable the discovery of new associations without the need to expand existing cohorts.

Keywords: Alzheimer's disease; neural networks; artificial intelligence; machine learning; GWAS; UK Biobank

Introduction

Alzheimer's disease (AD) affects ~30 million people in the world, making it the most common form of dementia [1]. It is characterized by the build-up of Aβ and tau proteins in the brain, leading to neuronal death and impaired cognitive function [2]. In the last 10 years, genome-wide association studies (GWAS) have revolutionized our understanding of the inherited basis of disease, and they have been critical in identifying multiple risk loci and novel disease pathways associated with AD involving the microglia and lysosome [3]. However, the classical GWAS analysis depends on sample size, and, despite the number of single-nucleotide polymorphisms (SNPs) identified until today, they still only explain a fraction of the heritability of the disease [4]. Gene-based methods have been developed to identify the joint effects of rare variants [5, 6] and common variants [7] and gene-level analysis from GWAS summary data [8]. However, there are currently no methods to perform gene-based discovery using machine learning methods and genetic data.

Along with the modern availability of large datasets [9–12], to complement and enhance current GWAS methods, we propose to use an approach based on machine learning to shed light on more complex patterns in genomic mechanisms involving gene interactions and nonlinear relationships. Machine learning methods, more particularly neural networks (NNs), have been instrumental in the advancement of multiple engineering industries due to their efficacy in analyzing complex data patterns [13, 14], especially where large amounts of data are available.

Received: September 23, 2024. **Revised:** November 29, 2024. **Accepted:** December 23, 2024

Compared to the success of classical GWAS, the success of NNs in gene discovery has been limited. Using knowledge from prior GWAS or a shortlisted set of SNPs obtained from linear models, machine learning methods such as stacked NNs [15], ensemble architectures [16], and support vector machines [17] have been used to achieve a better disease risk prediction when compared with polygenic risk scores. NNs have also recently been employed and tested on various complex traits and diseases including eye color and schizophrenia [18]. However, our aim in this work was to develop NNs specialized to perform a gene-level association study using SNP data available in the UK Biobank (UKB) [19]. The objective was to approach this task without the use of information from prior GWAS. Our method is gene-based and considers groups of SNPs within and around each gene in the genome to establish the association of the gene with the phenotype of interest. Since the method tested the association of a single gene at a time, the objective was not to achieve better disease prediction performance than previously mentioned methods, as this would not be possible, given the polygenic nature of AD. In this paper, we demonstrate the application of our new genome-wide association neural networks (GWANN) method as a complementary method to existing GWAS methods, to identify associations with family history of AD/dementia, a proxy that has been successfully used to identify new genes for AD in the UKB [20, 21]. We present the genetic associations to family history of AD found by the method and systematically support the results with *post hoc* enrichment analyses using transcriptomic data from postmortem AD brains, biological pathways and gene ontologies, protein–protein interaction (PPI) data, disease and trait gene sets, and data about target tractability for drug development.

Materials and Methods

Population

We utilized data from the UKB (http://www.ukbiobank.ac.uk). The data comprise health, cognitive, and genetic data collected from ∼500 000 individuals aged between 37 and 73 years from the UK at the study baseline (2006–2010) [19, 22]. We used imputed SNP genotype data as input to GWANN. UKB genotyping was conducted by Affymetrix using a BiLEVE Axiom array for 49 950 participants and further updated using an Affymetrix Axiom array for the remaining 438 427 individuals, based on the first array (95% marker content shared). The released genotyped data contained 805 426 markers on 488 377 individuals. Information on the genotyping process is available on the UKB website (http://www.ukbiobank.ac.uk/scientists-3/genetic-data) [22]. Genotype imputation was performed by combining the UK10K haplotype and Haplotype Reference Consortium as reference panels [23]. A number of individuals ($n = 856$) either with inconsistencies between their genetic predicted and reported sex or abnormal number of sex chromosomes were removed. In addition, 968 outliers were identified based on heterozygosity and missingness and removed. The dataset was further limited to only individuals of "White British" descent resulting in 409 703 remaining individuals. A genetic relationship matrix along with genome-wide complex trait analysis was used to identify 131 818 individuals with relatives within the dataset, using a relationship threshold of 0.025. Only one person from each pair of related individuals was retained. Only biallelic SNPs with minor allele frequency (MAF) $> 1\%$ and imputation quality info score > 0.8 were retained for the analysis, and all indels and multi-allelic SNPs were dropped. For the analysis, we used the imputed genotype dosages.

Definition of cases and controls

The cases were defined as individuals with AD diagnosis ($n = 1176$) or parental history of dementia ($n = 40\,934$). The parental histories of dementia were defined according to a previous study on family history of AD [20]. Individuals with other neurological disorders (Supplementary Table S1) [24] were removed from the control groups. We divided the entire range of ages into three groups (age-group1: 38–52, age-group2: 53–61, age-group3: 62–73 years) and paired them with the two possibilities of sex (male and female) to obtain six broad groups—(age-group1, male), (age-group1, female), etc. An equal number of controls ($n = 42\,110$) were randomly sampled while balancing for the six broad groups. Eighty percent ($n = 67\,380$) of the data were used to train the NNs, and 20% ($n = 16\,840$) of the data were reserved as a held-out test to evaluate the performance of the NNs and ascertain association with the phenotype.

Training the neural network model

Gene locations were mapped according to GRCh37/hg19. For every gene, SNPs within the gene and in the 2500 bp flanking region were considered. Since NNs are computationally more intensive than linear models, we set the limit to 2500 bp as a trade-off between increased computational time and including downstream and upstream SNPs in the analysis. This also minimized the chances of overlap between genes that are very close to each other. We divided every gene into windows of maximum 50 SNPs, and the final analysis was done on all windows of all genes. A different NN was trained for each window per gene in the entire genome. This resulted in 70 848 NNs. In addition to the SNPs, age (field 21003), sex (field 31), the first six genetic principal components obtained from UKB variables (field 22009), and education qualification (field 6138) were used as covariates. Education qualification was transformed into years of education using the International Standard Classification of Education encoding.

Each sample for GWANN consisted of SNPs and covariates for a homogenous group of 10 cases or controls, enabling the NNs of GWANN to identify similar patterns across the individuals in the group. The NN was trained to predict if a group was formed by cases or controls (Fig. 1). Since NNs are inherently stochastic, for each window, the method was run 16 times with different random seeds to get a stable aggregate performance metric on the held-out test set and to then determine the level of statistical significance of this metric being significantly above chance predictions of family history of AD. The aggregate metric was compared to a null distribution obtained from simulated data generated using the "dummy" method of PLINK 2.0 [25] to obtain a *P*-value. An empirical threshold, $\theta_1 = 1 \times 10^{-25}$, was determined such that 95% of the gene windows with *P*-value $< \theta_1$ would also satisfy *P*-value $< 7.06 \times 10^{-7}$—the Bonferroni-corrected genome-wide significance threshold—if the method were repeated another 16 times with different random seeds. This was to ensure that only the most confident hits were reported as significant associations. The negative log likelihood (NLL) of the NNs were used as the test metrics to evaluate significance and for all *post hoc* analyses. If multiple windows of a gene were significant, the window with the best test metric was selected. Further information about

