Biologically Enhanced Machine Learning Model to uncover Novel Gene-Drug Targets for Alzheimer's Disease

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

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

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

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

Given the complexity and multifactorial nature of Alzheimer's disease, investigating potential drug-gene targets is imperative for developing effective therapies and advancing our understanding of the underlying mechanisms driving the disease. We present an explainable ML model that integrates the role and impact of gene interactions to drive the genomic variant feature selection. The model leverages both the Alzheimer's knowledge base and the Drug-Gene interaction database (DGIdb) to identify a list of biologically plausible novel gene-drug targets for further investigation. Model validation is performed on an ethnically diverse study sample obtained from the Alzheimer's Disease Sequencing Project (ADSP), a multi-ancestry multi-cohort genomic study. To mitigate population stratification and spurious associations from ML analysis, we implemented novel data curation methods. The study outcomes include a set of possible gene targets for further functional follow-up and drug repurposing.

Keywords: genomics; Alzheimer's disease; feature importance; informatics; epistasis.

1. Introduction

Alzheimer's disease is the most common cause of dementia, and its prevalence is rapidly increasing due to extended lifespans worldwide.[1] With this surge, there is an urgent need to identify therapeutic targets, potential biomarkers, and risk predictive strategies.[2] Lack of success in recent clinical trials confirmed that AD pathology is very complex and a greater understanding of the underlying mechanisms that contribute to aging and neurodegenerative processes is critical.[3] AD is considered to have a large genetic component and is highly heritable.[4] The polygenic nature of AD presents an obstacle to early diagnosis and risk prediction.[2]

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

Research on AD is a national priority, with 6.5 million Americans affected at an annual cost of more than $250 billion and no definitive cure available.[5–7] This places a significant priority on discovery and approval of therapeutics treatment for AD.[7–9] Drug repurposing involves finding new therapeutic uses for existing drugs that are already on the market.[10] This can lead to significant savings in both time and cost compared to developing new drugs from scratch. Since the safety profiles of these drugs are already well-established, the process can bypass many early-stage trials, speeding up the timeline for reaching patients in need.[10] The Alzheimer's knowledge base (AlzKB) has been developed as a computational AD resource with a particular focus on drug discovery and drug repurposing.[7] It integrates data from 22 diverse sources that spans genes, pathways, drugs, and diseases related to AD to form a specialized open source graph-based knowledge base to aid discovery of complex translational associations for AD drug discovery. The nodes denotes entities (such as genes, pathways, drugs, and diseases) while the edges represent semantic relationships between nodes (entities) such as "*chemical_binds_gene*", "*gene_interacts_with_gene*", "*gene_regulates_gene*", etc. This work leverages the AlzKB's information on gene-gene interaction with known AD genes.

Understanding the role and impact of gene interactions on disease phenotypes is increasingly recognised as an essential aspect of genetic disease research.[11] Most disease-gene association methods do not account for gene-gene interactions, despite their crucial role in complex, polygenic diseases like AD.[2] Exploring the action, function, regulation, and control of proteins can elucidate a clearer understanding of disease processes, cellular functions, and regulatory networks.[12] This is critical in advancing towards precision medicine, given the necessity of anchoring therapeutic targets to a disease mechanism substantiated by genetic evidence.[13] Many of the key functions and life processes in biology are maintained to some extent by different types of protein-protein interactions (PPIs). Knowledge graphs, such as AlzKB, provide a rich heterogeneous network structure that leverages biological and molecular prior knowledge, to uncover possible novel gene-gene interactions that could aid the drug repurposing quest for AD. Drug-gene knowledge sources, such as Drug-Gene interaction database (DGIdb),[14] also provides a rich resource of known interactions between drugs and genes aggregated from multiple sources. This offers additional insights into the molecular mechanisms of drug actions and gene functions, aiding in understanding the underlying biology of diseases and outlining clinically relevant genes.

Machine learning (ML) models in combination with genome-wide association studies (GWAS) have shown promise for identifying novel genes that confer AD risk.[1] To this date, AD GWAS across multiple populations have identified more than 80 loci, with the majority studies conducted in European ancestry cohorts primarily due to large sample sizes.[4] The best known genetic risk factor is the inheritance of the $\epsilon 4$ allele of the apolipoprotein E ($APOE$) gene.[15] Other AD candidate genes have also been identified such as amyloid precursor protein (APP), microtubule-associated protein tau ($MAPT$).[2,15,16] Though ML models have the potential to exploit complex genetic interactions and provide insights into AD pathology, the heterogeneous landscape of AD etiology presents a key challenge.[15] Given the complex biomedical phenotypes that often characterize human diseases, it is becoming increasingly more accepted that epistatic interactions between genes could be more prevalent than previ-

ously assumed.[17,18] Epistatic interactions can be defined as interactions between two or more gene loci where the phenotype cannot be accurately predicted by simply adding the effects of individual gene loci.[19] Epistatic interactions have been detected in multiple GWAS of various disease phenotypes, including AD[20] and other neurological diseases.[21,22] Due to gaps in the current understanding of AD etiology and the complex interactions between genomic and other factors that contribute to its heterogeneity, a multi-modal approach is needed to promote a better mechanistic understanding of the disease.

We present an explainable ML model enhanced by PPI knowledge, specifically epistatic interaction, to identify potential novel non-AD genomic variants with drug targets for further investigation. The underlying hypothesis is that we can leverage the AlzKB and other knowledge sources to pinpoint a set of biologically plausible genes by exploring those with existing drug targets that exhibit a "gene_interacts_with_gene" relationship with known AD genes in the knowledge graph. A key novelty of the ML model is integration of biological knowledge at every level to yield meaningful explanations for model performance and genomic variant (single nucleotide polymorphism (SNP)) feature selection.

2. Methods

We present an ML explainable model, enhanced by the biological knowledge of epistatic interaction, to identify novel genomic variants that could be biomarkers for AD novel gene-drug targets. This framework (see Figure 1) consists of three key phases: (i) druggable gene priority feature selection leveraging AlzKB and DGIdb, (ii) AD study sample data curation, and (iii) ML feature selection and epistasis analysis.

