



# A Machine Learning Method to Identify Genetic Variants Potentially Associated With Alzheimer's Disease

Bradley Monk<sup>1,2</sup>, Andrei Rajkovic<sup>3</sup>, Semar Petrus<sup>4</sup>, Aleks Rajkovic<sup>5</sup>, Terry Gaasterland<sup>4</sup> and Roberto Malinow<sup>1,6\*</sup>

<sup>1</sup> Department of Neurosciences, Center for Neural Circuits and Behavior, School of Medicine, University of California, San Diego, San Diego, CA, United States, <sup>2</sup> Cognitive Science & Psychology IDP, University of California, San Diego, San Diego, CA, United States, <sup>3</sup> Department of Computer Science, Royal Holloway, University of London, Egham, United Kingdom, <sup>4</sup> Institute for Genomic Medicine, Scripps Institution of Oceanography, University of California, San Diego, San Diego, CA, United States, <sup>5</sup> Department of Pathology, Department of Obstetrics, Gynecology and Reproductive Sciences, University of California, San Francisco, San Francisco, CA, United States, <sup>6</sup> Section of Neurobiology, Division of Biological Sciences, University of California, San Diego, San Diego, CA, United States

## OPEN ACCESS

### Edited by:

Yun Xiao,  
Harbin Medical University, China

### Reviewed by:

Xiaodong Zheng,  
Anhui Medical University, China  
Michelle Lupton,  
QIMR Berghofer Medical Research  
Institute, Australia

### \*Correspondence:

Roberto Malinow  
rmalinow@ucsd.edu

### Specialty section:

This article was submitted to  
Computational Genomics,  
a section of the journal  
Frontiers in Genetics

Received: 29 December 2020

Accepted: 06 April 2021

Published: 14 June 2021

### Citation:

Monk B, Rajkovic A, Petrus S,  
Rajkovic A, Gaasterland T and  
Malinow R (2021) A Machine Learning  
Method to Identify Genetic Variants  
Potentially Associated With  
Alzheimer's Disease.  
Front. Genet. 12:647436.  
doi: 10.3389/fgene.2021.647436

There is hope that genomic information will assist prediction, treatment, and understanding of Alzheimer's disease (AD). Here, using exome data from ~10,000 individuals, we explore machine learning neural network (NN) methods to estimate the impact of SNPs (i.e., genetic variants) on AD risk. We develop an NN-based method (netSNP) that identifies hundreds of novel potentially protective or at-risk AD-associated SNPs (along with an effect measure); the majority with frequency under 0.01. For case individuals, the number of "protective" (or "at-risk") netSNP-identified SNPs in their genome correlates positively (or inversely) with their age of AD diagnosis and inversely (or positively) with autopsy neuropathology. The effect measure increases correlations. Simulations suggest our results are not due to genetic linkage, overfitting, or bias introduced by netSNP. These findings suggest that netSNP can identify SNPs associated with AD pathophysiology that may assist with the diagnosis and mechanistic understanding of the disease.

**Keywords:** machine learning, neural network, Alzheimer's, disease, polygenic

## INTRODUCTION

Alzheimer's disease (AD), the most common form of dementia, is heritable [58–79%, estimated from twin studies (Gatz et al., 2006)], and highly polygenic (Cauwenberghe et al., 2015). Mutations in three genes (*APP*, *PS1*, *PS2*) cause rare forms of the disease [accounting for ~1% of AD cases (Mendez, 2017)], which shows autosomal dominant transmission with high penetrance and displays an early onset [generally before age 60 (Carmona et al., 2018)]. In the more common form of the disease, late onset AD (LOAD), *APOE* has been established unequivocally as a susceptibility gene (Saunders et al., 1993) with several dozen other genetic loci receiving genetic support (Carmona et al., 2018; Jansen et al., 2019; Kunkle et al., 2019).

The neuropathology of AD is defined by the presence of extracellular senile plaques containing amyloid beta 42 and intracellular neurofibrillary tangles containing hyperphosphorylated tau protein (DeTure and Dickson, 2019). The neuropathological progression of disease has been best described using the Braak staging scheme (I–VI) (Braak et al., 2006). The most important genetic

variant in LOAD is the *APOE*  $\epsilon 4$  isoform, which predisposes patients to an earlier appearance of AD and a higher Braak score. The role of *APOE* or other identified genetic variants in the pathophysiology of AD is not well understood (Sisodia and George-Hyslop, 2002; Koffie et al., 2012; Karch and Goate, 2014; Shi et al., 2017). Currently available disease biomarkers can be expensive, labor intensive, and do not provide a definitive clinical diagnosis (Gustaw-Rothenberg et al., 2010; Hampel et al., 2018; Jack et al., 2018; Penner et al., 2019). The identification of additional LOAD-linked genetic variants could potentially increase diagnostic accuracy, increase our understanding of the disease, and unmask potential drug targets.

In 2009 two high-powered genome-wide association (GWAS) studies were published that identified, along with *APOE*, several single nucleotide polymorphism (SNPs) loci significantly linked to AD, including SNPs in *PICALM*, *CLU*, and *CR1* (Harold et al., 2009; Lambert et al., 2009). To continue the search for genetic targets linked to AD, the Alzheimer's Disease Sequencing Project (ADSP) was established as a collaboration between the National Human Genome Research Institute (NHGRI) and the National Institute on Aging (NIA). As part of this effort, whole-exome sequencing was performed on 5,740 LOAD cases and 5,096 cognitively normal, older individuals, serving as controls (Bis et al., 2018). The overarching goals of this initiative have been to identify novel genomic targets that contribute risk or confer protection toward AD outcomes, and to develop new insights as to why some at-risk individuals do not develop AD. Indeed, data from this project have been used to identify a number of novel SNPs linked to AD (Beecham et al., 2018; Bis et al., 2018; Raghavan et al., 2018; Ma et al., 2019; Patel et al., 2019; Zhang et al., 2019).

Recent studies have presented polygenic risk score (PRS) models for estimating AD risk (Escott-Price et al., 2015, 2017; Desikan et al., 2017; Zhang et al., 2020). In these models, GWAS summary data were used to identify AD-linked genomic variants and to assign each variant a coefficient based on their case-control asymmetries. While PRS models are a powerful method to identify individuals at risk for a disease, we believe they could provide another powerful utility – identifying novel genetic variants that confer AD risk or protection that escape GWAS identification for a number of factors, including rarity and potential interactions (linear and non-linear) with other variants. However, in order to capture these interactions the PRS model needs to be based on individual genotypes (not just GWAS summary data). Here we developed such a model based on individual AD case and control SNP data provided by ADSP. Specifically, artificial neural networks (Sejnowski, 2020) were trained using individual case and control genotypes to estimate polygenic risk. A primary aim of this study was to develop a machine learning-based method (netSNP) that can be used to identify the importance of individual SNPs in a complex polygenic classifier's decision making process.

netSNP can identify hundreds of AD-linked SNPs, most of which have a low population frequency ( $<0.01$ ). Supporting the validity of our method are the observations that the number of AD-linked SNPs identified by this method that are harbored by an individual diagnosed with AD is correlated with the age

at which that individual's AD was diagnosed as well as their brain pathology. In particular, the number of risk- (or protection-) linked SNPs an individual harbors correlates negatively (or positively) with the age at which an individual is diagnosed with AD and with their Braak score (i.e., individuals with more risk SNPs had AD at earlier ages and higher Braak scores; individuals with more protective SNPs had AD at later ages and lower Braak scores). Furthermore, scaling the SNPs with a netSNP-derived "importance factor" further increases the correlations. Thus, these correlations provide support for the view that this method correctly identifies AD-linked SNPs and correctly quantifies their importance.