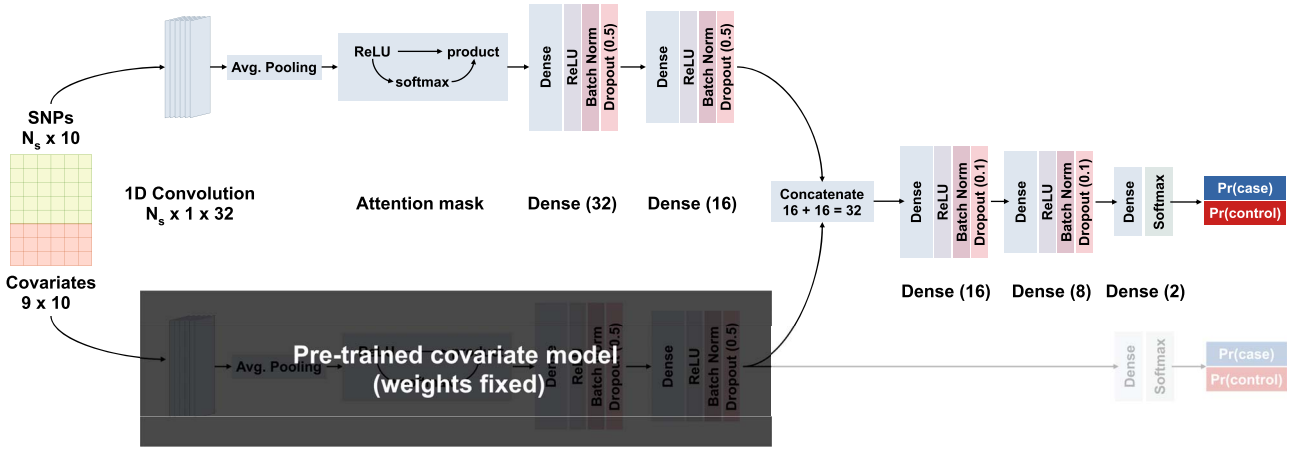

Figure 1. NN architecture used in the GWANN method. The top-left branch generates a 1D encoding from the SNP input, while the bottom-left branch does so for the covariate input. The right trunk merges the encodings of both branches to output whether the input belongs to cases or controls.

Table 1. GWANN hit genes associated with family history of AD.

Gene	Genomic interval	P-value	Gene	Genomic interval	P-value
APOE	19:45406673–45414451	6.95×10^{-160}	*SRGAP2B*	1:144042910–144042910	1.06×10^{-39}
BIN1	2:127863681–127867174	3.07×10^{-92}	*HSP90AB4P*	15:58981684–58987724	5.71×10^{-36}
NRG3	10:83832534–83855173	2.09×10^{-57}	*LINGO2*	9:28386903–28396747	1.11×10^{-34}
LRRC7	1:70399066–70428546	3.1×10^{-53}	*PALD1*	10:72301529–72311482	1.70×10^{-34}
ROR1	1:64591756–64611493	2.07×10^{-51}	*PCDH9*	13:67345123–67362138	7.27×10^{-32}
RPS6KC1	1:213372832–213408630	1.68×10^{-50}	*ADAM10*	15:59012608–59042081	1.69×10^{-31}
APH1B	15:63589709–63603802	9.95×10^{-47}	*SPI1*	11:47374633–47390692	1.01×10^{-30}
AKR1C6P	10:4942736–4956083	1.24×10^{-46}	*SMYD3*	1:245922632–245930742	1.80×10^{-28}
SORL1	11:121432788–121448972	9.62×10^{-46}	*SYNPO*	5:150015017–150033470	1.62×10^{-27}

P-values lower than 6.95×10^{-159} have been cropped to a value of 6.95×10^{-160}.

the neural network architecture, model training, and significance establishment can be found in Supplementary Information.

Results

Identification of genes related to family history of Alzheimer's disease using GWANN

On applying GWANN to family history of AD in the UKB, 32 genes passed the empirical significance threshold before pruning for linkage disequilibrium (LD). After identifying LD blocks (genes with $r^2 \geq 0.8$) among these genes, we retained the gene with the best test metric within a block as the hit gene. This resulted in narrowing down to 18 associated genes (Fig. 2, Table 1, Supplementary Table S2). Among these hits, *APOE*, *BIN1*, *ADAM10*, *SORL1*, *SPI1*, and *APH1B* have been previously associated with AD by large GWAS [3] (Fig. 3). In addition to these AD-associated genes, *LINGO2*, *LRRC7*, *NRG3*, *PCDH9*, *ROR1*, and *SMYD3* have been previously identified via SNP × SNP interaction studies to be associated with phosphorylated tau [3]. Six genes, *SYNPO*, *SRGAP2B*, *PALD1*, *AKR1C6P*, *HSP90AB4P*, and *RPS6KC1*, had no evidence for previous GWAS association with AD or AD-related traits. To further understand the 12 new GWANN hits, we obtained information about them from the Agora AD knowledge portal (Fig. 3b). Besides *PCDH9* and *AKR1C6P*, all hits had evidence for differential transcriptomic or proteomic expression between postmortem AD and healthy brains. RNAseq levels of *PCDH9* had evidence of association with a clinical consensus diagnosis of cognitive status at time of death (COGDX).

We also performed a GWAS using PLINK 2.0 [25] on the same data that was used for GWANN (TradGWAS). After LD pruning, TradGWAS identified *APOE* and *SORL1* as significant genes (P-value $< 5 \times 10^{-8}$), both of which were identified by GWANN. When compared with the genes identified by the largest AD GWAS run in the European population using the European Alzheimer & Dementia Biobank (EADB GWAS) [12], GWANN had an overlap of five genes (*APOE*, *SORL1*, *APH1B*, *BIN1*, *SPI1*), and TradGWAS had an overlap of two genes (*APOE*, *SORL1*) (Fig. 3c). We also looked at the overlap with the EADB GWAS hit genes using the top 100 genes from GWANN and TradGWAS (Fig. 3d). This showed an overlap of seven genes (*APOE*, *SORL1*, *BIN1*, *ABCA7*, *BCKDK*, *UFC1*, *CR1*) between the EADB GWAS and TradGWAS and seven genes (*APOE*, *SORL1*, *BIN1*, *ABCA7*, *APH1B*, *SPI1*, *CTSH*) between the EADB GWAS and GWANN. If an intergenic hit SNP in the EADB GWAS was not deterministically mapped to either the upstream or downstream gene, both were considered when calculating the overlap. The EADB GWAS reported 89 hit loci, but since we calculated the overlap on a gene level, we used the 84 unique genes that these loci were mapped to and added *APOE* to the list of hits. We also tested the top 100 genes for association in the "Asian or Asian British" and "Black or Black British" populations of the UKB (Supplementary Table S3). The number of samples in each population was >20 times smaller than the "White British" population, with 358 cases in the Asian population and 340 cases in the Black population. *APOE* and *AKR1C6P* were significant in the Asian population, and *APOE* was significant in the Black population at the Bonferroni threshold of P-value $< 5 \times 10^{-4}$.