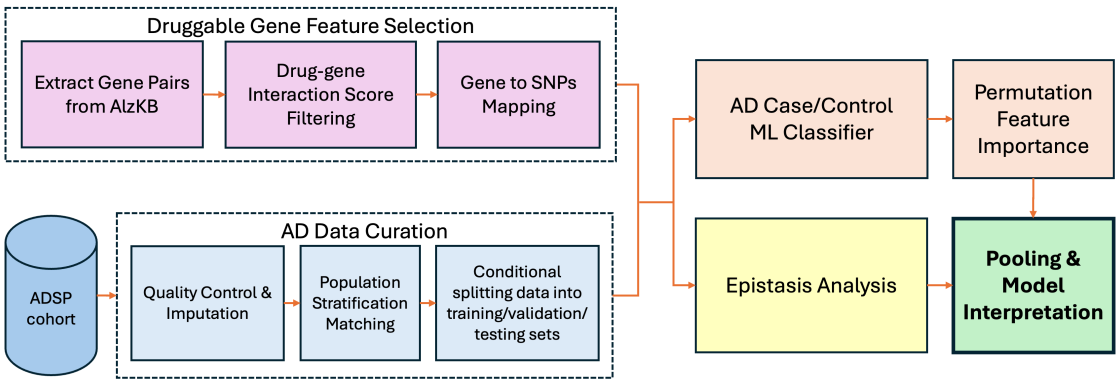


Fig. 1. Flowchart of overall study design.

2.1. *Druggable Gene Feature Selection using AlzKB and DGIdb*

The underlying hypothesis of this study is that nominated gene targets for identification of therapeutic targets for AD can be obtained from the search space of the non-AD genes (i.e. genes not currently known to be implicated for AD) that exhibit a gene-gene interaction with known AD genes in the AlzKB and have at least one drug target. The drug target condition is defined by the edges "*chemical_binds_gene*", "*chemical_upregulates_gene*" and "*chem-*

ical_downregulates_gene". (Note that in this work, AD genes imply all *protein-coding* genes directly linked to the "Alzheimer's Disease" node in AlzKB.) The "*gene_interacts_with_gene*" edges in AlzKB are based on protein-protein interactions.[23] 82 AD genes were connected to 1,805 non-AD protein-coding genes with drug targets, resulting in a total of 2,835 gene pairs. The 82 AD genes served as the baseline model gene list for the subsequent ML model analysis.

2.1.1. *Priority Druggable Gene Selection based on Drug-Gene Interaction Score*

To further prioritize the list of clinically relevant gene selection derived from the AlzKB, we define an additional gene druggability criteria based on the interaction score metric from DGIdb.[14] DGIdb is one of the most comprehensive resources incorporating knowledge about genomic modifications, diseases, and therapeutic targets.[14] The database utilizes experts curation and text-mining of an extensive list of over 40 drug, gene, and interaction sources to extract and rank drug-gene interactions. The interaction score is used to rank the significance and relevance of interactions between drugs and genes. It is calculated based on evidence strength (i.e., the strength of the evidence supporting the interaction from various sources), source credibility, interaction type, the number of supporting sources, and disease relevance. Hence, we retained (AD gene, non-AD gene) pairs from the AlzKB subset, if and only if, the maximum value of the interaction score of the non-AD gene exceeded a 75^{th} percentile threshold (i.e., 11.76 in this work). This yielded a final set of 44 AD genes interacting with 181 non-AD genes, a total of 285 gene pairs (see Figure 2(a) in Results section).

2.1.2. *Gene to SNPs Mapping*

Ensembl REST API is used to obtain the GRCh38 coordinates for the coding regions of each gene.[24] For each gene of interest, we extract all SNPs located within the regulatory regions (100kb upstream and 5kb downstream). The baseline model feature set consists of the union of all the SNPs mapped to each of the 82 AD genes. For the (AD, non-AD) gene-gene interaction datasets, the SNPs feature set is mapped per AD gene, i.e., the SNPs of the AD gene along with all SNPs belonging to each of the non-AD gene interacting with that AD gene.

2.2. *AD Data Sample and Curation*

The AD genotype data utilized in this study is drawn from the Alzheimer's Disease Sequencing Project (ADSP).[25] The ADSP aims to identify genetic variants that influence the risk of AD by sequencing the genomes of individuals (from ethnically diverse populations), focusing primarily on AD case/control phenotypes derived from clinical data. The study sample was extracted from the ADSP R4 v11 2023 release VCF dataset which originally had 346,763,200 variants and 36361 samples.

Pre-filtering quality control was done at two levels: at the variant level, based on sequencing statistics, and at the sample level, based on duplicate samples. To ensure the reliability of genetic analyses and focus on more impactful genetic variants, additional filtering steps were performed to remove singletons and exceedingly rare variants. Singletons imply variants present in only one individual, thus less likely to be relevant to the disease. For exceedingly

rare variants, the total number of counted alleles is very small relative to the number of samples. Thus, subsequent analysis focus on variants with enough occurrences to allow for meaningful statistical analysis. Variants with low call rates (missing call rate > 0.01) and samples with poor genotyping rates (missing call rate > 0.05) were also excluded. Only common variants (minor allele frequency (MAF) >1%) were retained resulting in a final variants count of 9,520,653 and 34971 samples. Imputation of missing values was done using mode-based imputation to avoid false positive signals as a small set of 400 variants had almost no homozygous calls.

2.2.1. *Population Stratification using Propensity Score Matching*

The ADSP R4 v11 2023 release spans 40 study cohorts made up of 5,218 subjects of African ancestry, 2,791 of Asian ancestry, 10,398 of Hispanic ancestry, and 16,191 Non-Hispanic White. Thus, another key consideration of the genomic data preprocessing is to insure that any bias due to population stratification is mitigated before quantitative analysis. Population stratification (PS) refers to the presence of systematic differences in allele frequencies between subpopulations in a population due to different ancestries. These differences can confound genetic association studies if not properly accounted for, leading to false associations or masking true associations between genetic variants and diseases.[26] A commonly used method to address PS is principal components analysis (PCA). This approach uses genotype data (independent loci) to compute the principal components, which are assumed to represent features of genomic ancestry that capture PS. The principal components are then used as covariates in subsequent analyses. However for complex ML analysis, usage of covariates is not applicable. To control for PS in this study, we developed a novel method that adjusts the dataset for ancestral heterogeneity by performing propensity score matching (PSM) on genomic PCA.

To obtain the PCA of the independent genomic loci, we extract a subset of the data based on these parameters: MAF > 0.02, Hardy-Weinberg Equilibrium (HWE) exact test p-value $>$ 1e-7, Linkage Disequilibrium (LD) with a variant window count of 100, a step size 10, and R2 cutoff of 0.1. Subsequently, we apply the PSM procedure using the top eight principal components derived from the PCA computation. The PSM conducts a logistic regression on the 8 PCA covariates to compute the propensity score. The matching is performed using *psmpy* package.[27] A key novelty of the matching process is that it ensures that the individual from the control subset has its closest counterpart in the disease subset based on the computed propensity score using k nearest neighbors matching. The final matched dataset (see Table 1) obtained had 22560 samples equally distributed between AD case and control phenotypes.