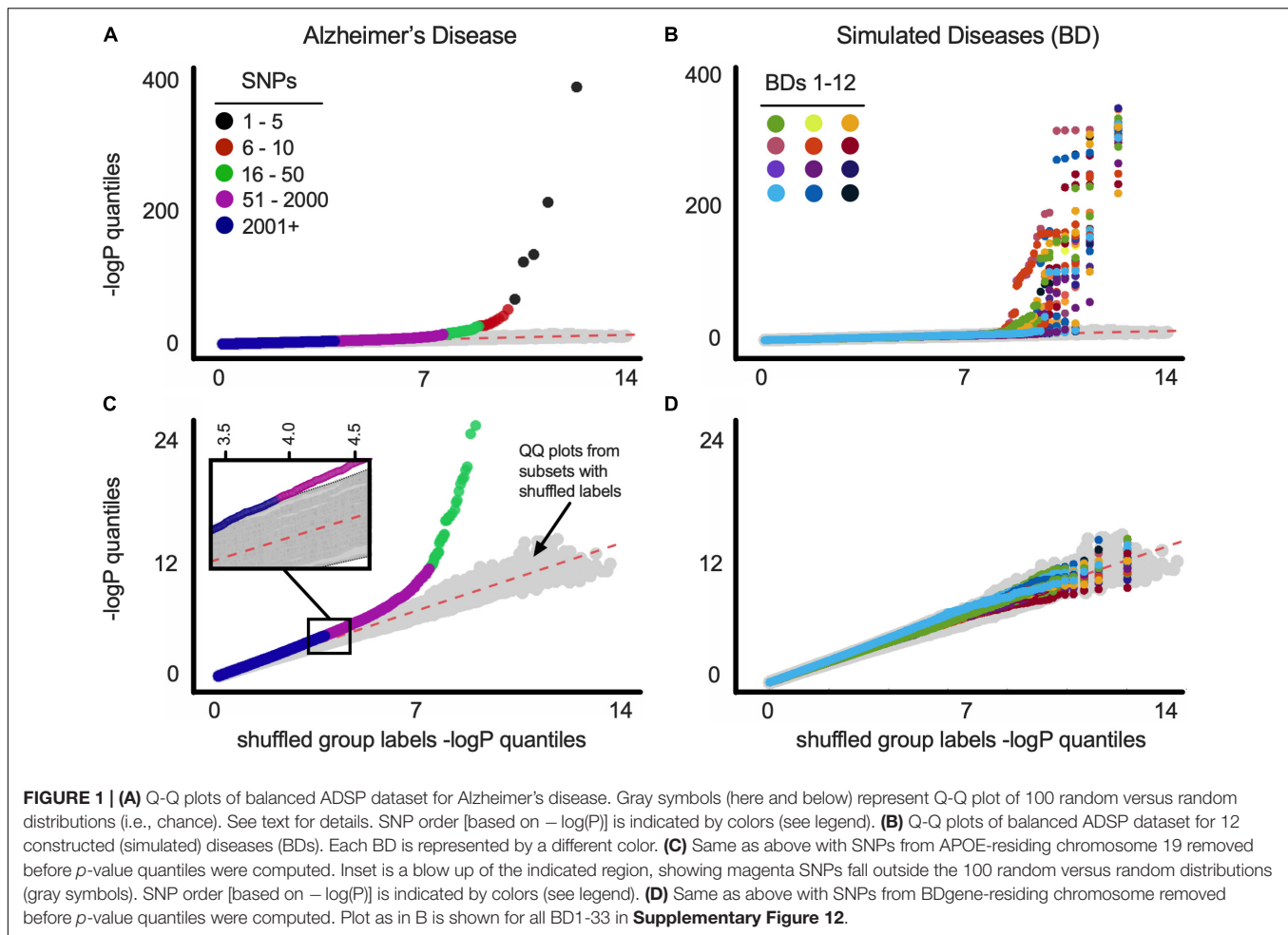
## RESULTS

### Dataset Pipeline, Case:Control Balancing, and SNP Properties

A large variant call format (VCF) datafile [ $\sim 200$  GB; Alzheimer's Disease Sequencing Project, ADSP (Beecham et al., 2017)] containing SNP information (i.e., reference or alternate allele for  $\sim 1.4$  million SNP sites) on  $\sim 11,000$  individuals over the age of 60 (Northern European descent;  $\sim 6,000$  diagnosed with AD, and  $\sim 5,000$  aged non-AD controls), was organized into a more manageable file ( $\sim 2$  GB; N.B.: a VCF datafile contains mainly zeros – indicating reference alleles – since  $>95\%$  of minor allele frequencies are  $<0.01$ ) permitting rapid queries regarding SNP content for any individual (see section "Materials and Methods"). The minor frequency allele (MFA) and reference allele count were determined at each SNP locus, separately for case and control groups. The Fisher's exact test was used to quantify the probability (FishP) that the observed case/control minor allele asymmetry could be due to chance.

The ADSP dataset consists of SNP information originating from 24 cohort groups (Beecham et al., 2017; Crane et al., 2017; Naj et al., 2018). We initially tested if a neural net (NN) classifier could be trained (Moller, 1993), with SNPs as features (50 SNPs with the lowest FishP values; 50 features were chosen as this was computationally tractable; see section "Materials and Methods"), to identify from which cohort an individual originated. Indeed, the classifier could identify cohort identity for each individual with  $\sim 50\%$  accuracy, much above the 4% expected by chance (**Supplementary Figures 1, 2**). This was of concern, because given the large case:control imbalance in many cohorts (see **Supplementary Figure 2**), the classifier could use cohort information to help identify patient AD status. Thus, the SNPs would indicate something about the cohort (e.g., platform- or probe-specific aspects of cohorts) rather than the disease. To avoid this potentially confounding issue, we balanced cohorts. In short, (a) only cohorts with at least 20% of the cohort being cases or controls were used; and (b) the same number of case and control individuals from each cohort was used in training sets (see **Supplementary Figure 2** and "Materials and Methods" for details).

A quantile-quantile (Q-Q) plot of the  $-\log(\text{FishP})$  values of a balanced dataset plotted against a similar computation of the same dataset with shuffled case-control labels shows that



most of the case-control minor allele asymmetries across the 1.4 million SNP loci can be explained by chance (i.e., lie on the  $x = y$  line; **Figures 1A–D**). For comparison we plotted 100 Q-Q plots, wherein  $-\log(\text{FishP})$  values from one shuffled dataset was plotted against  $-\log(\text{FishP})$  values from another shuffled dataset (**Figures 1A–D**, gray symbols). For the AD population, in a few SNPs from *APOE* and (its physically close linkage partner) *TOMM40* genes (Yu et al., 2007; Guerreiro and Hardy, 2012), the observed distribution of reference allele (Ref) and MFA in the case and control populations was far (orders of magnitude) from what can be accounted for by chance (**Figures 1A,C**).

To address the possibility that artifacts can account for SNPs off the  $x = y$  line (e.g., SNPs being linked to *APOE*, rather than to AD), we constructed 33 separate simulated diseases (“bad diseases,” BDs) using all ADSP individuals (cf., Bulik-Sullivan et al., 2015). Each BD was based on an existing gene (BDgene) that has two SNPs with frequencies in our population very close to *APOE* $\epsilon$ 4 (0.147, E4-like) and *APOE* $\epsilon$ 2 (0.076, E2-like); see **Table 1**, MAF (minor allele frequency) columns. Individuals with the BDgene genotype (i.e., having E4-like or E2-like SNPs) in the ADSP dataset were randomly ascribed to have BDs based on control/case odds ratio of *APOE* $\epsilon$ 4 (0.30) and *APOE* $\epsilon$ 2 (2.41) for AD. Individuals without BDgene SNPs were randomly assigned

based on the odds ratio of individuals without *APOE* $\epsilon$ 4 or  $\epsilon$ 2 (i.e., *APOE* $\epsilon$ 33 = 0.89). FishP values were computed for each SNP from true (AD) and simulated (BDs) diseases from balanced data sets, and Q-Q plots were generated (**Figures 1B,D**). Plots including all SNPs showed many with FishP values outside what could be accounted for by chance for both AD and BDs (**Figures 1A,B**). However, if SNPs from chromosome 19 (where *APOE* resides) or the chromosome with BDgene were removed, only SNPs for AD could not be accounted for by chance (**Figures 1C,D**). This result is consistent with the view previously observed that AD is a highly polygenic disorder (Cauwenbergh et al., 2015; Escott-Price et al., 2017) as there was a considerable asymmetry in MAF between case and control populations for over 2,000 SNPs (see **Figure 1D**). While artifacts related to data stratification can account for this behavior in Q-Q plots (Lander and Schork, 1994; Slatkin, 2007), cohort balancing and our simulations argue against such artifacts for our dataset, and support the existence of a large number of SNPs linked to AD, consistent with previous results (Cauwenbergh et al., 2015; Escott-Price et al., 2017).

## NN Construction and Performance

Once the cohorts were balanced, we calculated the FishP values for SNPs from a “training set” composed of 3,200 randomly

**TABLE 1** | netSNP identified tSNPs with greatest absolute average CVt when APOE locus variants were not excluded from the training set.

tSNPs predicted to confer most protection against AD						tSNPs predicted to confer the most risk for AD					
Chr	Pos	Gene	mCVt	FishP	MAF	Chr	Pos	Gene	mCVt	FishP	MAF
5	612,536	CEP72	-0.243	5.7E-03	0.002	11	10,327,875	ADM	0.289	4.1E-08	0.008
6	1,390,303	FOXF2	-0.182	1.4E-02	0.003	19	45,411,941	APOE ε4*	0.261	3.4E-111	0.135
4	110,638,764	PLA2G12A	-0.178	2.8E-02	0.002	7	23,213,734	KLHL7	0.217	5.1E-03	0.003
1	16,890,642	NBPF1	-0.174	3.3E-02	0.002	9	130,439,029	STXBP1	0.207	1.9E-02	0.001
11	1,017,294	MUC6	-0.157	1.4E-07	0.01	20	37,258,198	ARHGAP40	0.203	1.1E-02	0.002
1	40,961,395	ZFP69	-0.156	4.8E-04	0.002	15	41,862,356	TYRO3	0.197	8.4E-18	0.018
15	50,154,563	ATP8B4	-0.156	2.3E-02	0.004	6	146,276,263	SHPRH	0.195	6.6E-03	0.002
19	52,793,834	ZNF766	-0.155	2.8E-02	0.002	1	228,879,367	RHOU	0.195	6.4E-03	0.004
15	64,017,685	HERC1	-0.152	3.5E-03	0.004	9	131,398,647	WDR34	0.19	7.0E-03	0.002
19	45,412,079	APOE ε2*	-0.152	7.1E-38	0.079	19	52,497,235	ZNF615	0.188	3.3E-02	0.003
16	8,740,006	METTL22	-0.15	2.1E-03	0.002	12	85,450,243	LRRIQ1	0.188	1.0E-02	0.006
9	139,396,933	NOTCH1	-0.143	3.3E-02	0.004	15	25,963,545	ATP10A	0.185	1.2E-02	0.002
19	18,561,473	ELL	-0.137	7.3E-03	0.008	12	108,011,971	BTBD11	0.183	3.8E-03	0.007
11	57,467,411	ZDHHC5	-0.133	3.5E-03	0.002	9	107,533,232	NIPSNAP3B	0.181	1.1E-02	0.003
9	100,372,648	TSTD2	-0.131	2.2E-02	0.003	1	8,420,270	RERE	0.18	3.0E-02	0.004
1	65,120,426	CACHD1	-0.131	1.0E-02	0.002	8	10,480,495	RP1L1	0.179	3.1E-02	0.003
12	69,113,184	NUP107	-0.126	5.6E-03	0.006	4	5,682,993	EVC2	0.178	2.2E-02	0.004
5	145,508,644	LARS	-0.126	1.2E-02	0.006	5	140,530,973	PCDHB6	0.178	1.4E-02	0.002
7	6,561,105	GRID2IP	-0.125	2.9E-04	0.002	6	30,712,298	IER3	0.177	3.3E-03	0.007
19	43,268,061	PSG8	-0.125	2.9E-02	0.004	15	50,264,839	ATP8B4	0.176	1.1E-02	0.008
11	47,264,353	ACP2	-0.125	1.4E-03	0.004	16	3,604,305	NLRC3	0.176	2.7E-02	0.002
6	7,405,242	RIOK1	-0.124	1.5E-02	0.003	22	46,780,446	CELSR1	0.174	1.9E-02	0.003
3	146,167,089	PLSCR2	-0.123	2.3E-02	0.003	19	39,103,307	MAP4K1	0.173	1.9E-02	0.001
16	30,775,522	RNF40	-0.123	2.9E-02	0.006	1	89,579,827	GBP2	0.173	2.5E-02	0.005
9	139,008,644	C9orf69	-0.121	2.9E-03	0.003	12	50,500,080	GPD1	0.173	2.6E-03	0.002