Enriched biological pathways, gene ontology terms, diseases, and protein–protein interaction network

Gene set enrichment analysis (GSEA) [26] was applied to the GWANN summary test metrics for all genes to identify enriched

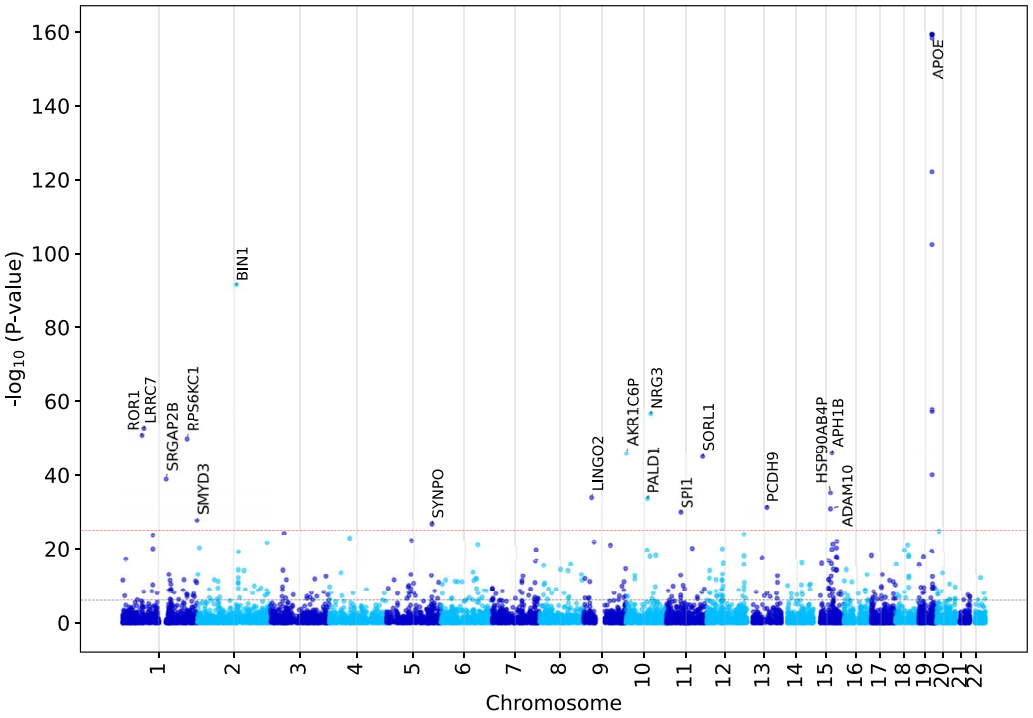

Figure 2. Manhattan plot after running GWANN on family history of AD. Significant hits were identified at an empirically determined *P*-value threshold of *P*-value $<1 \times 10^{-25}$ (top dotted line). After calculating the LD between significant genes, the gene with the best NLL within an LD block was identified as the hit gene. The *P*-values lower than 6.95×10^{-159} have been cropped to a value of 6.95×10^{-160}. The bottom dotted line marks the Bonferroni-corrected threshold for the number of gene windows that were tested, *P*-value = 7.06×10^{-7}.

pathways in Reactome, Wiki, Kyoto Encyclopaedia of Genes and Genomes (KEGG), and Gene Ontology (GO) gene sets obtained from MSigDB [27] (Fig. 4a–d, Supplementary Table S4). GSEA calculates the normalized enrichment score (NES) based on the test metric of all genes analyzed using a Kolmogorov–Smirnov-like test [26]. Hence, some pathways had a significant NES due to the cumulative contribution of genes that were nominally significant but not among the list of 18 GWANN hits. The enriched pathways with the GWANN hits present in the GSEA leading edge were extracellular matrix organization (*P*-value = 1.04×10^{-4}), signaling by receptor tyrosine kinases (*P*-value = 7.64×10^{-4}), axon guidance (*P*-value = 2.23×10^{-3}), diseases of signal transduction by growth factor receptors and second messengers (*P*-value = 1.11×10^{-2}), and ErbB signaling (*P*-value = 2.29×10^{-2}). Some of the most enriched GO terms with GWANN hits in the GSEA leading edge were regulation of neuron projection development (*P*-value = 7.03×10^{-8}), glutamatergic synapse (*P*-value = 2.64×10^{-6}), synapse organization (*P*-value = 5.66×10^{-6}), distal axon (*P*-value = 7.77×10^{-5}), and regulation of synapse structure or activity (*P*-value = 7.72×10^{-4}).

Disease and trait enrichment was performed using DisGeNET [28]. We used the top 100 genes (without LD pruning) ranked by the test metric for this analysis and filtered out diseases and traits with >5000 genes mapped to them (Fig. 4e, Supplementary Table S5). Some of the most enriched traits with the largest number of overlapping genes were AD (FDR = 2.56×10^{-4}), impaired cognition (FDR = 4.23×10^{-3}), autism spectrum disorders (FDR = 1.19×10^{-2}), and mental deterioration (FDR = 3.50×10^{-4}).

Using the same set of genes as used for the disease and trait enrichment, we generated a PPI network using STRING [29] (Fig. 4f). Some of the gene symbols were not recognized by the STRING protein database, leaving a set of 88 genes that were accepted. The resultant PPI network was significantly enriched with 72 edges (*P*-value $< 1 \times 10^{-16}$). Given a network of 88 proteins, the expected number of edges for a set of randomly selected proteins is 21, thereby rendering the GWANN PPI network to have significantly more connections than an equivalent network of random proteins. The network in Fig. 4f shows each protein colored by the group (Supplementary Table S6) of experimental factor ontology traits it enriched.

We also performed the same enrichments for TradGWAS (Supplementary Fig. S1). Pathways related to ErbB signaling and calcium signaling overlapped with GWANN. The other top enriched pathways were mainly related to cholesterol, lipids, and lipoproteins (Supplementary Fig. S1a–d). The PPI network showed similar levels of enrichment to the GWANN PPI network (*P*-value $< 1 \times 10^{-16}$, Supplementary Fig. S1e).

Enrichment of transcriptomic data from Alzheimer's disease postmortem brains using GWANN hits

We examined two studies on differential transcriptomic expression in AD brains. Patel *et al.* [30] reported differentially expressed genes (DEGs) between AD cases, controls, and non-AD mental disorders in multiple brain regions. We analyzed their DEGs for AD versus controls, non-AD mental disorders versus controls, and AD-specific DEGs. Patel *et al.* [31] identified DEGs for asymptomatic AD versus controls, symptomatic AD versus controls, and symptomatic versus asymptomatic AD in various brain regions. We applied GSEA to assess DEG set enrichment.