Table 1. Demographic summary of cases and controls in the final matched dataset

	Female (%)	Harmonized Age*	Race (%)						Ethnicity (%)		
			White	AA	Asian	Native/Amer. Ind.	Other	N/A	Hispanic/Latino	Non-Hispanic	N/A
Cases	60.48	33 to 90+	66.55	13.95	1.61	0.41	14.73	2.75	26.45	69.34	4.21
Controls	66.33	30 to 90+	47.81	25.61	1.25	0.33	16.11	8.89	36.54	61.99	1.46

*__Harmonized Age__: age at onset for cases, and age at last exam for controls (Age values of 90 or more are coded as "90+").
__Race__: uses NIH Racial Categories. AA denotes Black or African American, Native/Amer. Ind. denotes Native Americans and American Indian/Alaska Native.

2.2.2. *Conditional Splits of AD Data for Robust ML Analysis*

The last phase of the AD data curation involved an intentional split of the derived matched data so that the key fairness characteristics (mitigating population stratification) are not lost during the ML phase. Building a robust ML classifier model requires training and validation datasets as well as a test hold out set, that is not seen by the model during the training and validation phase, to ensure model generalization.[28] The two conditions that had to be preserved and consistent across the splits into three datasets were: (i) Matched case/control pairs and propensity score distributions, (ii) Distributions of significant SNPs reported by recent GWAS studies. The set of significant SNPs is based on the 2023 Lancet meta review[4] studies that listed 101 unique SNPs with a significance threshold p-value < 5e-8. The variant filtering for the matched dataset was based on these parameters: MAF > 0.1, HWE exact test p-value > 1e-7, LD with a variant window count of 100, a step size 10, and R2 cutoff of 0.8. Note that after the filtering phase, only 30 of the 101 SNPs were present in the matched data.

We designed an optimization algorithm using the Optuna platform[29] that satisfied the specified conditions for dataset splits. This entailed running 1000 Optuna trials by sampling different random seeds for training, validation and testing sets splits in equal ratio 1/3:1/3:1/3 for the matched case/control sample pairs. During each trial, 2 objectives with equal weights were evaluated: (i) maximizing the median –log(p) of 30 SNP set across splits; (ii) minimizing the absolute difference between the median –log(p) of 30 SNP. At the end of the optimization procedure, the best trial datasets were selected as the training/validation and test sets for subsequent analyses.

2.3. *ML feature selection and Epistasis Analysis*

2.3.1. *ML AD Case/Control Classifier and Feature Importance*

To identify genomic biomarkers that may indicate potential gene-drug targets, we assessed their predictive power in constructing a ML model for AD case vs control classification using the ADSP matched data. We performed 44 experiments for the gene-gene interaction sets using its corresponding SNPs PPI input feature set. Let $G_{AD} = g_{AD_1}, g_{AD_2}, ..., g_{AD_l}$ be the set of AD genes, where l=44. $G_{nonAD} = g_{AD_1}, g_{AD_2}, ..., g_{AD_m}$ be the set of non-AD genes, where m=181. $I \subseteq G_{AD} \times G_{nonAD}$ denotes the set of interacting (AD, non-AD) gene pairs; $I = \{(g_{\mathrm{AD}_i}, g_{\mathrm{nonAD}_j}) \mid g_{\mathrm{AD}_i} \in G_{\mathrm{AD}}, g_{\mathrm{nonAD}_j} \in G_{\mathrm{nonAD}}\}$, where $|I| = 285$. Let SNP(g) be the set of SNPs for gene g. For each AD gene $g_{AD_i} \in G_{AD}$, its SNPs PPI feature set is the union of g_{AD_i} and the SNPs of all non-AD genes g_{nonAD_j} that interact with g_{AD_i}, $\mathrm{Input\_SNPs}(g_{\mathrm{AD}_i}) = \mathrm{SNP}(g_{\mathrm{AD}_i}) \cup \bigcup_{(g_{\mathrm{AD}_i}, g_{\mathrm{nonAD}_j}) \in A} \mathrm{SNP}(g_{\mathrm{nonAD}_j})$. This is the input data for each of the 44 AD gene PPI experiments. The baseline performance was determined by building the AD case/control with the SNPs derived from all the 82 AD genes, G^{+}_{AD} (see Section 2.1). $\mathrm{Baseline\_SNPs} = \bigcup_{g_{AD} \in G^{+}_{\mathrm{AD}}} \mathrm{SNP}(g_{AD})$.

The AD case/control classifier model was implemented using both automated ML with tree-based pipeline optimizer 2 (TPOT2[30]) platform and the Extreme Gradient Boosting (XGBoost) algorithm.[31] TPOT2 allows for the selections of the best-performing ML model for a given problem in an agnostic manner. The classification pipelines are generated from the sub-

set of ML methods and data pre-processing operators imported from the scikit-learn Python library. During the optimization process, various combinations of pre-processing operator are combined with ML methods into a pipeline in a tree-based manner. XGBoost is a tree-based model implementation of the gradient boosting framework, which combines the predictions of multiple weak learners (usually decision trees) to produce a strong overall model. For fair comparison of both models, the hyperparameters tuning for Xgboost was performed using Optuna method[29] with the objective function set to maximize Receiver Operating Characteristic - Area Under the Curve (ROC AUC) metric across hyperparameters search space of the TPOT2 configuration. For both methods, the training and validation datasets were used for tuning and optimization, and once the final model and hyperparameter set was determined, the final performance was evaluated using testing set.

To identify which variants were driving the predictive power of each model, we performed permutation feature importance (PFI) to compute the univariate contribution of each variant (feature). Note that the PFI coefficient value, which estimates the main effect of each SNP, was calculated exclusively using the testing dataset.

2.3.2. *Epistasis Analysis*

The aim of the epistasis analysis was to compute the level of strength of interactions between SNPs that contribute to the disease, rather than individual SNPs or the additive effects of SNP subsets. Exhaustive searches of epistatic interactions are computationally expensive due to the high dimensionality of genomic datasets, We computed the epistatic interaction using BitEpi, a parallelized bitwise algorithm, which allowed for fast, exhaustive computation of higher-order interactions between SNPs.[32] The genotypes are encoded in bytes (8-bits) with the first 2 bits denoting the combination (e.g. $0/0 \rightarrow 00$, $0/1 \rightarrow 01$) and the remainder bits set as 0. Bitwise operations are subsequently applied to combine genotypes of up to 4 SNPs to create contingency tables and compute the entropy-based metrics (association power (β) and interaction effect size (α)). The β metric reflects the combined association power of the SNPs considered, while the α metric indicates the gain in association power due to the epistatic effect of those SNPs. The α (also known as information gain) metric quantifies the level of strength of interaction of the SNP sets. For all $SNPs \in I$ gene pairs, we computed α for each individual SNP (18778 variants) and its two way interactions (176,297,253 SNP pairs) using the matched testing dataset.