Rows 26:1000 available online

Rows 26:1000 available online

\* Previously published AD-linked gene

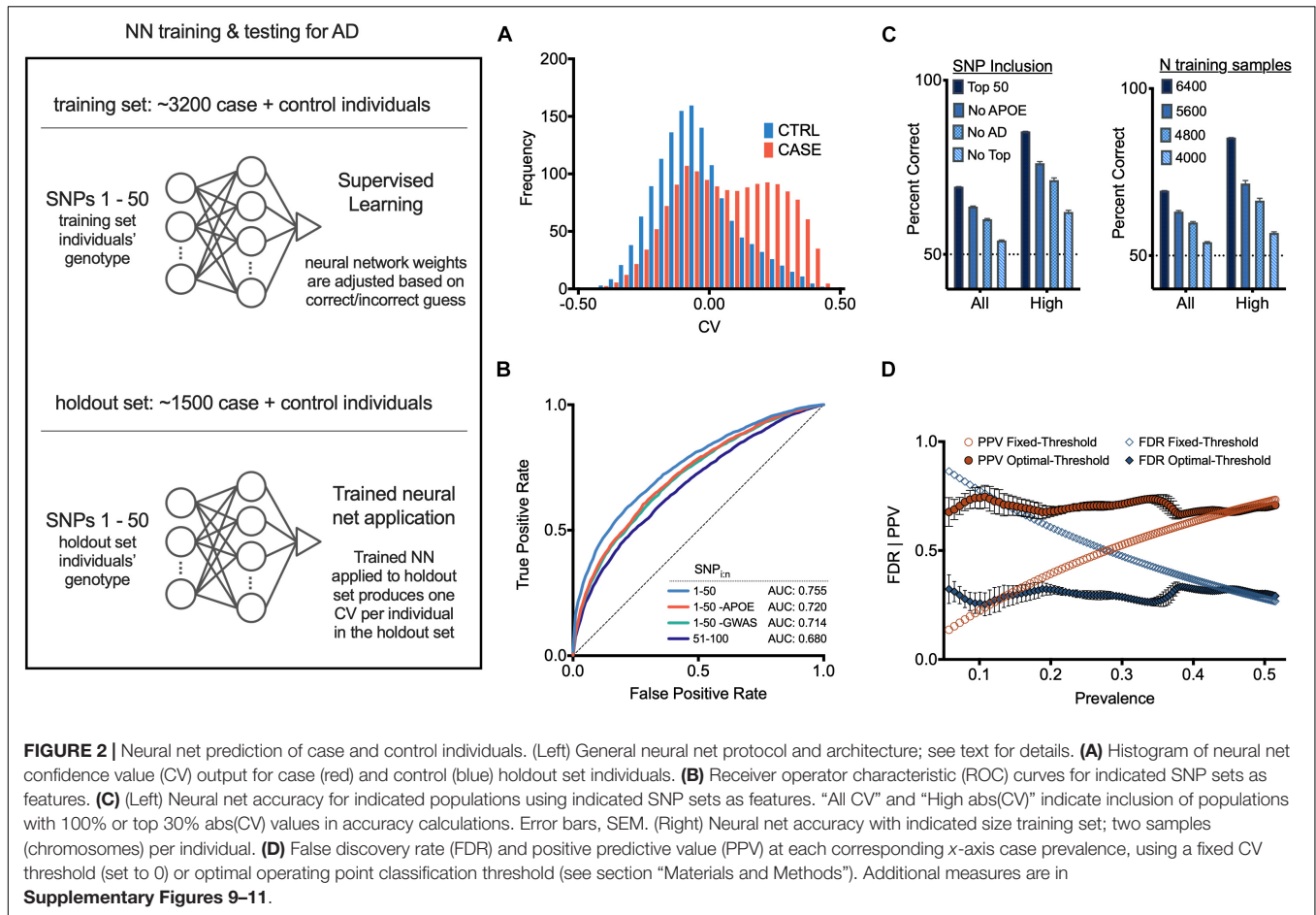
\* Previously published AD-linked gene

chosen individuals (case + controls; equal number of each) and used the 50 SNPs with the lowest FishP values in the training set to train an NN classifier to predict if an individual was a case or control (Figure 2, left; see section “Materials and Methods”). Briefly (Demuth et al., 2014), an artificial NN was trained to classify cases vs. controls using genotypes (for 50 SNPs) of individuals in the training set. The NN was initialized with random weights connecting each node, so the initial prediction  $y$  was random (each  $y$  was a real number scaled between  $-0.5$ , the control label, and  $+0.5$ , the case label). This prediction, also known as a classifier value (CV), was evaluated against the true label (case or control) using a loss function, and the network weights were updated using an optimization function. Throughout training the optimizer adjusts NN weights, working to minimize the loss function. Training concluded when the NN weights were considered optimal (within the constraints of the stopping criteria and cross validation; see section “Materials and Methods”), at which point the NN weights remain fixed. Thus, additional input to the NN would yield CV predictions, but would *not* change network weights or alter the model in any regard.

After the NN was trained as described above, it was applied to the 50 SNPs of each individual from the “holdout set”

(1,500 individuals randomly chosen who were not included in the *training set*), providing a CV for each. Overall these CVs correlated well with actual AD status of each individual (case, red; control, blue; Figure 2A and Supplementary Figure 4). Using a classification threshold of zero (such that any positive CV was predicted as *case*, and any negative CV was predicted as *control*), the classifier accuracy was 67.3% (SD = 0.3%, see section “Materials and Methods”). The NN performance with 50 SNPs was significantly better than what could be achieved using only SNPs from the *APOE* locus (62.2%). It also performed better than a logistic regression model using the same 50 SNPs (64.2%,  $p < 10e-20$ , McNemar Test). When only considering individuals with CVs closer to  $-0.5$  or  $+0.5$ , the accuracy of the NN increased. For individuals with CVs in the outer quartiles, prediction accuracy was 76.4% (SD = 0.5%); for those with a CV ranked in the upper 12.5% and lower 12.5% quantiles, the classification accuracy was 82.6% (SD = 0.6%) (see Figure 2C).

We next trained an NN using a set of 50 SNPs (a) not containing *APOE* gene SNPs, or (b) not containing the 22 previously published AD-associated SNPs (Carmona et al., 2018), or (c) with the 51–100 lowest FishP values. The resulting accuracy and receiver operator characteristic (ROC) curves, which provide a measure of the sensitivity and specificity of a



method (Koen et al., 2016), were all above chance in predicting AD status of an individual (Figures 2B,C). Reducing the size of the training set reduced the accuracy in a roughly linear fashion (Figure 2C), suggesting that the NN accuracy did not asymptote at 3,200 individuals, and that gathering SNP information from more individuals would increase NN accuracy. The area under the ROC curve (AUC) for our NN model with 50 SNPs was 0.755. Further analysis of NN hyperparameters such as the number of SNPs, which SNPs were employed, NN architecture, etc., may improve NN performance; we note that producing an optimal NN was not the primary goal of this study. Other methods, such as PCA (Jolliffe, 1986; Selzam et al., 2018), or Random Forest (Goldstein et al., 2011) analyses were not examined.

Due to the cohort counterbalancing requirement (see above), the prevalence of AD in our training set was 0.5. Since disease prevalence in most populations will almost certainly be lower than 0.5, we quantified signal detection metrics for a range of disease prevalence rates from 0.05 to 0.5 (0.05 is the approximate AD prevalence at age 75), using the optimal operating point (OOP) for each respective base rate (see section "Materials and Methods"). Using the OOP, the false discovery rate was largely independent of prevalence for values from ~0.05 to 0.5 (Figure 2D). Similarly, the same optimal threshold maintained a largely constant positive predictive rate (Figure 2D). Thus,

computing an NN with training data composed of an equal number of cases and controls can be used despite a low disease prevalence.

### netSNP Description and Application

While NNs can perform well in solving complex problems, determining the importance of different NN input features (in this case, different SNPs) is difficult to assess. With this in mind, we developed a method (netSNP) using a modification of the standard NN protocol, aimed to assess the impact of any SNP on conferring AD risk or protection. Specifically, we derived a quantitative measure for the impact of an SNP on the output of an NN.

netSNP is a modification of the *Permutation Importance* method used in machine learning (Altmann et al., 2010; Molnar, 2021), which we have adapted for use with polygenic models. In general Permutation Importance is used to address the question "What variables have the biggest impact on the predictions of a trained neural network classifier?" Permutation Importance computations are performed after a model has already been fitted, and works using a basic strategy: a single predictor variable is modified in the input data, leaving all the other predictor variables unchanged, and examining how this affects classifier

performance. This procedure is then repeated, one variable at a time, for all the predictor variables used in the model. This permits one to determine the relative effect of each predictor variable. The netSNP method uses a similar strategy. For a specific SNP, netSNP addresses this question “if this SNP is artificially made homozygous for the MFA, what impact does it have on the classifier output?” If the average CV shifts to the right (e.g., goes from 0.1 to 0.3) when a SNP is set to homozygous for the MFA, netSNP deems this SNP to confer risk. If the average CV shifts to the left (e.g., goes from 0.1 to  $-0.2$ ) when a SNP is set to homozygous for the MFA, netSNP deems this SNP to confer protection.

To demonstrate the netSNP method on a specific example, we used netSNP to compute the impact of the *APOE* genotype on AD risk. From a balanced dataset, we randomly chose a training set composed of 3,200 cases+controls (which contained individuals with all *APOE* genomic variants; i.e., *APOE*  $\epsilon 22$ ,  $\epsilon 23$ ,  $\epsilon 24$ ,  $\epsilon 33$ ,  $\epsilon 34$ , and  $\epsilon 44$ ). This set was used to train an NN (which we call NN $\epsilon$ ) to identify cases or controls based on their top 50 SNPs (see section “Materials and Methods” and **Figure 3** left panel, top). After this training session, NN $\epsilon$  was not modified in the subsequent analysis of the *APOE* genomic variants. We then applied NN $\epsilon$  to a holdout set of 1,500 individuals (**Figure 3**, left panel, bottom), producing 1,500 CV outputs with a distribution shown in **Figure 3A** (dashed line; this is used as a *baseline* for comparisons). We next reasoned that the impact of a specific *APOE* genotype on NN $\epsilon$  predictions could be assessed by artificially modifying the *APOE* genotype of every holdout set individual to that specific *APOE* genotype. For instance, to assess the impact of the  $\epsilon 22$  genotype, we artificially assigned every holdout set individual the *APOE*  $\epsilon 22$  genotype (keeping non-*APOE* genotypes of each individual unaltered). After applying NN $\epsilon$  to these modified genotypes, the distribution of CV outputs was strongly shifted leftward compared to the *baseline* distribution (**Figure 3A**, compare blue distribution to dashed line). Alternatively, if we assigned all holdout set individuals the  $\epsilon 44$  genotype, the CV distribution shifted significantly rightward from *baseline* (**Figure 3A**, compare orange distribution to dashed line). Falling between the  $\epsilon 22$  and  $\epsilon 44$  distributions were the CV distributions when NN $\epsilon$  was applied to holdout set individuals assigned either the  $\epsilon 23$ , or  $\epsilon 33$ , or  $\epsilon 24$ , or  $\epsilon 34$  genotype (**Figure 3A**).