In the first study (Patel *et al.*—A) [30], all brain regions showed enrichment for AD versus controls and "only AD" versus controls, while the cerebellum and parietal lobes were enriched for "non-AD" versus controls (Table 2). Key GWANN hits in the GSEA leading edge were *PCDH9*, *APOE*, *SORL1*, and *LRRC7*. The temporal lobe, despite enrichment, had no GWANN hits in the leading edge but included *UFC1* and *MAP2K1* (nominal GWANN hits). AD versus

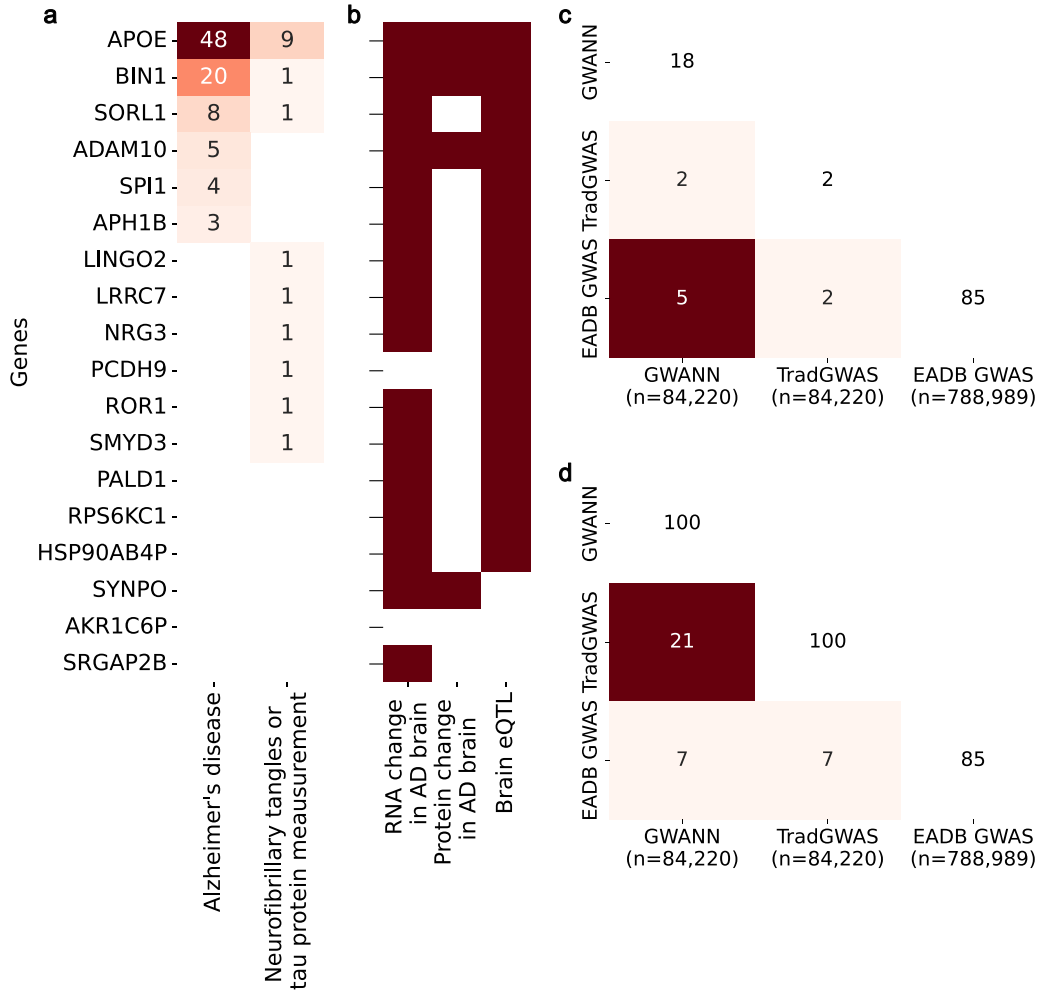

Figure 3. Overlap of GWANN hits with previous studies. (a) Heatmap showing the count of previous GWAS where the GWANN hits were identified to be associated with the phenotypes on the x-axis. (b) Heatmap showing the presence of significant evidence for the terms on the x-axis for the GWANN hits. (c) Heatmap showing the overlap between GWANN hits (GWANN), a GWAS run using PLINK 2.0 on the same data as GWANN (TradGWAS), and the largest European AD GWAS (EADB GWAS) [12]. (d) Similar heatmap to (c) but instead of using the GWANN and TradGWAS hits, the top 100 genes from both methods were considered for the overlap with the EADB GWAS hit genes. The sample size of each method is mentioned in the x-axis of the heatmaps, and the diagonals show the number of genes of each method considered while calculating the overlap.

controls and "only AD" versus controls showed identical enrichment in the temporal lobe. In the second study (Patel *et al.*—B) [31], asymptomatic AD versus controls showed no enrichment. The entorhinal cortex and temporal lobes were enriched for symptomatic AD versus controls and symptomatic AD versus asymptomatic AD. GWANN hits in the leading edge were *BIN1*, *SORL1*, *SYNPO*, and *SRGAP2B*.

The same enrichments were also performed for TradGWAS (Supplementary Table S7). The brain regions and conditions enriched were similar to GWANN.

Potential of GWANN targets for Alzheimer's disease drug discovery

To assess the tractability of the GWANN hits for aiding drug discovery, we used TargetDB [32] to score them based on information collected from literature and knowledge about their chemistry, biology, structure, and genetics. *ADAM10*, *APOE*, *SMYD3*, *BIN1*, *SORL1*, and *ROR1* were reported to be tractable, and *SPI1*, *LRRC7*, *APH1B*, *NRG3*, and *PCDH9* were reported as challenging but tractable (Supplementary Table S8). Among the tractable genes, *ROR1* has a drug, cirmtuzumab, associated with it, which is currently under clinical trials for different cancers and neoplasms [33].

Discussion

We developed GWANN and applied it to identify genes associated with family history of AD using data from the UKB. In doing so, we were able to identify 18 genes significantly associated with the phenotype. The *post hoc* enrichment analyses showed enriched biological and disease pathways relevant to AD and neurodegeneration. Several GWANN hits were also identified as tractable drug targets.

Role of hit genes and enriched biological pathways in neurodegeneration and Alzheimer's disease

While some of the GWANN hits have not been identified in previous GWAS, many of them, or their associated biological pathways, have been linked to AD or neurodegeneration. The six well-known AD genes identified by GWANN have not been discussed here. In the following paragraphs, the pathways and gene ontologies