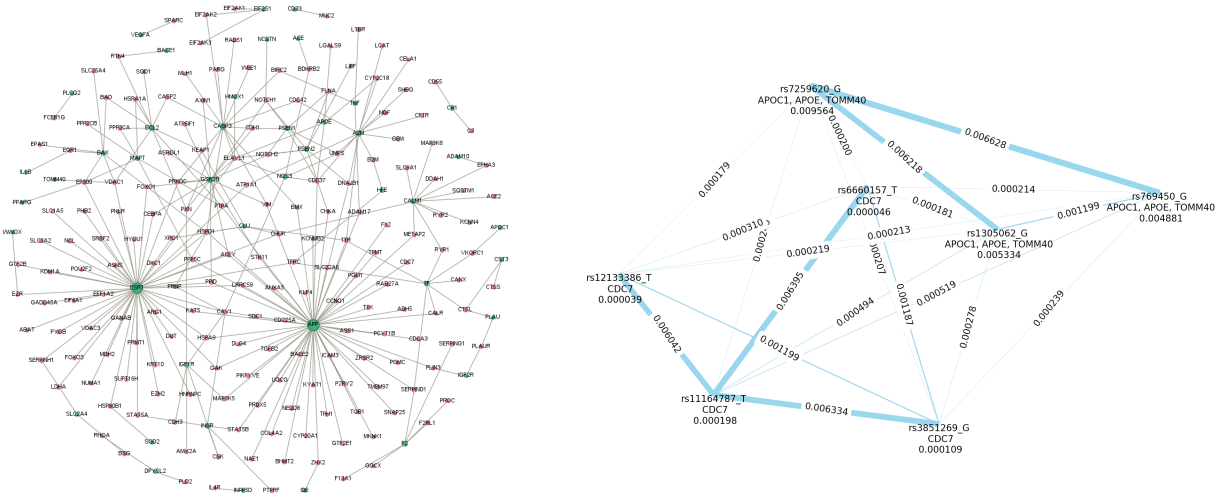
2.3.3. *Pooling & Model Interpretation*

The final phase of the learning framework pools the results obtained from both the epistatic interaction analysis and the ML feature selection to determine a final set of potential genomic biomarkers for AD novel gene-drug targets. Though the ML classification models are able to assess the predictive power of a set of SNP variants to distinguish AD case from control phenotype, it may fall short in the explainability phase of the key drivers. Feature importance scores, as quantified by PFI coefficient scores, are based on main effects of each feature (SNP in this case). The goal is to extract the list of top ranking SNPs exhibiting relatively strong level of interactions and assess their predictive power in distinguishing AD case from control

phenotypes. This will provide additional evidence of their effect on the ML classification models for each AD gene and its set of interacting non-AD gene pairs.

3. Results and Analysis

Figure 2(a) illustrates the outcome of druggable gene feature selection phase which yielded 285 $(g_{\mathrm{AD}_i}, g_{\mathrm{nonAD}_j})$ pairs.



(a) Network of 285 (AD, non-AD) gene pairs (*gene_interacts_with_gene*) filtered by 75^{th} percentile drug interaction score from DGIdb. Green dots denote the 44 AD genes while red, the 181 non-AD genes.

(b) Visualization of network of top 5 epistatic interactions, as quantified by information gain using α values for the main effect and 2-way interactions from BitEPI in ADSP dataset.

Fig. 2. Visualization of (a) 285 (AD, non-AD) gene pairs, (b) SNPs that exhibit high epistatic interactions.

Table 2 illustrates the performance of all the AD case/control classification experiments for both XGBoost and TPOT2 models. Out of the 44 experiments conducted for each of the AD genes (and its set of interacting non-AD genes), only the top 15 best performing experiments (based on XGBoost ROC AUC) are listed in Table 2. See Table S1 in Supplementary file [a] for complete details of all experimental results. From Table 2, we observe that the baseline model (union of all SNPs from 82 AD genes) outperformed all the other models, as expected, with 64.23% for TPOT2 model and 63.65% for XGBoost. While the TPOT2 models seemed to have the better ROC AUC performance overall, the pipelines were very complex and hard to interpret. Hence, based on complexity/performance trade-off, we selected the XGBoost models for further evaluation of the individual contribution of each SNP variant to overall model performance. Among the best performing models, we observe that *APOC1*, *APOE*, and *TOMM40*

[a]Supplementary information is available at: `https://github.com/EpistasisLab/PSB25_ADSP_GIG`

Table 2. AD case/control classification (XGBoost vs TPOT2) performance outcomes for baseline (all AD genes) model, top 15 gene-gene interaction sets, and subset of non-AD genes SNPs that exhibit strong epistatic interaction. Gene-gene interaction sets are sorted by ROC AUC scores from the final XGBoost model.

Gene set	**# non-AD genes**	**# SNPs**	**ROC AUC**		**Recall**		**Precision**		**Accuracy**	
			XGB	TPOT2	XGB	TPOT2	XGB	TPOT2	XGB	TPOT2
All AD genes	-	9539	63.65	64.23	57.53	58.38	59.70	60.17	59.35	59.87
APOC1	1	57	63.15	64.13	53.16	53.94	60.96	61.98	59.56	60.43
APOE	4	389	63.05	64.10	55.53	52.74	60.50	62.10	59.64	60.28
TOMM40	1	141	62.79	64.18	53.40	52.15	60.30	61.98	59.12	60.08
ESR1	53	3399	60.03	52.28	64.63	54.10	56.95	50.05	57.89	50.05
GSK3B	15	1199	59.59	46.17	62.95	0.00	56.60	0.00	57.34	50.00
APP	64	5256	59.29	49.15	64.04	54.10	56.95	50.05	57.82	50.05
CASP3	15	884	59.04	59.80	62.31	70.59	56.39	56.15	57.06	57.73
DPYSL2	2	269	59.02	59.19	63.38	64.92	55.98	56.21	56.77	57.17
BCL2	10	558	58.98	58.94	62.71	70.72	56.18	55.07	56.90	56.52
A2M	14	1112	58.67	59.39	59.76	61.99	56.33	56.70	56.72	57.33
BAX	7	589	58.53	59.04	61.46	45.96	55.53	50.00	56.12	50.00
WWOX	1	1320	58.46	58.67	63.54	64.12	55.79	56.31	56.60	57.18
INSR	5	468	58.43	59.35	61.89	61.09	55.82	56.77	56.45	57.29
CALM1	11	1878	58.42	59.82	61.14	71.14	55.57	55.28	56.13	56.80
TF	8	593	58.36	59.82	61.81	62.23	55.69	57.51	56.32	58.13
non-AD genes*	104	1867	60.06	49.94	65.21	53.67	57.05	50.01	58.06	50.01
High α SNPs†	6	56	62.92	63.65	60.90	59.15	59.19	60.01	59.45	59.87