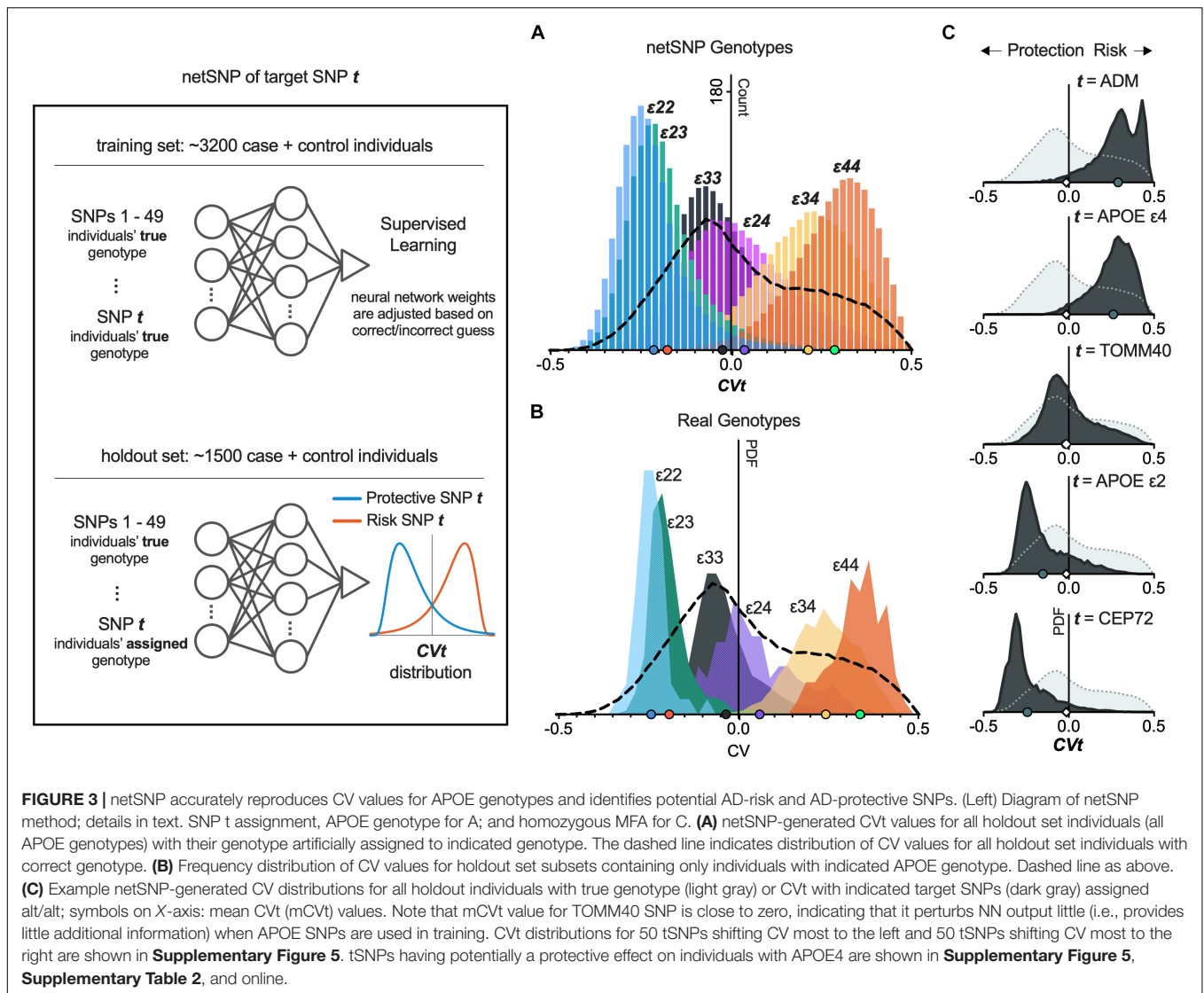
We next performed a critical test of netSNP: to determine if the above (i.e., the colored CV distributions in **Figure 3A**) corresponded to distributions when NN $\epsilon$  was applied for individuals who did have distinct *APOE* genotypes. To test for this, we created holdout sets with individuals with only one *APOE* genotype (i.e., one holdout set included only *APOE*  $\epsilon 22$  individuals, another holdout set only *APOE*  $\epsilon 23$  individuals, etc.). We then used NN $\epsilon$  to compute CVs for individuals in each of these holdout sets (using the true genotypes for each individual, for 50 SNPs). The resulting CV distribution for true *APOE* genotype holdout sets moved from left to right as *APOE* changed from  $\epsilon 22$  to  $\epsilon 44$  (**Figure 3B**), closely matching the CV distributions from above, where *APOE* status was assigned to all individuals in the holdout set (compare **Figures 3A,B**). This result suggests that netSNP

can accurately assess the impact of individual SNPs on a classifier output.

Since *APOE* SNPs are known to significantly impact AD risk, this result also suggests that the netSNP method could be used to estimate the impact of many different target SNPs of interest (which we call *tSNPs*) on AD risk. We achieved this for each *tSNP* by performing the following procedure (analogous to the procedure used to test the impact of *APOE* genotypes above; see **Figure 3**, left panel): From a balanced dataset, we randomly chose a training set composed of 3,200 cases+controls. This set was used to train an NN (which we call NN $t$ ) to identify cases or controls based on their true genotypes for 50 SNPs (the top 49 SNPs based on FishP value, and the *tSNP* of interest). Then we constructed a holdout set of 1,500 individuals, and applied NN $t$  on each individual, using the same 50 SNPs used in training, and using the true genotypes of each individual. This produced 1,500 baseline CV values. Finally we constructed a holdout set of 1,500 individuals, and used the same 50 SNPs, using their true genotypes for each individual for 49 SNPs, but the *tSNP* was set to be homozygous for the *tSNP* MFA. We then applied the NN $t$  producing 1,500 CV values (which we call *CV $t$* ) which can be plotted in a frequency distribution (**Figure 3**, bottom). We repeated this procedure for many *tSNPs* (see section “Materials and Methods”; **Figure 3C** shows *CV $t$*  distributions for several *tSNPs*). Intuitively, we reasoned that if an SNP had an effect on AD risk, then when evaluated as a *tSNP*, the *CV $t$*  distribution would be shifted compared to the baseline CV distribution – shifts to the left would indicate the MFA SNP is AD-protective (**Figure 3**, left panel, “Protective SNP  $t$ ” distribution); shifts to the right would indicate the MFA SNP incurs AD risk; the larger the shift, the greater the impact on AD. We test this proposal below.

We used netSNP to test 4,000 individual SNPs as *tSNPs*; we chose those SNPs with the 4,000 lowest FishP values. Each *tSNP* was evaluated 20 times (see section “Materials and Methods”) from which a mean *CV $t$*  (*mCV $t$* ) is computed over all holdout set individuals for all 20 runs. Evaluating *APOE* $\epsilon 4$  as a *tSNP* with netSNP resulted in a *CV $\epsilon 4$*  distribution that was shifted to the right (**Figure 3C**, 2nd from top; same as **Figure 3A**, green), as expected. Surprisingly, the MFA of an adrenomedullin (*ADM*) SNP shifted the CV distribution more to the right than *APOE* $\epsilon 4$  (*mCV $\epsilon 4$*  =  $0.26 \pm 0.001$ ; *mCV $_{ADM}$*  =  $0.29 \pm 0.001$ ). Also, a number of SNPs shifted NN $t$  output CVs more to the left than *APOE* $\epsilon 2$  (e.g., *mCV $\epsilon 2$*  =  $-0.15 \pm 0.001$ ; *mCV $_{CEP72}$*  =  $-0.24 \pm 0.001$ ; see **Table 1** for *tSNPs* with the most extreme *mCV $t$* ). Thus netSNP appears to identify a number of SNPs that can considerably shift NN output CV, potentially identifying SNPs that confer AD protection (shifting CV to the left) and AD risk (shifting CV to the right).

To exclude the artifactual possibility that netSNP was dependent on *APOE*, we repeated the netSNP method with *APOE* (and *TOMM40*) SNPs excluded from the 49 SNPs with the lowest FishP values as features in training NN $t$  (although *APOE* was tested as a target *tSNP*). Results were very similar to the above, with hundreds of *tSNPs* shifting CV to the right (potentially AD risk SNPs) and hundreds of *tSNPs* shifting CV to the left (potentially AD protective SNPs; **Supplementary Table 1**).



In general, this method provides a quantitative measure of the impact (as indicated by *mCVt* values) of specific SNPs on NN output, and potentially (see below) the effect of such SNPs on developing AD.

### NN and CV as Predictors of AD and Its Pathophysiology

While CV values (computed with or without APOE as an NN feature) predict well the likelihood of an individual being diagnosed with AD (**Supplementary Figure 4**), we aimed to determine if CV values correlate with the pathophysiology underlying AD. We reasoned that individuals diagnosed with AD at an earlier age may have a more aggressive form of the disease, which could be a consequence of their genetics, and this might be detected by more positive CV values; equivalently, AD diagnosis at an older age may correlate with less aggressive AD pathophysiology, and may have more negative CV values. This reasoning is supported by previous findings with APOE

genotypes (Corder et al., 1993), which we found to also be true in our dataset (**Figure 4A**). Linear regression fitting shows that, for case individuals, as their APOE $\epsilon 2$  allele count increases, so does their observed disease onset age [ $F(2,4750) = 86$ ,  $\beta$  (slope; indicating number of years per  $\epsilon 2$  allele) = 3.8,  $p < 2.6e-20$ ; general linear model, see section “Materials and Methods”]; conversely, the number of APOE $\epsilon 4$  alleles reduces the age of AD diagnosis [ $F(2,4750) = 1910$ ,  $\beta = -8.4$ ,  $p < 1e-300$ ]. With this reasoning in mind, we tested and found that the age at which cases were diagnosed with AD could be predicted by their CV values [as computed in section “NN Construction and Performance”; more positive CV for younger age of AD diagnosis,  $F(2,4752) = 571$ ,  $\beta = -27$ ,  $p < 2.3e-119$ ; **Figure 4B**]. Furthermore, an individual’s CV was positively correlated with Braak score, for case individuals receiving autopsy [ $F(2,2025) = 154$ ,  $\beta = 1.8$ ,  $p < 4.1e-34$ ; **Figure 4G**]. These effects were also highly significant if APOE was not included in the NN calculation of CV [CV vs. age,  $F(2,4752) = 422$ ,  $\beta = -22.4$ ,

$p = 5.3e-90$ ; CV vs. Braak,  $F(2,2025) = 59$ ,  $\beta = 1.7$ ,  $p = 1.8e-14$ ]. These findings support the view that the NN output value CV, as described above in section “NN Construction and Performance,” is related to the pathophysiology of AD.