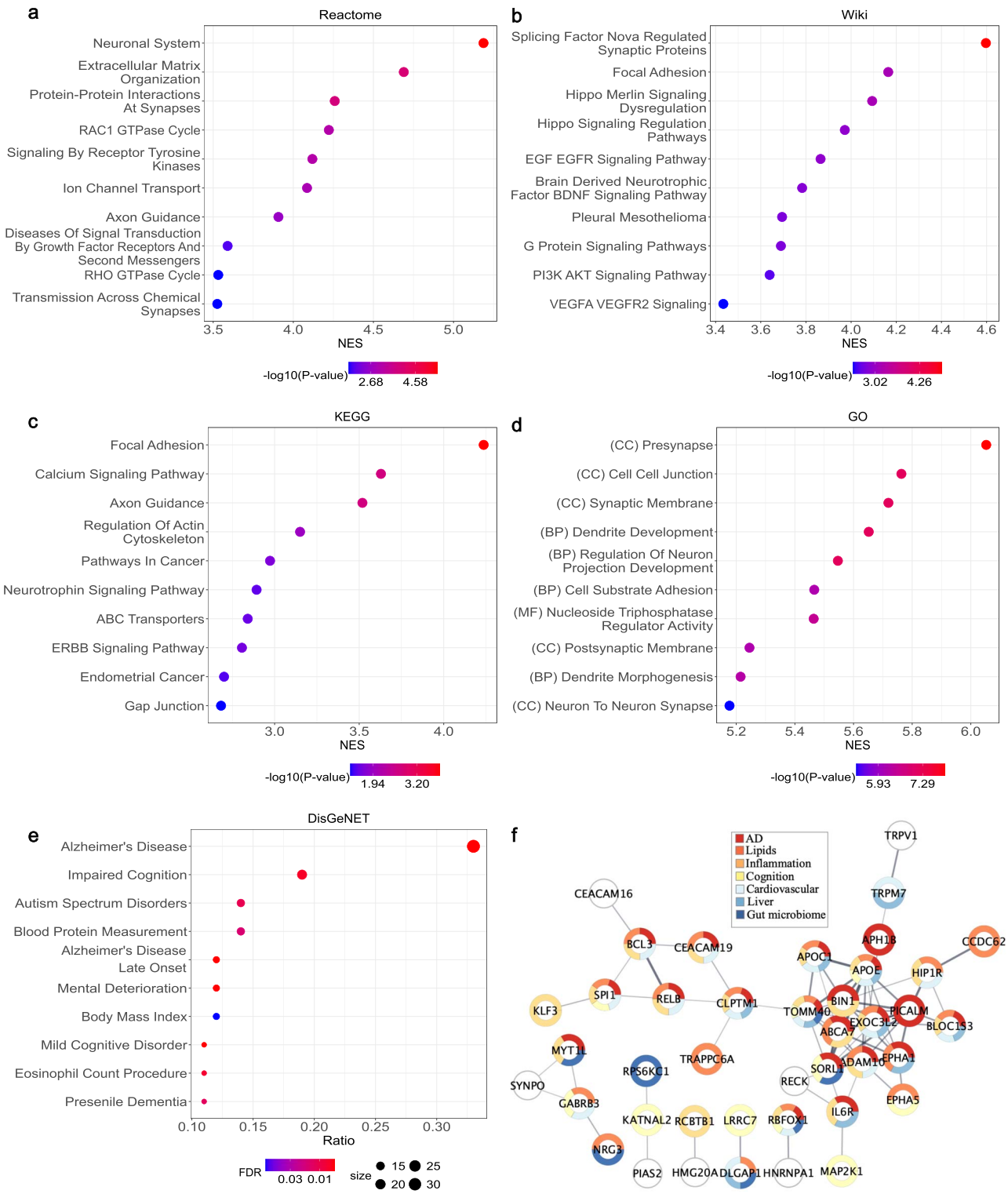

Figure 4. *Post hoc* enrichment analysis after GWANN analysis. (a–d) Gene set enrichment analysis for (a) Reactome, (b) Wiki, (c) KEGG, and (d) GO using GWANN summary metrics. (e, f) Genes were ranked according to the metric $1 - \mathrm{NLL}_{\mathrm{NN}}$, where $\mathrm{NLL}_{\mathrm{NN}}$ was the negative log likelihood of the neural network for a given gene. (e) Disease and trait enrichment using the top 100 genes. (f) Enriched PPI network (P-value $< 1 \times 10^{-16}$) for the top 100 genes. The colors within the nodes highlight the trait categories enriched by the protein encoded by the gene.

mentioned in parentheses were enriched and contain the gene being discussed.

Our *post hoc* analyses link *LINGO2* to synapse organization (GO:0050808) and structure (GO:0050803). Previous findings suggest that *LINGO2* promotes lysosomal degradation of amyloid-β protein precursor, thereby potentially protecting against AD [34, 35]. This could explain its importance in maintaining a healthy synapse by facilitating the clearance of amyloid-β, as identified

Table 2. Enrichment of DEGs identified in two transcriptomic studies on AD brains.

		AD versus Cont	Non-AD versus Cont	Only AD versus Cont
Patel *et al.*—A [30]	Cerebellum	**3.48×10^{-2}** *(PCDH9)*	**3.48×10^{-2}** *(PCDH9)*	**3.48×10^{-2}** *(PCDH9)*
	Frontal	**1.24×10^{-2}** *(SORL1)*	5.65×10^{-2}	**1.81×10^{-2}** *(SORL1)*
	Parietal	**8.75×10^{-7}** *(APOE, SORL1, PCDH9)*	**1.02×10^{-15}** *(APOE, BIN1, LRRC7, SORL1)*	**2.50×10^{-3}** *(PCDH9)*
	Temporal	**1.24×10^{-2}**	9.26×10^{-1}	**1.24×10^{-2}**
		AD versus Cont	**AsymAD versus Cont**	**AD versus AsymAD**
Patel *et al.*—B [31]	Cerebellum	8.49×10^{-1}	4.88×10^{-1}	4.82×10^{-1}
	Entorhinal	**4.20×10^{-3}** *(BIN1, SORL1, SYNPO)*	8.89×10^{-1}	**4.20×10^{-3}** *(BIN1, SRGAP2B, SYNPO)*
	Frontal	8.89×10^{-1}	1.92×10^{-1}	1.50×10^{-1}
	Temporal	**1.88×10^{-2}** *(SYNPO)*	9.07×10^{-1}	**2.56×10^{-2}** *(SYNPO)*

The genes mentioned in parentheses were the GWANN hits in the leading edge of the GSEA for each condition and brain region. Cont - controls, Non-AD - non-AD mental disorders, AsymAD - asymptomatic AD.

by the previously mentioned studies. While not genome-wide significant for AD, *LINGO2* shows nominal significance in GWAS for nonhypertensive AD and brain atrophy [36, 37]. *SYNPO*, another GWANN hit, is involved in synaptic plasticity [38] and autophagic clearance of p-Ser262 *MAPT* [39]. Our *post hoc* enrichment analyses highlighted its localization to the actin cytoskeleton (GO:0015629). The role of the actin cytoskeleton in facilitating autophagy [40] and the involvement of *SYNPO* with the organization (GO:0007015) and binding (GO:0003779) of this cellular component could explain how it aids in the clearance of phosphorylated tau. It was also identified to be downregulated in patients with dementia of Lewy bodies and Parkinson's disease dementia, suggesting its role in other neurodegenerative disorders with similar pathology to AD [41]. *ROR1*, a new GWANN hit, contributes to synaptic health through cytoskeletal involvement [42]. It encodes a receptor tyrosine kinase (GO:0004713) associated with the actin cytoskeleton (GO:0015629). Actin filaments help in maintaining the integrity of the neuronal cytoskeleton, and *ROR1* overexpression has been shown to prevent the degradation *in vitro*, even in the presence of amyloid□ by preserving the actin network [42]. Epigenetically, *ROR1* shows differential hydroxymethylation between late-onset AD patients and controls, correlating with MMSE and MoCA cognitive scores [43].