**non-AD genes SNPs set selected based on $\alpha > 0.003$ from BitEpi.*

†*High α SNPs selected based on top 50 α from BitEpi.*

PPI gene sets had relatively high performance, though the number of their corresponding interacting non-AD genes SNPs was very small. This suggests that model performance could be attributed mainly to SNP(g_{AD}). The gene sets for *ESR1*, *GSK3B*, *APP*, and *CASP3* had a larger number of interacting non-AD genes (and SNPs) and performed relatively well (ROC AUC of $59 - 60\%$).

Figure 3 reports the PFI values for each top performing gene sets based on the XGBoost models. For *APOC1*, *APOE* and *TOMM40* gene sets, SNP rs7259620G has the largest main effect followed by rs769450G and rs449647A, with PFI values of 0.084, 0.057, 0.09 respectively. The rs769450G is a known intronic variant associated with AD risk[33] . There are limited studies demonstrating the association of rs7259620G with AD risk.[33,34] However, the rs449647A (*TOMM40* intronic variant) currently has no GWAS, functional or clinical annotation available. For the remaining 12 gene sets, none of the SNPs exhibited an informative contribution of significant value, with all contributions being less than 1%. Though the performance of these models is 58 - 60% ROC AUC which is relatively close in performance to the top models with quantifiable independent effects (*APOC1*, *APOE*, *TOMM40*). This suggests that the driver for the model performance is likely due to the interaction effect of its variants.

Figure 4(a) presents the PFI values for the baseline (SNPs(G^{+}_{AD})) model. The SNP with the highest PFI score, by a large margin, is the same set of three SNPs from the *APOC1*, *APOE* and *TOMM40* gene sets (Fig. 3). This provides additional evidence that the key driver of model performance for those gene sets were most likely due to the already know AD risk

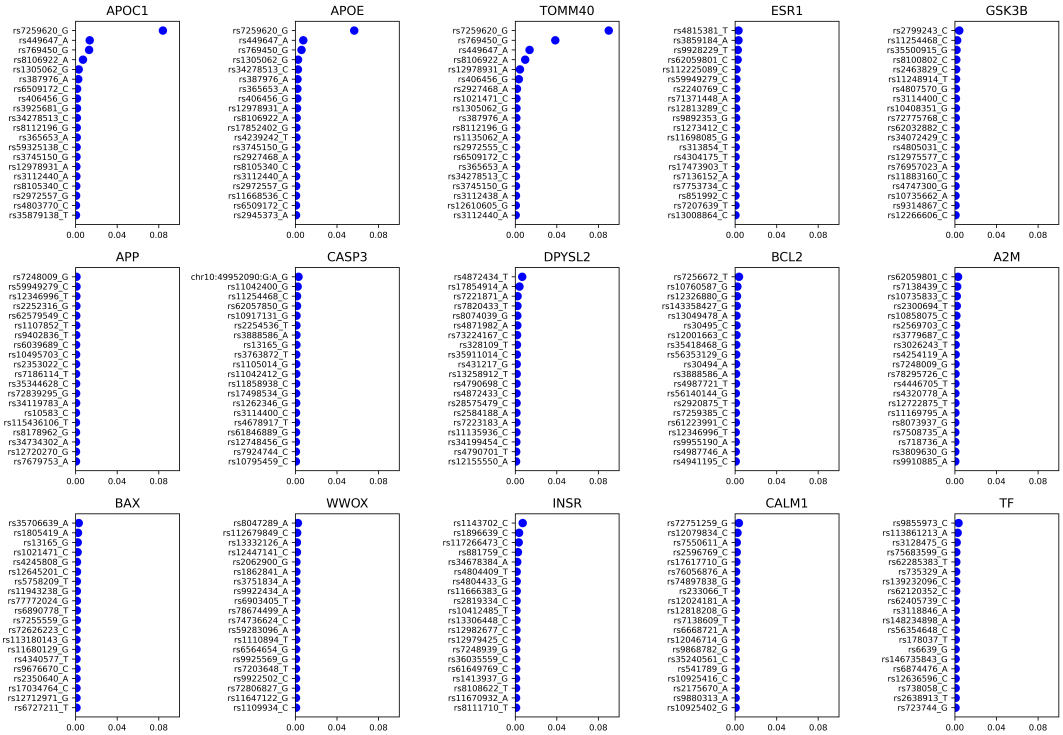
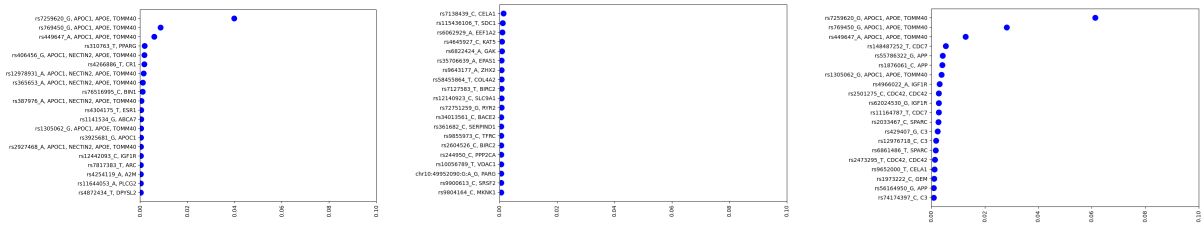


Fig. 3. Permutation feature importance scores for top 15 gene sets from XGBoost model.



(a) Baseline model (all AD) (b) Non-AD SNPs ($\alpha > 0.003$) (c) High α SNPs (Top 50)

Fig. 4. Comparison of XGBoost permutation feature importance scores for selected SNPs sets

genomic variants.