## netSNP as Predictor of AD-Linked *tSNPs* and AD Pathophysiology

We next tested if netSNP can identify AD-linked SNPs and can quantify their impact on the likelihood of developing AD. We considered a set of *tSNPs* for which their computed *mCVt* values were significantly ( $p < 0.05$ ) outside the range of *mCVt* values generated by randomly choosing target SNPs from the set of all  $1.4 \times 10^6$  ADSP SNPs (see section “Materials and Methods”). This resulted in 851 *tSNPs* with *mCVt*  $< 0$  (provisionally indicated “AD-protective *tSNPs*”) and 672 *tSNPs* with *mCVt*  $> 0$  (“AD-risk *tSNPs*”), the majority (64%) with MAF under 0.01. Only some of the previously published AD-linked SNPs (which we exclude from the subsequent validation analysis) are in these sets (see **Table 1**). Using a general linear model, we found that the number of “AD-protective *tSNPs*” harbored by each case individual correlated positively with their age of AD diagnosis [ $F(2,4750) = 13.9$ ,  $\beta = 0.072$ ,  $p < 1.9e-4$ ; **Figure 4C**], while the number of “AD-risk *tSNPs*” they harbored correlated inversely with age of AD diagnosis [ $F(2,4750) = 18.2$ ,  $\beta = -0.11$ ,  $p < 1.9e-05$ ; **Figure 4C**]. Providing *tSNPs* with a *CVt* weight increased the positive correlation between *CVt*-weighted “AD protective *tSNPs*” [ $F(2,4750) = 400$ ,  $\beta = 22$ ,  $p < 1.6e-85$ ], or the negative correlation between *CVt*-weighted “AD risk *tSNPs*” [ $F(2,4750) = 404$ ,  $\beta = -25$ ,  $p < 2.4e-86$ ; **Figure 4D**] and age of AD diagnosis. Interestingly, the number of previously published AD risk SNPs (excluding *APOE* and *TOMM40* SNPs) per individual did not correlate with age of AD diagnosis ( $p = 0.32$ ; **Figure 4E**). However, if netSNP is used to calculate *CVt* for each of these SNPs, the number of *CVt*-weighted SNPs did correlate inversely with age of AD diagnosis [ $F(2,4750) = 419$ ,  $\beta = -22$ ,  $p < 2.2e-89$ ; **Figure 4E**], supporting the view that *CVt* provides a quantitative measure of the impact of an SNP on AD pathophysiology. We were concerned that the netSNP method may ascribe *CVt* values to SNPs based on genetic linkage to *APOE* $\epsilon 2$  or  $\epsilon 4$ , therefore we performed simulations using BD populations (see section “Materials and Methods”: *netSNP Validation Simulations*). These simulations support the view that the netSNP method is not choosing AD “protective” and AD “at-risk” SNPs based on genetic linkage or some other bias introduced in the netSNP procedure.

We also examined the relation of netSNP-identified *tSNPs* to the Braak scores that individuals (cases and controls) received during autopsy. The number of netSNP-identified “AD protective *tSNPs*” harbored per person displayed a negative correlation with Braak scores [ $F(2,2698) = 349$ ,  $\beta = -0.08$ ,  $p < 3.0e-73$ ; **Figure 4H**], while the number of netSNP-identified “AD risk *tSNPs*” harbored per person displayed a positive correlation with Braak scores [ $F(2,2698) = 272$ ,  $\beta = 0.09$ ,  $p < 3.4e-58$ ; **Figure 4H**]. These significant correlations, and the effect of providing

*CVt* weights, were obtained if *APOE* and *TOMM40* SNPs were (**Figure 4**) or were not (**Supplementary Figure 8**) included in the training matrix, indicating that the observed correlations were not driven by *APOE* (or SNPs in linkage disequilibrium with *APOE*; see **Supplementary Figure 13** and “Materials and Methods”).

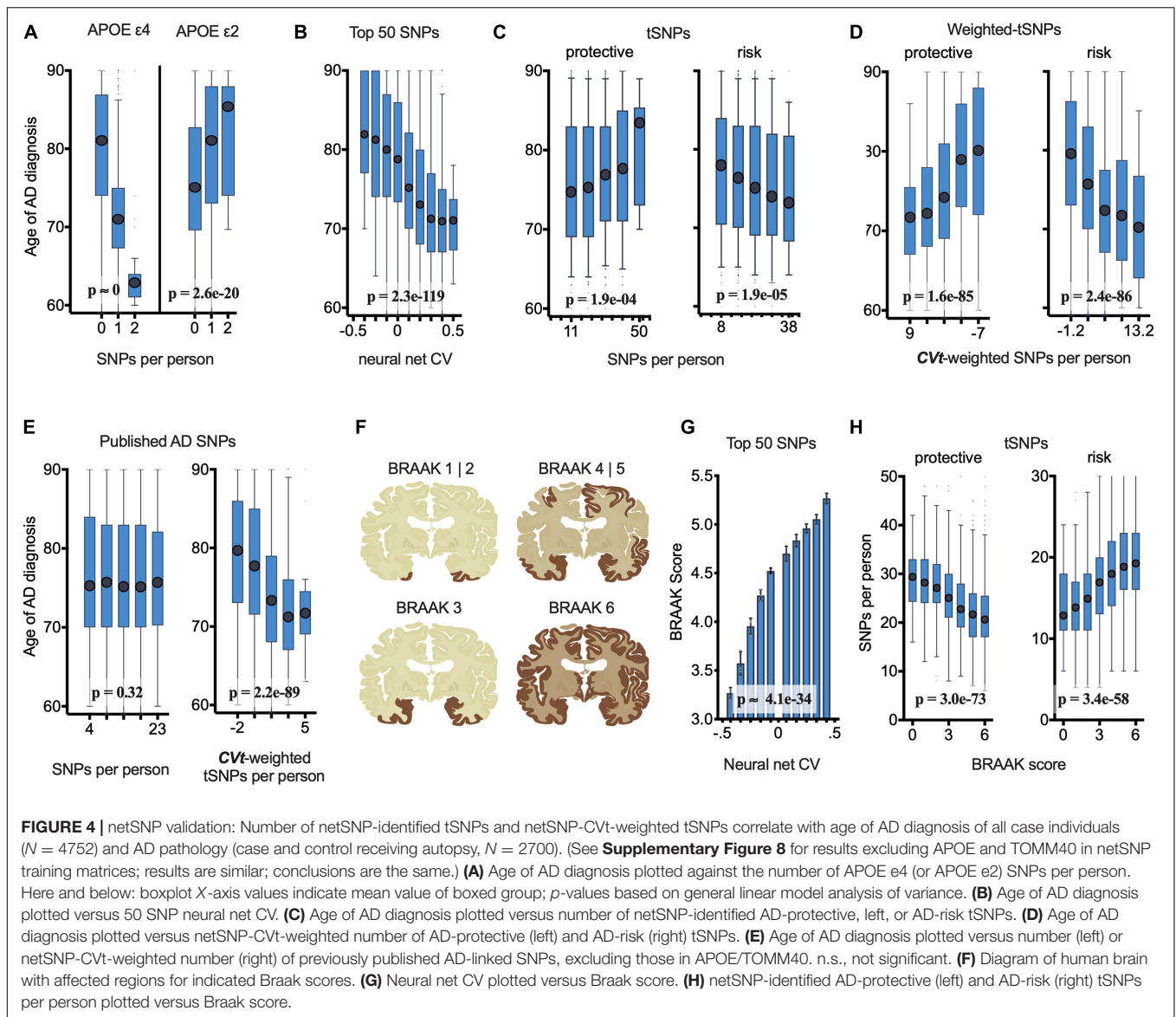
## DISCUSSION

Here we applied a standard and modified neural network tool to a large LOAD dataset and examined the association of SNPs to AD. We found that a standard NN trained with 50 SNPs can identify an individual's cohort identity above chance; thus data were subsequently analyzed using only cohorts that were case:control balanced. Comparing Q-Q plots for AD and simulated constructed diseases (based on real genes with SNPs that have true population frequency as *APOE* $\epsilon 2$  and  $\epsilon 4$ ) supports previous suggestions (Escott-Price et al., 2015, 2017) that there exist considerably more SNPs than the  $\sim 20$  previously identified as AD-associated.

An NN trained with 50 SNPs can predict dataset cases with accuracy greater (albeit, slightly) than if using only *APOE* SNPs genotypes, or a basic logistic regression model. NN accuracy was related approximately linearly with training set size, suggesting increasing dataset size will increase NN accuracy. NN accuracy was above chance if an NN was trained without (a) *APOE* SNPs, or (b) previously published AD-linked SNPs, or (c) 50 SNPs displaying the greatest case control asymmetry. These findings further support the view (Escott-Price et al., 2015, 2017) that more than the previously identified SNPs contain information regarding AD.

We developed netSNP, which investigated the impact of specific SNPs on NN output. In netSNP, once an NN was trained, the holdout set genotype was artificially assigned at a single (or multiple) target SNP(s); in the general case the target SNP was assigned as homozygous to the minor frequency allele; the effect of the artificially introduced genotype was reflected by how much the NN output value was modified. netSNP recapitulated well the effect of different *APOE* genotypes on NN output. netSNP identified several hundred SNPs with weight values (i.e., *mCVt*) significantly outside values produced by randomly chosen SNPs. Some netSNP-identified SNPs had more extreme weight values than *APOE* $\epsilon 2$  or  $\epsilon 4$ . Notably, FishP values of SNPs with extreme *mCVt* values were not low in general, likely because too few individuals carry these SNPs. Yet their impact on NN output was large, possibly by leveraging non-linear interactions embedded in an NN. Notably, ADM (containing an SNP with the largest *mCVt* value despite a MAF = 0.009) was elevated in AD brains (Ferrero et al., 2017), contributed to age-related memory loss in mice (Larrayoz et al., 2017), was elevated in aging human brains (Larrayoz et al., 2017), and had been proposed as a novel drug target for AD (Ferrero et al., 2018). We also examined ABHD17A, as it relates to findings indicating that reduced function of this enzyme increases synaptic PSD-95 levels (Jeyifous et al., 2016; Yokoi et al., 2016), which protect





synapses from beta amyloid (Malinow, unpublished observation). netSNP predicted that an *ABHD17A* SNP was protective for individuals with *APOE* $\epsilon 4$  (see **Supplementary Table 2**). Indeed, we found that  $\epsilon 4$  carrier case individuals with this *ABHD17A* SNP received an AD diagnosis almost 6 years later than such individuals without this SNP [76.6 years ( $N = 19$ ) vs. 70.8 years ( $N = 1831$ ),  $p < 0.0001$ ;  $t$ -test], which is consistent with this SNP being protective against AD in *APOE* $\epsilon 4$  carriers. These findings support the view that netSNP can identify AD-relevant SNPs.