NRG3, along with known AD hits (*APOE, BIN1, APH1B, ADAM10*), is involved in receptor tyrosine kinase signaling (R-HSA-1250342, R-HSA-1963642) and synapse organization/signaling (GO:0099177, GO:0050808). A single-cell RNAseq study identified the *NRG3-ERBB4* ligand–receptor pair as crucial for intercellular communication in AD brains, with their ablation reducing excitatory synapse formation [44]. *ERBB4* showed nominal significance in our analysis (P-value = 2.59×10^{-10}). A hypothesis-driven study found *NRG3* SNPs and haplotypes significantly associated with AD risk and age of onset [45]. Hence, despite not being genome-wide significant in previous studies, the *NRG3* gene and associated biological pathways have been shown to possess a link to AD.

PCDH9 facilitates neural cell adhesion (GO:0007156, GO:0098742), contributes to forebrain development (GO:0030900), and is associated with the distal axon (GO:0150034). Its variants have been nominally associated with AD [46] and essential tremor [47] in previous GWAS. *SMYD3* has been shown to be significantly elevated in AD patients' prefrontal cortex and tauopathy mouse models, with its inhibition helping rescue cognitive defects and restore synaptic function in pyramidal neurons [48].

Selection of the significance threshold

We ran the method 16 times to obtain a stable metric, defining hit stability as the percentage of intersection between paired runs. The empirical P-value threshold was selected as the largest value ensuring 95% stability for eight runs, which would guarantee at least 95% stability for 16 runs (Supplementary Information, Supplementary Fig. S2b). This approach reduced false positives but increased false negatives. Raising the threshold to 1×10^{-15} would have included well-known AD genes like *PICALM*, *EPHA1*, and *ABCA7*. Using the Bonferroni threshold of 7.06×10^{-7} would have added *ACE*, *CD2AP*, *IL34*, and *APP*. However, these less stringent thresholds would have also included many genes (53% and 65%, respectively) without previous GWAS evidence for AD association. While some might genuinely relate to a family history of AD, others would increase false positives. We therefore chose the conservative empirical threshold of 1×10^{-25} to limit false positives and report the most confident hits.

Comparison of methods and datasets

We studied the overlap of hit genes and top 100 genes between GWANN, TradGWAS, and the EADB GWAS (Fig. 3c and d). There was a larger overlap between the hits of (i) GWANN and EADB GWAS ($n = 5$) as compared to the overlap between the hits of (ii) TradGWAS and EADB GWAS ($n = 2$). However, for the top 100 genes, the overlap was the same ($n = 7$) in (i) and (ii). A possible reason for the smaller overlap between the hits in (ii), despite employing similar methods, can potentially be attributed to the lower power in TradGWAS. Additionally, the overlap between the top 100 genes for (iii) TradGWAS and GWANN ($n = 21$) was larger than (i) or (ii). This would suggest that there seems to be a greater effect of dataset similarity as compared to the method. Although the EADB GWAS included the signal from the UKB, the inclusion of additional datasets made the signal sufficiently different as compared to the data analyzed by TradGWAS and GWANN. The effective dataset used in the EADB GWAS ($n = 788\,989$) was almost 10 times the size of the GWANN or TradGWAS data ($n = 84\,220$), which also contributed to the power of the analysis. We also tested multiple

approaches to calculate the overlap between the different methods and observed the same pattern (Supplementary Fig. S3).

Alongside the traditional GWAS methods that are directly comparable to GWANN, it is worth discussing alternative but related applications of NNs and machine learning methods. Graph NNs have been applied to known gene–disease associations along with additional sources of information including known disease mechanisms, gene connectivity, and gene annotations to prioritize the known disease-associated genes [49]. Another study used gradient boosting to build disease prediction models using clinical biomarkers, plasma protein levels, and other quantitative traits in the UKB [50]. They used their model's prediction probabilities to generate augmented case–control cohorts based on different probability thresholds and perform a phenome-wide association study on these augmented cohorts, resulting in the identification of putative novel genes associated with different diseases. These methods demonstrate alternative approaches to identifying new gene–disease associations in existing cohorts by augmenting the findings of traditional linear methods utilizing data modalities other than genetics.

Limitations and considerations

GWANN did not identify some well-known AD hits like *CLU*, *CR1*, and *TREM2*, even with increased *P*-value thresholds. This could be due to several factors. Firstly, we limited the analysis to SNPs with MAF > 0.01. Excluding lower frequency variants could be a possible explanation for missing genes with rarer AD-associated variants, like *TREM2*. Secondly, to address the computational burden, the analysis was limited to a narrow genomic region, considering only SNPs within genes and 2500 bp flanking regions. This exclusion of most intergenic SNPs led to a lot of potential AD-associated SNPs not being considered in the analysis. Additionally, we also limited the number of SNPs per NN to a maximum of 50 to avoid having multiple NN architectures to accommodate the wide range of SNPs between genes (1–10 000 SNPs). However, it would be more beneficial to include a much larger range of SNPs to utilize the true potential of NNs in identifying nonlinear relationships. Thirdly, the use of a 1:1 case-to-control ratio to prevent overfitting to controls (majority class) resulted in a reduced sample size, potentially weakening signals from some genes. These differences in SNP selection, sample size, and methodology compared to traditional GWAS likely contributed to GWANN's inability to identify known AD loci.

GWANN's inability to provide SNP-level statistics for hit genes limits its comparability with standard GWAS methods and compatibility with post-GWAS analysis tools. While packages like SHAP [51] and Captum [52] offer methods to assign importance to neural network input features, implementing these across multiple runs of GWANN proved complex. The NLL of neural networks for each gene serves as an alternative to traditional effect size estimates, with smaller NLLs suggesting stronger associations. However, the NLL does not indicate the effect direction. Additionally, since we had to run the method more than once, it contributed to increasing the computational cost. Hence, effort is required to make the method scalable and efficient.

Finally, we acknowledge that while the analyzed cohort had diagnosed AD cases, the majority were those with a family history. Family history has been previously used as a proxy for AD [20, 21], but the findings warrant validation in external cohorts with diagnosed AD cases. We attempted to validate the findings in the Asian and Black populations of the UKB, but the validation was underpowered due to very small sample sizes. While it does not serve as a substitute for external validation, in the absence of it, the series of *post hoc* enrichment analyses serve as an additional source of confidence for our findings.

Conclusion

We applied our method to family history of AD using data from the UKB and introduced a new method to complement the success of existing GWAS methods. GWANN identified genes associated with family history of AD that have previously not been identified by GWAS. A series of *post hoc* enrichment analyses provided evidence for differential expression of RNA and proteins associated with the hits between the brains of AD patients and healthy controls. Among the new hits, *LINGO2*, *NRG3*, *PALD1*, *PCDH9*, *SMYD3*, and *SYNPO* have evidence of association with AD or other neurodegenerative disorders from previous *in vitro* and *in vivo* studies. Additionally, enrichment of biological pathways and gene ontologies provided possible explanations for the role of these genes in the processes contributing to AD. Furthermore, *SMYD3*, *LRRC7*, *NRG3*, *PCDH9*, and *ROR1* were identified as tractable targets for drug development. Overall, the findings suggest the potential of GWANN to augment the effort of existing methods in understanding the pathogenesis of AD and other diseases.