The epistasis interaction analysis outcome is presented in Table 3 for the top 15 SNPs ranked by the information gain of its two-way interaction (α). (see Table S2 in supplementary file for complete list [b]). The effect size of the top two-way SNP combinations, while slightly smaller, was comparable to the top individual SNP effect size indicating that non-additivity could be a contributing factor in explaining AD genomic mechanism. From Table 3, we observe that the strongest pair of interacting SNP variants with α=0.0066 was (rs7259620G, rs769450G). These two SNPs also had largest main effects (α=0.0096, 0.0053) and were two of SNPs driving ML performance for the top 3 gene sets (*APOC1*, *APOE* and

[b]Supplementary information is available at: `https://github.com/EpistasisLab/PSB25_ADSP_GIG`

TOMM40) and baseline model (see Figs. 3, 4(c) and4(a)) (see Table S3 in supplementary file for complete list of PFI values [b]). There were also SNPs from non-AD genes that had strong interaction values: (rs6660157T, rs11164787T $\alpha = 0.0064$), (rs3851269G, rs11164787T $\alpha = 0.0062$), and (rs12133386T, rs11164787T $\alpha = 0.006$). These were all affiliated with the non-coding region of the *CDC7* gene. A visualization of the network of the 7 SNP variants involved in the top 5 interactions (see Figure 2(b)) reveals two groups of interactions: *CDC7* non-coding region, and *APOE* intronic with non-coding *TOMM40* intronic regions. The visualization also reveals slightly weaker interactions (α =0.0012) present for *CDC7* non-coding (rs12133386T, rs3851269G) and (rs3851269G, rs6660157T), and for *TOMM40*/*APOE* (rs1305062G, rs769450G).

Table 3. Top 15 epistatic interaction results from BitEpi

SNP_A	SNP_B	α_{AB}	α_A	α_B	β_{AB}	β_A	β_B	Gene_A	Gene_B
rs7259620_G	rs769450_G	0.0066	0.0096	0.0048	0.5162	0.5096	0.5049	*APOC1*, *APOE*, *TOMM40*	*APOC1*, *APOE*, *TOMM40*
rs6660157_T	rs11164787_T	0.0064	4.60E-05	0.0002	0.5066	0.5000	0.5002	*CDC7*	*CDC7*
rs3851269_G	rs11164787_T	0.0063	0.0001	0.0002	0.5065	0.5001	0.5002	*CDC7*	*CDC7*
rs1305062_G	rs7259620_G	0.0062	0.0053	0.0096	0.5158	0.5053	0.5096	*APOC1*, *APOE*, *TOMM40*	*APOC1*, *APOE*, *TOMM40*
rs12133386_T	rs11164787_T	0.0060	3.90E-05	0.0002	0.5062	0.5000	0.5002	*CDC7*	*CDC7*
rs11166498_G	rs11164787_T	0.0060	6.00E-05	0.0002	0.5062	0.5000	0.5002	*CDC7*	*CDC7*
rs12816187_A	rs9652000_T	0.0060	0.0001	0.0004	0.5064	0.5001	0.5004	CELA1	CELA1
rs1305062_G	rs449647_A	0.0059	0.0053	0.0036	0.5113	0.5053	0.5036	*APOC1*, *APOE*, *TOMM40*	*APOC1*, *APOE*, *TOMM40*
rs2473295_T	rs2501275_C	0.0059	0.0006	0.0002	0.5065	0.5006	0.5002	*CDC42*	*CDC42*
rs2473296_C	rs2501275_C	0.0059	0.0005	0.0002	0.5065	0.5005	0.5002	*CDC42*	*CDC42*
rs7529485_C	rs11164787_T	0.0059	2.70E-05	0.0002	0.5061	0.5000	0.5002	*CDC7*	*CDC7*
rs1883421_C	rs2501275_C	0.0059	0.0005	0.0002	0.5064	0.5005	0.5002	*CDC42*	*CDC42*
rs12116952_G	rs2501275_C	0.0059	0.0006	0.0002	0.5065	0.5006	0.5002	*CDC42*	*CDC42*
rs1063116_A	rs2501275_C	0.0059	0.0005	0.0002	0.5064	0.5005	0.5002	*CDC42*	*CDC42*
rs2501291_G	rs2501275_C	0.0059	0.0006	0.0002	0.5065	0.5006	0.5002	*CDC42*	*CDC42*

Figure 5 illustrates the estimated AD distribution for using contingency table plots for selected 2-way interactions with large α values. These plots display the number of samples for each genotype combinations for the selected SNPs in both case and control cohorts. The plots for the (rs7259620G, rs769450G) pair have substantially increased AD rate when both SNP are homozygous for the alternative allele. More complex associations were observed for (rs6660157T, rs11164787T) pair. Increased AD risk is observed when rs11164787T is homozygous for the reference allele and rs11164787 is homozygous for the alternative allele. When both rs11164787T and rs6660157T are heterozygous, and when rs11164787T is homozygous for the alternative allele and rs11164787 is homozygous for the reference allele. Select genotypes combinations for non-coding region of *CDC42* also associated with increased risk of AD: when both rs760923G and rs760923G are homozygous for the reference allele, when both rs760923G and rs760923G are heterozygous, and when both rs760923G and rs760923G homozygous for the alternative allele.

When the SNPs from non-AD genes of meaningful α values from epistasis analysis are pooled to build the AD case/control classifier, it yields a comparable performance (60.06% ROC AUC) for the XGBoost model (see last row in Table 2). However, the PFI result (Figure 4(b)) which quantifies the univariate contribution of each SNP ($< 2\%$) to model performance still fall shorts in explainability of model performance. This provides additional evidence that the key contributors for model performance is beyond univariate contributions of these SNPs. The non-additive effects of these SNPs could be a factor.