To validate netSNP we considered variables not used in any netSNP computations: age of an individual's AD diagnosis (cf., Mars et al., 2020) and Braak score. The number of netSNP-identified "AD-protective SNPs" harbored by an individual correlated significantly with the age an individual was diagnosed with AD and inversely with Braak score; while the number of

netSNP-identified "AD-risk SNPs" harbored by an individual correlated significantly inversely with the age an individual was diagnosed with AD and positively with Braak score. Scaling each netSNP-identified SNP with CVt increased the significance of these correlations. Notably, applying netSNP-derived CVt weights to previously reported AD SNPs (each thought to have a small effect on AD pathophysiology) converted their correlation to age of diagnosis from not significant to significant, suggesting that netSNP can accurately assess small-effect SNPs. The correlations examined in this validation test hold if *APOE* or *TOMM40* are not used in the training step of netSNP, indicating that the netSNP-identified SNPs as well as the netSNP-generated CVt weights are not dependent on a bias imposed by *APOE* SNPs (or SNPs in linkage disequilibrium with *APOE*, **Supplementary Figure 13**) in netSNP. Further validation of netSNP and net-SNP-identified SNPs suggested

to be “protective” or “at-risk” in this study will require tests using an independent AD dataset as well as biological experimentation.

Our data suggest the set, as a whole, of netSNP-identified SNPs are highly predictive of AD age of onset and physiological severity, and their relative importance may be indicated by the netSNP-derived *mCVt* weight. The netSNP-identified SNPs would each, on average, be expected to have a small impact on the disease (on average  $\sim 1/200$  that of *APOE* $\epsilon$ 4; but see above for *ABHD17A* SNP). Insight into AD provided by such small-effect SNPs will require computational methods that can analyze disease and biochemical pathways from large groups of genes. Such tools may be aided by incorporation of *mCVt* values.

In general, our findings suggest that netSNP may be useful in identifying pathophysiologically relevant genes in AD; it may be equally applicable to other conditions. It will be important to test these methods on a completely independent AD dataset with similar ethnic make-up (and compare those results with results in this study), as well as AD datasets with different ethnic backgrounds, for this method to be generally applicable to the multicultural nature of the United States and world population (Martin et al., 2019).

## MATERIALS AND METHODS

### Alzheimer's Disease Sequencing Project Dataset

The dataset used in these analyses was generously provided by the Alzheimer's Disease Sequencing Project (ADSP), and has been previously described in detail in other manuscripts (Harold et al., 2009; Raghavan et al., 2018) and online at niagads.org. To summarize, individuals in this dataset were from well-characterized cohorts, including  $\sim 6,000$  individuals diagnosed with late-onset Alzheimer's disease (mean age of diagnosis: 75.4) and  $\sim 5,000$  elderly controls without dementia (mean age: 86.1, at the date of last visit to AD practitioner). Whole-exome sequencing data for each individual went through a quality-control “cleaning” process by two independent sources (Baylor and Broad Institutes), and was provided in variant call format (.vcf); genotype data was accompanied by several phenotypic and qualitative metrics (e.g., each individual's sex, age, race, cohort, etc.). For  $\sim 28\%$  of individuals an autopsy was performed and their Braak staging score was reported (Braak et al., 2006). Data are available for download upon administrative approval from the NIA Genetics of Alzheimer's Disease Storage Site (NIAGADS).

### VCF Data Compression

Raw SNP data were passed through an automated preprocessing pipeline that involved reducing the dataset size by  $\sim 100$ -fold using sparse matrices and annotating SNPs of interest. The raw data were downloaded to a secure local hard drive as VCFs. VCFs were formatted as a matrix with rows being loci and columns being samples. This matrix was converted into a structure like an adjacency list. Sample IDs were replaced with seven-digit IDs.

Flags passed through the VCFs were converted to numeric flags. Counts of homozygous and heterozygous samples, as well as the sample names and genotypes were recorded per locus. The dataset was binned into three bins according to the following criteria: first, if the genotype was heterozygous (noted as 1), or homozygous (noted as 2) for the alternate allele. The second, if the genotype was homozygous for the reference allele (noted as 0). Third, if there was missing data for that sample (noted as  $-1$ ). The combination of the bins and information contained within makes the  $\sim 100$ -fold compression conversion a lossless process. The resulting matrices were relatively small and thus easier to query/manipulate than VCFs.

### General Data Processing

Unless otherwise stated, data processing and analyses were conducted using MATLAB scientific computing software (Mathworks, 2020a,b). A compressed version of the data (as described in the section above) was imported into the MATLAB workspace. The data were then prepared for machine learning by splitting the data into training and holdout datasets. As the data were split, an attempt was made to balance cases and controls from each cohort. Cohorts that had too few cases or controls ( $<20\%$  of each other; or fewer than 20 individuals) were omitted (see **Supplementary Figure 2**). After splitting and counterbalancing, a Fisher's exact test was performed for each SNP to assign a *p*-value to the case:control asymmetries. SNPs were then sorted, ascending, by *p*-value.

### Artificial Neural Network Classification

In most instances, the model training matrix (feature matrix) consisted of individual genotypes for the 50 top SNPs after sorting SNPs by the training group's Fisher's exact test *p*-value. The rows and columns of this feature matrix represented individuals and SNPs, respectively, with each cell indicating whether a person was a homozygous reference, heterozygous, or homozygous alternate (see **Supplementary Figure 3**).

For polygenic classification we used a multilayer pattern recognition neural network (Mathworks, 2020a). This feed-forward neural net architecture can be trained to predict target classes (i.e., “labels” or “conditions” like case/control) based on a set of training features (Demuth et al., 2014). Labels for pattern recognition networks in a binary classification problem consist of a vector of 0 s and 1 s, where a 0 represents the negative condition (i.e., control), while a 1 represents the positive condition (i.e., case). In our formulation a pattern recognition network includes the following parameterization:

$$\text{patternnet}(nLayers, fTrain, fPerf)$$

where *nLayers* is the row vector of length *n*, representing the number of hidden layers; each *n*th value specifies the number of neurons in a given layer (e.g., [50, 10] would have two hidden layers of 50 neurons and 10 neurons, respectively). *fTrain* specifies the network training function (e.g., BFGS Quasi-Newton). *fPerf* specifies the performance function (e.g., cross-entropy).

We used a scaled conjugate gradient (SCG) training function for the polygenic classification task ( $fTrain = SCG$ ). The SCG network training function updates network weights and bias values using conjugate gradient backpropagation, and can be used to train any network with derivatives for weight, input, and transfer functions (Moller, 1993). With regard to network training speed, SCG is significantly faster than other conjugate gradient methods, because it does not require line searches during each machine learning iteration ( $\sim 0.1$  core hours per training session). Parameterization of the training function involves:

$fTrain$  ( $maxEpochs$ ,  $minGrad$ ,  $maxFails$ ,  $WtSigma$ ,  $Lambda$ ) where  $maxEpochs$  is the maximum number of epochs to train (e.g., 1000),  $minGradient$  is the minimum performance gradient (e.g.,  $1e-6$ ),  $maxFails$  is the maximum validation failures allowed (e.g., 10),  $WtSigma$  is the change in weight for second derivative approximation (e.g.,  $5.0e-5$ ), and  $Lambda$  regulates the indefiniteness of the Hessian (e.g.,  $5.0e-7$ ). Unless otherwise noted, the model was implemented in the MATLAB (*Mathworks - Deep Learning Toolbox*) scientific programming environment and parameterized with the following values:

```
patternnet(nLayers = (50, 10), fTrain = "SCG",
  fPerf = "cross-entropy")
SCG(maxEpochs = 1000, minGrad = 1e - 6, maxFails = 10,
  WtSigma = 5e - 5, Lambda = 5e - 7)
cross-entropy(reg = 0.1, norm = (-0.5, 0.5)).
```

The last steps involve preparing the data for network training: (1) individuals are randomly split into a training, validation, or holdout group; (2) a Fisher's exact test is used to compute the  $p$ -value associated with the case:control asymmetry in the training set at each variant locus; (3) the list of SNPs are sorted, ascending by  $p$ -value; and (4), some number of SNPs (e.g., the top 50) are selected for generating an individual-by-SNP matrix, where each cell contains the genotype of a given person at a given SNP locus. Finally, with the feature matrices prepared, and the model fully parameterized, neural net training can commence:

$$net = train(patternnet, Xt, Yt, Xv, Yv)$$

Again, *patternnet* represents the parameterized model (and all instructions for model training),  $Xt$  and  $Xv$  represent the individual-by-SNP feature matrix for the training and validation groups, respectively, and  $Yt$  and  $Yv$  are binary arrays indicating whether each person is a case or control (i.e., the condition labels). The model is trained as described above, and the final output is a fitted neural network model (a set of network weights).

## netSNP Validation Test Using BD Populations

We conducted simulations to rule out the possibility that the netSNP method may choose SNPs based on genetic linkage to APOE  $\epsilon 2$  or  $\epsilon 4$ ; i.e., significant tSNPs could

display at-risk or protective properties despite their not being pathophysiologically associated with AD. Furthermore, other details of the netSNP method may predispose cases to artifactual correlations with age of AD diagnosis and Braak scores (We note, however, that neither the age of AD diagnosis, nor their Braak score, was used in any calculations performed in section "NN Construction and Performance" or "netSNP Description and Application").

We thus tested for the correlations shown in section "NN and CV as Predictors of AD and Its Pathophysiology." for BDs 1–12 (see above; **Supplementary Table 3** and **Supplementary Figure 7**). Age of diagnosis of BD was ascribed based on APOE SNPs effects in age of AD diagnosis (using MATLAB empirical cumulative distribution functions). For each BD, a balanced dataset was constructed (as for AD, see section "Dataset Pipeline, Case:Control Balancing and SNP Properties"), and BD "protective" and "at-risk" tSNPs were identified as described for AD in section "NN and CV as Predictors of AD and Its Pathophysiology." Next, we considered the set of individuals ascribed BD. We computed a correlation probability, based on a general linear model, between their age of BD diagnosis and the number of BD tSNPs or number of BD CVt-weighted tSNPs. Results for one BD (based on a BD constructed from BDgene CHSY1; **Supplementary Figure 7D**) is compared with results for AD (**Supplementary Figure 7C**). A summary of results for the 12 separate BDs, and AD for comparison, are shown in **Supplementary Table 3**. Note that for no BD was there a significant correlation (right columns). These simulations support the view that the netSNP method is not choosing AD "protective" and AD "at-risk" SNPs based on genetic linkage or some other bias introduced in the netSNP procedure.