Key Points

- We present a new method to perform gene-level disease association using neural networks, called genome-wide association neural networks (GWANN).
- GWANN was applied to family history of Alzheimer's disease (AD) in the UK Biobank and identified 18 genes significantly associated with family history of AD—six well-known AD-associated genes and 12 potentially novel hits.
- *Post hoc* enrichment analyses of GWANN results revealed biologically relevant pathways and processes associated with AD and neurodegeneration.
- Several of the genes identified by GWANN were reported as tractable targets for drug development, suggesting potential avenues for AD therapeutics.
- The findings suggest the possibility of using GWANN as a complementary method to GWAS, to enable the identification of novel signals in existing cohorts.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Acknowledgements

We thank the UK Biobank participants and the UK Biobank team for their work in collecting, processing, and disseminating these data for analysis. This research was conducted using data from the UK Biobank Resource under the approved project 15181. The results published here are in part based on data obtained from Agora (https://agora.adknowledgeportal.org), a platform initially developed by the NIA-funded AMP-AD consortium that shares evidence in support of AD target discovery.

Funding

This work was supported by Alzheimer's Research UK [grant ID ARUK-PhD2022-031]; King Abdulaziz University and the University of Oxford Centre for Artificial Intelligence in Precision Medicine (KO-CAIPM); Janssen Research and Development (Johnson & Johnson); the John Fell Foundation [grant ID 0010659]; and the Virtual Brain Cloud from European Commission [grant number H2020-SC1-DTH-2018-1]. C.A. is funded by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health.

Data availability

The genetic and phenotype data for the analysis were obtained from the UK Biobank resource under the approved project 15181. Data on brain eQTLs, RNA change in the brain, and protein change in the brain were obtained from https://agora.adknowledgeportal.org (site version 3.3.0, data version syn13363290-v66). Differentially expressed genes used in *post hoc* enrichment analysis for (i) AD versus controls, (ii) non-AD versus controls, and (iii) only AD versus controls were obtained from https://doi.org/10.3233/jad-181085. Differentially expressed genes used in *post hoc* enrichment analysis for (i) AD versus controls, (ii) asymptomatic AD versus controls, and (iii) AD versus asymptomatic AD was obtained from https://doi.org/10.1016/j.bbi.2019.05.009. Gene position mapping was obtained from https://vu.data.surfsara.nl/index.php/s/Pj2orwuF2JYyKxq/download. GWAS Catalog Data format v1.0, Data release 2024-01-19 was used. Pathway data were obtained from https://www.gsea-msigdb.org/gsea/msigdb/human/collections.jsp#C5 (v2023.2). Original code used in the analysis can be found at https://github.com/titoghose/GWANN.

Disclosures

W.S. received funding from Johnson & Johnson. A.N.H. received funding from Johnson & Johnson, GlaxoSmithKline, and Ono Pharma. T.Z. received funding from Novo Nordisk. All other authors report no biomedical financial interests or potential conflicts of interest.

References

1. Gauthier S, Rosa-Neto P, Morais JA. *et al.* World alzheimer report 2021: journey through the diagnosis of dementia. 2021.
2. Querfurth HW, LaFerla FM. Alzheimer's disease. *N Engl J Med* 2010;**362**:329–44. https://doi.org/10.1056/NEJMra0909142.
3. Buniello A, MacArthur JAL, Cerezo M. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* 2019;**47**:D1005–12. https://doi.org/10.1093/nar/gky1120.
4. Manolio TA, Collins FS, Cox NJ. *et al.* Finding the missing heritability of complex diseases. *Nature* 2009;**461**:747–53. https://doi.org/10.1038/nature08494.
5. Visscher PM, Brown MA, McCarthy MI. *et al.* Five years of GWAS discovery. *Am J Hum Genet* 2012;**90**:7–24. https://doi.org/10.1016/j.ajhg.2011.11.029.
6. Wu MC, Lee S, Cai T. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 2011;**89**:82–93. https://doi.org/10.1016/j.ajhg.2011.05.029.
7. Ma S, Dalgleish J, Lee J. *et al.* Powerful gene-based testing by integrating long-range chromatin interactions and knockoff genotypes. *Proc Natl Acad Sci* 2021;**118**:e2105191118. https://doi.org/10.1073/pnas.2105191118.
8. de Leeuw CA, Mooij JM, Heskes T. *et al.* MAGMA: Generalized gene-set analysis of GWAS data. *PLoS Comput Biol* 2015;**11**:e1004219. https://doi.org/10.1371/journal.pcbi.1004219.
9. Lambert JC, Ibrahim-Verbaas CA, Harold D. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* 2013;**45**:1452–8. https://doi.org/10.1038/ng.2802.
10. Kunkle BW, Grenier-Boley B, Sims R. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat Genet* 2019;**51**:414–30. https://doi.org/10.1038/s41588-019-0358-2.
11. Wightman DP, Jansen IE, Savage JE. *et al.* A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat Genet* 2021;**53**:1276–82. https://doi.org/10.1038/s41588-021-00921-z.
12. Bellenguez C, Küçükali F, Jansen IE. *et al.* New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat Genet* 2022;**54**:412–36. https://doi.org/10.1038/s41588-022-01024-z.
13. Rusk N. Deep learning. *Nat Methods* 2016;**13**:35–5. https://doi.org/10.1038/nmeth.3707.
14. Zou J, Huss M, Abid A. *et al.* A primer on deep learning in genomics. *Nat Genet* 2019;**51**:12–8. https://doi.org/10.1038/s41588-018-0295-5.
15. Kim S. *et al.* Stacked neural network for predicting polygenic risk score. *Sci Rep* 2024;**14**:11632. https://doi.org/10.1038/s41598-024-62513-1.
16. Hermes S, Cady J, Armentrout S. *et al.* Epistatic features and machine learning improve Alzheimer's disease risk prediction over polygenic risk scores. *J Alzheimers Dis* 2024;**99**:1425–40. https://doi.org/10.3233/JAD-230236.
17. Platt DE, Guzmán-Sáenz A, Bose A. *et al.* AI-enabled evaluation of genome-wide association relevance and polygenic risk score prediction in Alzheimer's disease. *iScience* 2024;**27**:109209. https://doi.org/10.1016/j.isci.2024.109209.
18. van Hilten A, Kushner SA, Kayser M. *et al.* GenNet framework: Interpretable deep learning for predicting phenotypes from genetic data. *Commun Biol* 2021;**4**:1–9. https://doi.org/10.1038/s42003-021-02622-z.
19. Sudlow C, Gallacher J, Allen N. *et al.* UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;**12**:e1001779. https://doi.org/10.1371/journal.pmed.1001779.
20. Marioni RE, Harris SE, Zhang Q. *et al.* GWAS on family history of Alzheimer's disease. *Transl Psychiatry* 2018;**8**:99. https://doi.org/10.1038/s41398-018-0150-6.
21. Jansen IE, Savage JE, Watanabe K. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat Genet* 2019;**51**:404–13. https://doi.org/10.1038/s41588-018-0311-9.
22. Bycroft C, Freeman C, Petkova D. *et al.* The UK biobank resource with deep phenotyping and genomic data. *Nature* 2018;**562**:203–9. https://doi.org/10.1038/s41586-018-0579-z.
23. Huang J, Howie B, McCarthy S. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat Commun* 2015;**6**:8111. https://doi.org/10.1038/ncomms9111.