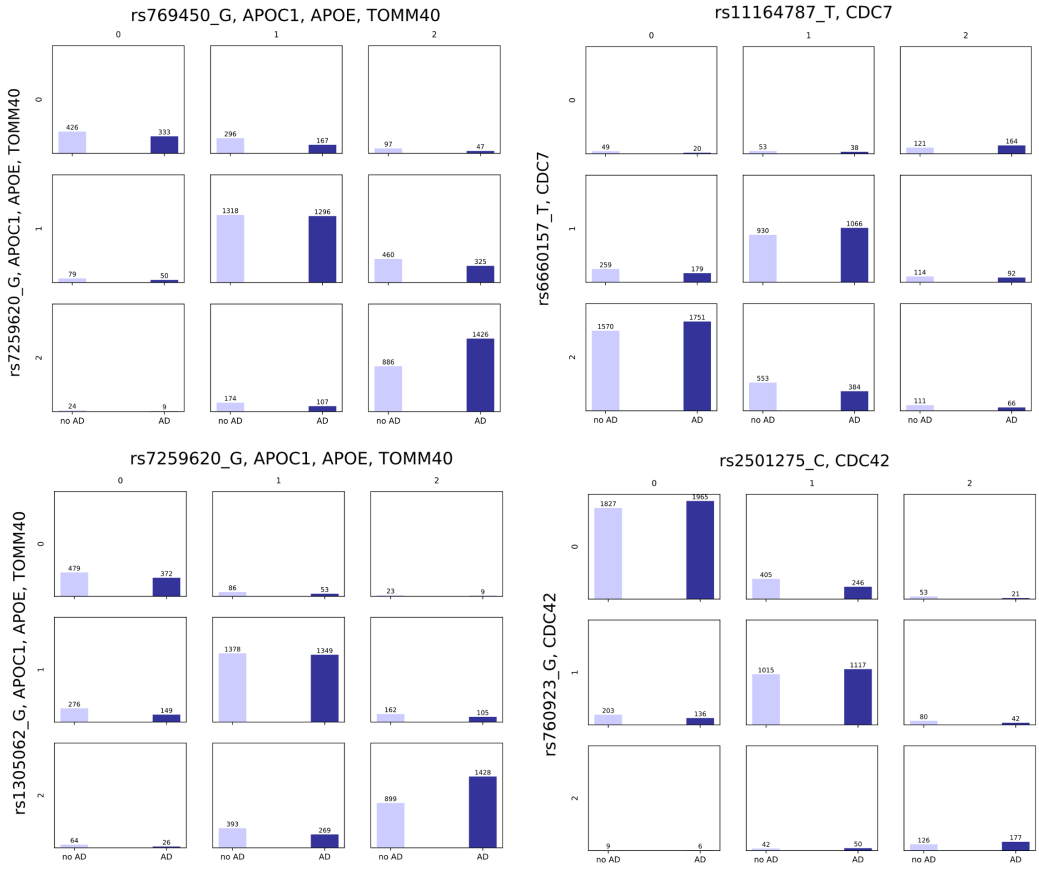


Fig. 5. Genotypes combinations for selected SNP pairs with high information gain (α values).

4. Discussion

In this study, we have implemented an ML SNPs feature selection model that integrates epistatic interaction and leverages the Alzheimer's knowledge base (AlzKB), Drug-Gene interaction database (DGIdb) to identify a list of biologically plausible novel gene-drug targets for further investigation. The prior biological knowledge of gene-gene interactions in AlzKB is based on protein-protein interactions.[23] The model is validated using an ethnically diverse study sample obtained from the Alzheimer's Disease Sequencing Project (ADSP). A primary goal of the ADSP is to further the understanding of the genetic architecture of AD and related dementias and subsequently, turn genetic findings into meaningful therapeutic targets.[25] Given the complexity of the dataset fueled by the multiple ancestry in the sample population, ML analysis directly applied could yield spurious associations that are not related to the disease mechanism but possible ancestry differences. Hence, a key contribution of this work is the extensive novel preprocessing steps applied on the AD case/control ADSP genomic data to mitigate of population stratification. We applied a novel method that combines PCA with propensity score matching. The mitigation of system bias was validated by computing the genomic inflation factor using an external GWS study analysis.

The robustness and generalization of the ML model outcomes is enforced by the conducting conditional splits of the datasets into training, validation and testing sets such that the

matching benefits are not compromised. Conducting the model performance evaluation, feature selection, and epistasis analysis exclusively on the test set, increases the confidence in the generalization and reproduciblity of the results obtained. We utilized two ML methods: TPOT2, an automated ML tool that explores multiple classification algorithms using genetic algorithm and selects the most optimal, and XGBoost, a scalable and highly effective tree-boosting algorithm. Though the TPOT2 performed better overall, subsequent analysis was done based on the XGBoost models, as they were relatively simple and efficient compared to the complexity of the TPOT2 pipelines. Permutation analysis of top models revealed that some SNPs of known AD risk genes are drivers of the performance. Specifically, for the best performing models (all AD genes, *APOE*, *APOC1*, and *TOMM40*) the most informative variant is rs7259620G located 2KB upstream from the *APOE* gene region. GWAS with 17,480 European individuals found an association of the *APOE* rs7259620 G allele with increased AD risk (OR=1.68, p=2x10-23[33]). The second high ranking SNP (rs769450G) is a common intronic variant associated with AD risk. Several large GWAS have also found highly significant associations with various traits including AD.[33] For the rs449647A, a *TOMM40* intronic variant, there was no GWAS, functional or clinical annotation available. Understanding the functional implications of rs449647A could potentially shed light on its contribution to disease risk or progression. However, for other top performing models, the PFI analysis could not quantify any SNP has having substantial univariate contribution to explain model performance.

To identify whether the genotype combinations can better explain phenotype variance in these experiments, we ran an exhaustive pairwise epistasis analysis with BitEpi, a highly scalable and efficient method. Among the combinations of genotypes with strong informative contributions are previously identified SNPs from intronic and noncoding upstream regions of *APOE* and *TOMM40*, and novel SNPs from noncoding regions of *CDC7* (linked to *APP* through gene-gene interactions in AlzKb) and *CDC42* (linked to A2M through gene-gene interactions in AlzKb) (see Table 3, Fig. 5). The non-AD gene SNPs haven't been previously reported in GWAS studies and do not have any functional or clinically relevant affiliations with AD. Epistasis analysis uncovered some novel SNPs, not related to the AD genes, which when pooled into the ML analysis demonstrated comparable predictive power to baseline all AD genes model (see Table 2, Figure 4(b)). This suggests a biological plausible set of genes for further investigation as potential drug-target genes for AD.

This work highlights the limitations of basic ML model interpretation methods, which tend to focus solely on main effects while overlooking impact of epistatic interactions that may contribute to model performance. Evaluating model-based 2- and 3- way PFI is an exhaustive procedure and not scalable for high-dimensional genomic data. We propose that integrating ML analysis with epistasis detection could address this challenge and facilitate advancements in uncovering disease mechanisms and identifying potential therapeutic targets."

Acknowledgements

We wish to thank Yuki Bradford and Rachit Kumar for their help with the data imputation process.