## Statistics

Statistical methods described per figure below.

For each BD constructed, individuals in the ADSP population were assigned a BD based on their genotype; those with APOE $\epsilon 2$ -like SNPs were randomly assigned as control with OR 2.41; those with APOE $\epsilon 4$ -like SNPs were assigned as case with OR 0.30. Those without either SNPs were assigned randomly to control with OR 0.89 (see **Supplementary Table 3**). To generate random Q-Q plots, 100 datasets were generated with randomly scrambled case-control labels. Fisher's exact test  $p$ -values were then computed for those 100 scrambled sets. Scrambled sets were plotted against each other to generate the C.I. region (gray dots) and also plotted against the actual data (colored dots).

Hundred random groups were generated with cases and controls counterbalanced within cohorts to formulate neural network training matrices. As described above, in each run one of these random groups was selected and an artificial NN was trained using the 50 SNPs with the lowest Fisher's exact test  $p$ -value among training group individuals. NN classifier performance on the holdout set was then evaluated. A histogram of each individual's mean NN classifier value (CV). Shows receiver operator characteristic (ROC) curves using SNP sets as features and normalizing CVs to range between 0 and 1: curve

“1–50” used 50 SNPs with the lowest training group  $p$ -values; “1–50 -*APOE*” used 50 SNPs with the lowest training group  $p$ -values omitting *APOE* and *TOMM40*; “1–50 -GWAS” used 50 SNPs with the lowest training group  $p$ -values omitting SNPs that previously met genome-wide significance in the literature; “51–100” used SNPs with the 51st–100th lowest training group  $p$ -values. The left panel shows the mean correct predictions in percent for each condition in (Figure 2B); the right panel was generated like the left panel’s “Top 50,” except the experimental manipulation varied the number of samples in the training group (1 sample = 1 chromosome), as indicated in the figure legend. CVs were generated like in “2B 1–50,” and normalized to a range between –0.5 and 0.5. The classification threshold was fixed at 0 and the false discovery rate (FDR) and positive predictive value (PPV) were then computed at each corresponding  $x$ -axis case prevalence. The FDR and PPV were also computed using the optimal operating point (OOP):

$$S = \frac{\text{Cost}(P|N) - \text{Cost}(N|N)}{\text{Cost}(N|P) - \text{Cost}(P|P)} * \frac{N}{P}$$

where  $\text{Cost}(N|P)$  is the cost of misclassifying a case,  $\text{Cost}(P|N)$  is the cost of misclassifying a control, where  $P = TP + FN$ , and  $N = TN + FP$  ( $TP$ , true positive;  $TN$ , true negative;  $FP$ , false positive;  $FN$ , false negative). The OOP was then determined by moving a line with slope  $S$  from  $FPR = 0$ ,  $TPR = 1$  (the top left of the ROC) down-and-right, until it intersected with the ROC curve (Mathworks, 2020b).

The histograms shown in (Figure 3A) are the result of training an NN using individuals of all *APOE* subtypes, and applying this NN on holdout set individuals assigned to each of the six *APOE* genotypes. That is, after the NN is trained as described above in *General Data Preprocessing*, all holdout individuals are assigned the *APOE*ε22 genotype and a histogram is generated; then all holdout individuals are assigned the *APOE*ε23 genotype and another histogram is generated, etc. We call this genotype assignment procedure the *netSNP method* (described below) which we show can be used as a general method for assessing the importance of any SNP on NN performance. For comparison, histograms shown in (Figure 3B) are the result of training an NN using a balanced set of individuals, and computing CVs for holdout set subgroups of individuals with the *APOE* genotypes limited to one of *APOE*ε22, ε23, ε24, ε33, ε34, or ε44. **netSNP method:** 4,000 target SNPs were chosen based on them having the lowest Fisher’s exact test  $p$ -value. For each target SNP, the netSNP method can produce a *NAT*, *REF*, *ALT*, and *DIF* value for each individual. For a single target SNP, obtaining these values involved the following steps: (1) a target SNP was selected to be part of a 50-SNP training matrix. (2) A random subset (~70%) of a balanced set of individuals served as a training set. (3) From this training set, a Fisher’s exact test  $p$ -value was calculated for each of the (~1.4 million) SNPs. (4) A single target SNP was paired with the 49 SNPs with the lowest  $p$ -value to generate a neural network training matrix. (5) The neural network was trained as described above in the *General Data Processing* methods. (6) A CV score was

generated for each of the individuals in the holdout set (*NAT* score). (7) All holdout individuals were assigned the homozygous reference genotype for the target SNP and again a CV was generated (*REF* score). (8) All holdout individuals were assigned the homozygous alternate allele (minor frequency allele) for the target SNP and a CV was generated (*ALT* score). (9) The difference between the *ALT* and *REF* scores were computed (*DIF* score). This procedure was performed 20x for each target SNP; for a given target SNP, each individual’s average *ALT* score represents that individual’s *CVt* score for the given target SNP. In this study we tested if *CVt* value could be considered a weighted measure of the impact of target SNP  $t$  on the NN. Similar to how histograms are generated for, after the NN was trained as described above in *General Data Preprocessing*, all holdout individuals were assigned the homozygous genotype for minor frequency allele of the target SNP for the indicated gene (see Table 1 for chromosome and position of the target SNP for each indicated gene).

Boxplots in (Figure 4A) were generated by grouping case individuals based on whether they had a homozygous reference, heterozygous, or homozygous minor frequency for the indicated allele, and plotted the median AD age-of-onset (+/- interquartile range, IQR; whiskers = range; dots = outliers). Boxplots were generated by pooling case individuals into six bins that were uniformly discretized based on the NN CV value, on the number of protective (left) or risk (right) target SNPs each individual had, or *CVt*-weighted target SNPs, and then plotted the median AD age-of-onset (+/- IQR; whiskers = range; dots = outliers) for each of these bins. Figure 4E (left) was generated like Figure 4C, considering previously published (without *APOE*) AD SNPs. Figure 4D (right) was generated like Figure 4C, providing a netSNP-computed *CVt* for each previously published (without *APOE*) AD SNPs. Brain sections in (Figure 4F) depict Braak staging – a method used to classify the degree of pathology in Alzheimer’s disease – commonly used in post-mortem clinical diagnosis of AD by performing brain autopsy; images here intend to summarize the general disease sequelae as shown in actual brain images from Braak et al. (2006). The bar plot in (Figure 4G) was generated by identifying individuals that had *mCVt* scores across all *tSNPs* that fall into each of the indicated bins, and the mean Braak stage of the individuals in each bin was plotted. Boxplots in (Figure 4H) pool individuals based on ADSP-reported Braak values and plot the median number of target SNPs (+/- IQR; whiskers = range; dots = outliers) found in individuals with a brain pathology that fall into one of these six Braak stages; as in (Figures 4C,D), effects are shown separately for SNPs that potentially confer protection (left panel) and risk (right panel).  $p$ -values were computed using a general linear model, where  $p$ -value represents the probability of the slope coefficient having such a magnitude, under the null hypothesis.

## DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: NIH ADSP Embargo – Access granted via application. Requests to access these datasets should be

directed to <https://dss.niagads.org/documentation/applying-for-data/application-instructions/>.

## ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

BM and RM designed the study. BM, SP, TG, and RM prepared and preprocessed the data. BM, AnR, AlR, and RM performed the statistical analyses. BM and RM generated figures. BM, AlR, and RM wrote the manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

D. H. Chen Foundation Grant “Preventing Alzheimer’s Disease,” NIH1R01-EY022306. The Alzheimer’s Disease Sequencing Project (ADSP) is comprised of two Alzheimer’s Disease (AD) genetics consortia and three National Human Genome Research Institute (NHGRI) funded Large Scale Sequencing and Analysis Centers (LSAC). The two AD genetics consortia are the Alzheimer’s Disease Genetics Consortium (ADGC) funded by NIA (U01 AG032984), and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) funded by NIA (R01 AG033193), the National Heart, Lung, and Blood Institute (NHLBI), other National Institutes of Health (NIH) institutes, and other foreign governmental and non-governmental organizations. The Discovery Phase analysis of sequence data is supported through UF1AG047133 (to Schellenberg, Farrer, Pericak-Vance, Mayeux, and Haines); U01AG049505 to Seshadri; U01AG049506 to Boerwinkle; U01AG049507 to Wijisman; and U01AG049508 to Goate, and the Discovery Extension Phase analysis is supported through U01AG052411 to Goate, U01AG052410 to Pericak-Vance, and U01 AG052409 to Seshadri and Fornage. Data generation and harmonization in the Follow-up Phases is supported by U54AG052427 (to Schellenberg and Wang). The ADGC cohorts include: Adult Changes in Thought (ACT), the Alzheimer’s Disease Centers (ADCs), the Chicago Health and Aging Project (CHAP), the Memory and Aging Project (MAP), Mayo Clinic (MAYO), Mayo Parkinson’s Disease controls, University of Miami, the Multi-Institutional Research in Alzheimer’s Genetic Epidemiology Study (MIRAGE), the National Cell Repository for Alzheimer’s Disease (NCRAD), the National Institute on Aging Late Onset Alzheimer’s Disease Family Study (NIA-LOAD), the Religious Orders Study (ROS), the Texas Alzheimer’s Research and Care Consortium (TARCC), Vanderbilt University/Case Western Reserve University (VAN/CWRU), the Washington Heights-Inwood Columbia Aging Project (WHICAP) and the Washington University