24. Newby D, Winchester L, Sproviero W. *et al.* The relationship between isolated hypertension with brain volumes in UK biobank. *Brain Behav* 2022;**12**:e2525. https://doi.org/10.1002/brb3.2525.
25. Chang CC, Chow CC, Tellier LC. *et al.* Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* 2015;**4**:s13742-015-0047–8. https://doi.org/10.1186/s13742-015-0047-8.
26. Subramanian A, Tamayo P, Mootha VK. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* 2005;**102**:15545–50. https://doi.org/10.1073/pnas.0506580102.
27. Liberzon A, Subramanian A, Pinchback R. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 2011;**27**:1739–40. https://doi.org/10.1093/bioinformatics/btr260.
28. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res* 2020;**48**:D845–55. https://doi.org/10.1093/nar/gkz1021.
29. Szklarczyk D, Kirsch R, Koutrouli M. *et al.* The STRING database in 2023: Protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* 2023;**51**:D638–46. https://doi.org/10.1093/nar/gkac1000.
30. Patel H, Dobson RJB, Newhouse SJ. A meta-analysis of Alzheimer's disease brain transcriptomic data. *J Alzheimers Dis* 2019;**68**:1635–56. https://doi.org/10.3233/JAD-181085.
31. Patel H, Hodges AK, Curtis C. *et al.* Transcriptomic analysis of probable asymptomatic and symptomatic alzheimer brains. *Brain Behav Immun* 2019;**80**:644–56. https://doi.org/10.1016/j.bbi.2019.05.009.
32. De Cesco S, Davis JB, Brennan PE. TargetDB: A target information aggregation tool and tractability predictor. *PLoS One* 2020;**15**:e0232644. https://doi.org/10.1371/journal.pone.0232644.
33. Ochoa D, Hercules A, Carmona M. *et al.* Open targets platform: Supporting systematic drug-target identification and prioritisation. *Nucleic Acids Res* 2021;**49**:D1302–10. https://doi.org/10.1093/nar/gkaa1027.
34. Miller JA, Woltjer RL, Goodenbour JM. *et al.* Genes and pathways underlying regional and cell type changes in Alzheimer's disease. *Genome Med* 2013;**5**:48. https://doi.org/10.1186/gm452.
35. de Laat R, Meabon JS, Wiley JC. *et al.* LINGO-1 promotes lysosomal degradation of amyloid-β protein precursor. *Pathobiol Aging Age Relat Dis* 2015;**5**:25796. https://doi.org/10.3402/pba.v5.25796.
36. Nazarian A, Arbeev KG, Yashkin AP. *et al.* Genetic heterogeneity of Alzheimer's disease in subjects with and without hypertension. *GeroScience* 2019;**41**:137–54. https://doi.org/10.1007/s11357-019-00071-5.
37. Furney S, Simmons A, Breen G. *et al.* Genome-wide association with MRI atrophy measures as a quantitative trait locus for Alzheimer's disease. *Mol Psychiatry* 2011;**16**:1130–8. https://doi.org/10.1038/mp.2010.123.
38. Deller T, Bas Orth C, Del Turco D. *et al.* A role for synaptopodin and the spine apparatus in hippocampal synaptic plasticity. *Ann Anat Anat Anz Off Organ Anat Ges* 2007;**189**:5–16. https://doi.org/10.1016/j.aanat.2006.06.013.
39. Ji C, Tang M, Zeidler C. *et al.* BAG3 and SYNPO (synaptopodin) facilitate phospho-MAPT/tau degradation via autophagy in neuronal processes. *Autophagy* 2019;**15**:1199–213. https://doi.org/10.1080/15548627.2019.1580096.
40. Kast DJ, Dominguez R. The Cytoskeleton–Autophagy Connection. *Curr Biol* 2017; **27**:R318–26, DOI: https://doi.org/10.1016/j.cub.2017.02.061
41. Datta A, Chai YL, Tan JM. *et al.* An iTRAQ-based proteomic analysis reveals dysregulation of neocortical synaptopodin in Lewy body dementias. *Mol Brain* 2017;**10**:36. https://doi.org/10.1186/s13041-017-0316-9.
42. Chanda K, Jana NR, Mukhopadhyay D. Receptor tyrosine kinase ROR1 ameliorates Aβ1–42 induced cytoskeletal instability and is regulated by the miR146a-NEAT1 nexus in Alzheimer's disease. *Sci Rep* 2021;**11**:19254. https://doi.org/10.1038/s41598-021-98882-0.
43. Chen L, Shen Q, Xu S. *et al.* 5-Hydroxymethylcytosine signatures in circulating cell-free DNA as diagnostic biomarkers for late-onset alzheimer's disease. *J Alzheimers Dis* 2022;**85**:573–85. https://doi.org/10.3233/JAD-215217.
44. Dang Y, He Q, Yang S. *et al.* FTH1- and SAT1-induced astrocytic ferroptosis is involved in Alzheimer's disease: Evidence from single-cell transcriptomic analysis. *Pharmaceuticals* 2022;**15**:1177. https://doi.org/10.3390/ph15101177.
45. Wang K-S, Xu N, Wang L. *et al.* NRG3 gene is associated with the risk and age at onset of Alzheimer disease. *J Neural Transm* 2014;**121**:183–92. https://doi.org/10.1007/s00702-013-1091-0.
46. Sherva R, Gross A, Mukherjee S. *et al.* Genome-wide association study of rate of cognitive decline in Alzheimer's disease patients identifies novel genes and pathways. *Alzheimers Dement (N Y)* 2020;**16**:1134–45. https://doi.org/10.1002/alz.12106.
47. Clark LN, Gao Y, Wang GT. *et al.* Whole genome sequencing identifies candidate genes for familial essential tremor and reveals biological pathways implicated in essential tremor aetiology. *EBioMedicine* 2022;**85**:104290. https://doi.org/10.1016/j.ebiom.2022.104290.
48. Williams JB, Cao Q, Wang W. *et al.* Inhibition of histone methyltransferase Smyd3 rescues NMDAR and cognitive deficits in a tauopathy mouse model. *Nat Commun* 2023;**14**:91. https://doi.org/10.1038/s41467-022-35749-6.
49. Middleton L, Melas I, Vasavda C. *et al.* Phenome-wide identification of therapeutic genetic targets, leveraging knowledge graphs, graph neural networks, and UK biobank data. *Sci Adv* 2024;**10**:eadj1424. https://doi.org/10.1126/sciadv.adj1424.
50. Garg M, Karpinski M, Matelska D. *et al.* Disease prediction with multi-omics and biomarkers empowers case–control genetic discoveries in the UK biobank. *Nat Genet* 2024;**56**:1821–31. https://doi.org/10.1038/s41588-024-01898-1.
51. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems* 2017;4768–77.
52. Kokhlikyan N, Miglani V. Martin M. *et al.* Captum: A unified and generic model interpretability library for PyTorch, 2020.

Briefings in Bioinformatics, 2025, **26(1)**, bbae704
https://doi.org/10.1093/bib/bbae704
Problem Solving Protocol