References

1. B. Qorri, M. Tsay, A. Agrawal, R. Au and J. Geraci, Using machine intelligence to uncover alzheimers disease progression heterogeneity, *Exploration of Medicine* **1**, 377 (2020).
2. Y. Lagisetty, T. Bourquard, I. Al-Ramahi, C. G. Mangleburg, S. Mota, S. Soleimani, J. M. Shulman, J. Botas, K. Lee and O. Lichtarge, Identification of risk genes for alzheimer's disease by gene embedding, *Cell genomics* **2** (2022).
3. J. Pleen and R. Townley, Alzheimer's disease clinical trial update 2019–2021, *Journal of Neurology* **269**, 1038 (2022).
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. Grueso and R. Viejo-Sobera, Machine learning methods for predicting progression from mild cognitive impairment to alzheimer's disease dementia: a systematic review, *Alzheimer's research & therapy* **13**, 1 (2021).
6. A. Nandi, N. Counts, J. Bröker, S. Malik, S. Chen, R. Han, J. Klusty, B. Seligman, D. Tortorice, D. Vigo *et al.*, Cost of care for alzheimer's disease and related dementias in the united states: 2016 to 2060, *npj Aging* **10**, p. 13 (2024).
7. 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).
8. W. K. Self and D. M. Holtzman, Emerging diagnostics and therapeutics for alzheimer disease, *Nature medicine* **29**, 2187 (2023).
9. H. W. Haddad, G. W. Malone, N. J. Comardelle, A. E. Degueure, S. Poliwoda, R. J. Kaye, K. S. Murnane, A. M. Kaye and A. D. Kaye, Aduhelm, a novel anti-amyloid monoclonal antibody, for the treatment of alzheimer's disease: A comprehensive review, *Health Psychology Research* **10** (2022).
10. A. Jiménez, M. J. Merino, J. Parras and S. Zazo, Explainable drug repurposing via path based knowledge graph completion, *Scientific Reports* **14**, p. 16587 (2024).
11. A. Renaux, C. Terwagne, M. Cochez, I. Tiddi, A. Nowé and T. Lenaerts, A knowledge graph approach to predict and interpret disease-causing gene interactions, *BMC bioinformatics* **24**, p. 324 (2023).
12. S. Jin, X. Zeng, F. Xia, W. Huang and X. Liu, Application of deep learning methods in biological networks, *Briefings in bioinformatics* **22**, 1902 (2021).
13. C. X. Alvarado, M. B. Makarious, C. A. Weller, D. Vitale, M. J. Koretsky, S. Bandres-Ciga, H. Iwaki, K. Levine, A. Singleton, F. Faghri *et al.*, omicsynth: An open multi-omic community resource for identifying druggable targets across neurodegenerative diseases, *The American Journal of Human Genetics* **111**, 150 (2024).
14. 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).
15. J. I. Castrillo, S. Lista, H. Hampel and C. W. Ritchie, Systems biology methods for alzheimer's disease research toward molecular signatures, subtypes, and stages and precision medicine: application in cohort studies and trials, *Biomarkers for Alzheimer's Disease Drug Development* , 31 (2018).
16. J. Schwartzentruber, S. Cooper, J. Z. Liu, I. Barrio-Hernandez, E. Bello, N. Kumasaka, A. M. Young, R. J. Franklin, T. Johnson, K. Estrada *et al.*, Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new alzheimer's disease risk genes, *Nature genetics* **53**, 392 (2021).

17. J. H. Moore, The Ubiquitous Nature of Epistasis in Determining Susceptibility to Common Human Diseases, *Human Heredity* **56**, 73 (11 2003).
18. T. F. Mackay and J. H. Moore, Why epistasis is important for tackling complex human disease genetics, *Genome Med.* **6**, p. 124 (June 2014).
19. J. H. Moore and S. M. Williams, Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis, *BioEssays* **27**, 637 (2005).
20. T. J. Hohman, W. S. Bush, L. Jiang, K. D. Brown-Gentry, E. S. Torstenson, S. M. Dudek, S. Mukherjee, A. Naj, B. W. Kunkle, M. D. Ritchie, E. R. Martin, G. D. Schellenberg, R. Mayeux, L. A. Farrer, M. A. Pericak-Vance, J. L. Haines and T. A. Thornton-Wells, Discovery of gene-gene interactions across multiple independent data sets of late onset alzheimer disease from the alzheimer disease genetics consortium, *Neurobiology of Aging* **38**, 141 (2016).
21. Q. Sha, Z. Zhang, J. C. Schymick, B. J. Traynor and S. Zhang, Genome-wide association reveals three SNPs associated with sporadic amyotrophic lateral sclerosis through a two-locus analysis, *BMC Med. Genet.* **10**, p. 86 (September 2009).
22. M. Steffens, T. Becker, T. Sander, R. Fimmers, C. Herold, D. A. Holler, C. Leu, S. Herms, S. Cichon, B. Bohn, T. Gerstner, M. Griebel, M. M. Nöthen, T. F. Wienker and M. P. Baur, Feasible and successful: genome-wide interaction analysis involving all 1.9 x 10(11) pair-wise interaction tests, *Hum. Hered.* **69**, 268 (March 2010).
23. 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).
24. F. J. Martin, M. R. Amode, A. Aneja, O. Austine-Orimoloye, A. G. Azov, I. Barnes, A. Becker, R. Bennett, A. Berry, J. Bhai *et al.*, Ensembl 2023, *Nucleic acids research* **51**, D933 (2023).
25. 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).
26. J. N. Hellwege, J. M. Keaton, A. Giri, X. Gao, D. R. Velez Edwards and T. L. Edwards, Population stratification in genetic association studies, *Curr. Protoc. Hum. Genet.* **95**, 1.22.1 (October 2017).
27. A. Kline and Y. Luo, Psmpy: a package for retrospective cohort matching in python, in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2022.
28. K. Al-Jabery, T. Obafemi-Ajayi, G. Olbricht and D. Wunsch, *"Computational Learning Approaches to Data Analytics in Biomedical Applications"* (Academic Press, 2019).
29. T. Akiba, S. Sano, T. Yanase, T. Ohta and M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019.
30. 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.
31. T. Chen and C. Guestrin, Xgboost: A scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016.
32. A. Bayat, B. Hosking, Y. Jain, C. Hosking, M. Kodikara, D. Reti, N. A. Twine and D. C. Bauer, Fast and accurate exhaustive higher-order epistasis search with BitEpi, *Scientific reports* **11**, p. 15923 (2021).
33. A. Nazarian, A. I. Yashin and A. M. Kulminski, Genome-wide analysis of genetic predisposition to alzheimer's disease and related sex disparities, *Alzheimer's research & therapy* **11**, 1 (2019).

34. N. Sinnott-Armstrong, Y. Tanigawa, D. Amar, N. Mars, C. Benner, M. Aguirre, G. R. Venkataraman, M. Wainberg, H. M. Ollila, T. Kiiskinen *et al.*, Genetics of 35 blood and urine biomarkers in the uk biobank, *Nature genetics* **53**, 185 (2021).