Sequencing Project (WUSP), the Columbia University Hispanic-Estudio Familiar de Influencia Genetica de Alzheimer (EFIGA), the University of Toronto (UT), and Genetic Differences (GD). The CHARGE cohorts are supported in part by National Heart, Lung, and Blood Institute (NHLBI) infrastructure grant HL105756 (Psaty), RC2HL102419 (Boerwinkle), and the neurology working group is supported by the National Institute on Aging (NIA) R01 grant AG033193. The CHARGE cohorts participating in the ADSP include the following: Austrian Stroke Prevention Study (ASPS), ASPS-Family study, and the Prospective Dementia Registry-Austria (ASPS/PRODEM-Aus), the Atherosclerosis Risk in Communities (ARIC) Study, the Cardiovascular Health Study (CHS), the Erasmus Rucphen Family Study (ERF), the Framingham Heart Study (FHS), and the Rotterdam Study (RS). ASPS is funded by the Austrian Science Fund (FWF) grant numbers P20545-P05 and P13180 and the Medical University of Graz. The ASPS-Fam is funded by the Austrian Science Fund (FWF) project I904, the EU Joint Programme – Neurodegenerative Disease Research (JPND) in frame of the BRIDGET project (Austria, Ministry of Science) and the Medical University of Graz and the Steiermärkische Krankenanstaltengesellschaft. PRODEM-Austria is supported by the Austrian Research Promotion Agency (FFG) (Project No. 827462) and by the Austrian National Bank (Anniversary Fund, project 15435). ARIC research was carried out as a collaborative study supported by NHLBI contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). Neurocognitive data in ARIC are collected by U01 2U01HL096812, 2U01HL096814, 2U01HL096899, 2U01HL096902, 2U01HL096917 from the NIH (NHLBI, NINDS, NIA, and NIDCD), and with previous brain MRI examinations funded by R01-HL70825 from the NHLBI. CHS research was supported by contracts HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, and grants U01HL080295 and U01HL130114 from the NHLBI with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629, R01AG15928, and R01AG20098 from the NIA. FHS research was supported by NHLBI contracts N01-HC-25195 and HHSN268201500001I. This study was also supported by additional grants from the NIA [R01s AG054076, AG049607, and AG033040, and NINDS (R01 NS017950)]. The ERF study as a part of EUROSPAN (European Special Populations Research Network) was supported by European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947) and also received funding from the European Community’s Seventh Framework Programme (FP7/2007-2013)/grant agreement HEALTH-F4-2007-201413 by the European Commission under the programme “Quality of Life and Management of the Living Resources” of 5th Framework Programme (no. QLG2-CT-2002-01254). High-throughput analysis of the ERF data was supported by a joint grant from the Netherlands Organisation for Scientific Research and the Russian Foundation

for Basic Research (NWO-RFBR 047.017.043). The Rotterdam Study is funded by Erasmus University Medical Center and Erasmus University Rotterdam, the Netherlands Organisation for Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the municipality of Rotterdam. Genetic datasets are also supported by the Netherlands Organisation for Scientific Research NWO Investments (175.010.2005.011, 911-03-012), the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), and the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO), Netherlands Consortium for Healthy Aging (NCHA), project 050-060-810. All studies are grateful to their individuals, faculty, and staff. The content of these manuscripts is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the U.S. Department of Health and Human Services. The four LSACs are: the Human Genome Sequencing Center at the Baylor College of Medicine (U54 HG003273), the Broad Institute Genome Center (U54HG003067), The American Genome Center at the Uniformed Services University of the Health Sciences (U01AG057659), and the Washington University Genome Institute (U54HG003079). Biological samples and associated phenotypic data used in primary data analyses were stored at Study Investigators institutions, and at the National Cell Repository for Alzheimer's Disease (NCRAD, U24AG021886) at Indiana University funded by NIA. Associated Phenotypic Data used in primary and secondary data analyses were provided by Study Investigators, the NIA funded Alzheimer's Disease Centers (ADCs), and the National Alzheimer's Coordinating Center (NACC, U01AG016976) and the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS, U24AG041689) at the University of Pennsylvania, funded by NIA, and at the Database for Genotypes and Phenotypes (dbGaP) funded by NIH. This research was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine. Contributors to the Genetic Analysis Data included Study Investigators on projects that were individually funded by NIA, and other NIH institutes, and by private United States organizations, or foreign governmental or non-governmental organizations.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.647436/full#supplementary-material>

## REFERENCES

- Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics* 26, 1340–1347. doi: 10.1093/bioinformatics/btq134
- Beecham, G. W., Bis, J. C., Martin, E. R., Choi, S.-H., DeStefano, A. L., van Duijn, C. M., et al. (2017). The Alzheimer's disease sequencing project:

**Supplementary Figure 1** | A multi-layer feed forward neural network was trained to classify individuals by cohort identity. The scaled conjugate gradient (SCG) algorithm was the primary learning algorithm used to minimize neural network weights, here and in all other applications of a neural network in this manuscript unless otherwise noted.

**Supplementary Figure 2** | ROC curves of NN performance were generated based on neural net output after being trained on individual cohort labels, and using individual genotypes (for the top 50 SNPs) as training features.

**Supplementary Figure 3** | For neural net architecture used to classify individuals as cases or controls see methods for **Supplementary Figure 1**.

**Supplementary Figure 4** | See legend.

**Supplementary Figure 5** | Distributions of **CVt** values are based on netSNP outputs (see **netSNP method** above). To determine if any **tSNPs** might confer protection from **APOE $\epsilon$ 4** effects, neural nets weights were fit using true training group genotypes; then the netSNP test was performed simultaneously for **APOE $\epsilon$ 4** and the **tSNP**, such that in each holdout group individual the **APOE** locus was set to  $\epsilon$ 4 and the target SNP locus were set to the homozygous minor alleles, and neural net output (**CVt**) was evaluated.

**Supplementary Figure 6** | The **CVt** distributions of holdout group individuals for different **tSNPs** were generated using the netSNP method (see **netSNP method** above); here **APOE** and **TOMM40** were excluded as training features.

**Supplementary Figure 7** | Same methods used to generate **Figure 4**, with the only modification being that **APOE** and **TOMM40** were excluded as training features, except Panel-A which necessarily includes **APOE** as a training feature. The left boxplot of Panel-E is the same as in **Figure 4** since the raw per-person count of known AD genes is independent of neural network output and netSNP manipulations.

**Supplementary Figure 8** | See text: results section "netSNP as Predictor of AD-Linked **tSNPs** and AD Pathophysiology." For cases, ages of BD diagnosis was assigned based on AD age distribution for analogous genotype [e.g., those with **SNP2** (**E4**-like) were given age of diagnosis with a distribution as individuals with **APOE $\epsilon$ 4** are diagnosed with AD]. For each BD, netSNP analysis was performed as it was for AD (**netSNP method**, above). Correlations between age of BD diagnosis and 12 of netSNP-identified SNPs and CVt-weighted 12 of netSNP-identified SNPs was conducted as for AD (**Figure 4**).

**Supplementary Figure 9** | Same methods used to generate **Figure 2D** using a classification threshold fixed at zero. NCASE, the number of individuals in the case condition; PCASE, the number of individuals predicted as case; NCTRL, the number of individuals in the control condition; PCTRL, the number of individuals predicted as control; TP, true positive; TN, true negative; FP, false positive; FN, false negative; TPR, true positive rate; TNR, true negative rate; FPR, false positive rate; FNR, false negative rate; PPV, positive predictive value; NPV, negative predictive value; FOR, false omission rate; FDR, false discovery rate; AUC, area under the curve; ACC, accuracy; FOS, F1-score; PLR, positive likelihood ratio; NLR, negative likelihood ratio; DOR, diagnostic odds ratio; MCC, Matthews correlation coefficient.

**Supplementary Figure 10** | Same methods used to generate **Figure 2D** using the OOP as the classification threshold. Same abbreviations as in **Supplementary Figure 9**.

**Supplementary Figure 11** | Definitions in **Supplementary Figure 9** shown in the form of a confusion matrix.

**Supplementary Figure 12** | See **Figure 1C**.

**Supplementary Figure 13** | See Figure legend.

study design and sample selection. *Neurol. Genet.* 3:e194. doi: 10.1212/nxg.0000000000000194

- Beecham, G. W., Vardarajan, B., Blue, E., Bush, W., Jaworski, J., Barral, S., et al. (2018). Rare genetic variation implicated in non-Hispanic white families with Alzheimer disease. *Neurol. Genet.* 4:e286. doi: 10.1212/nxg.0000000000000286
- Bis, J. C., Jian, X., Kunkle, B. W., Chen, Y., Hamilton-Nelson, K. L., Bush, W. S., et al. (2018). Whole exome sequencing study identifies novel rare and

- common Alzheimer's-associated variants involved in immune response and transcriptional regulation. *Mol. Psychiatr.* 25, 1–17. doi: 10.1038/s41380-018-0112-7
- Braak, H., Alafuzoff, I., Arzberger, T., Kretschmar, H., and Tredici, K. D. (2006). Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. *Acta neuropathol.* 112, 389–404. doi: 10.1007/s00401-006-0127-z
- Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., et al. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* 47, 291–295. doi: 10.1038/ng.3211
- Carmona, S., Hardy, J., and Guerreiro, R. (2018). The genetic landscape of Alzheimer disease. *Handb. Clin. Neurol.* 148, 395–408. doi: 10.1016/b978-0-444-64076-5.00026-0
- Cauwenberghe, C. V., Broeckhoven, C. V., and Sleegers, K. (2015). The genetic landscape of Alzheimer disease: clinical implications and perspectives. *Genet. Med.* 18, 421–430. doi: 10.1038/gim.2015.117
- Corder, E., Saunders, A., Strittmatter, W., Schmechel, D., Gaskell, P., Small, G., et al. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* 261, 921–923. doi: 10.1126/science.8346443
- Crane, P. K., Foroud, T., Montine, T. J., and Larson, E. B. (2017). Alzheimer's disease sequencing project discovery and replication criteria for cases and controls: data from a community-based prospective cohort study with autopsy follow-up. *Alzheimers Dement.* 13, 1410–1413. doi: 10.1016/j.jalz.2017.09.010
- Demuth, H. B., Beale, M. H., Jess, O. D., and Hagan, M. T. (2014). *Neural Network Design*. Available online at: <https://dl.acm.org/doi/book/10.5555/2721661> (accessed February 24, 2021).
- Desikan, R. S., Fan, C. C., Wang, Y., Schork, A. J., Cabral, H. J., Cupples, L. A., et al. (2017). Genetic assessment of age-associated Alzheimer disease risk: development and validation of a polygenic hazard score. *PLoS Med.* 14:e1002258. doi: 10.1371/journal.pmed.1002258
- DeTure, M. A., and Dickson, D. W. (2019). The neuropathological diagnosis of Alzheimer's disease. *Mol. Neurodegener.* 14:32. doi: 10.1186/s13024-019-0333-5
- Escott-Price, V., Myers, A. J., Huentelman, M., and Hardy, J. (2017). Polygenic risk score analysis of pathologically confirmed Alzheimer disease. *Ann. Neurol.* 82, 311–314. doi: 10.1002/ana.24999
- Escott-Price, V., Sims, R., Bannister, C., Harold, D., Vronskaya, M., Majounie, E., et al. (2015). Common polygenic variation enhances risk prediction for Alzheimer's disease. *Brain* 138, 3673–3684. doi: 10.1093/brain/awv268
- Ferrero, H., Larrayoz, I. M., Gil-Bea, F. J., Martínez, A., and Ramírez, M. J. (2018). Adrenomedullin, a novel target for neurodegenerative diseases. *Mol. Neurobiol.* 55, 8799–8814. doi: 10.1007/s12035-018-1031-y
- Ferrero, H., Larrayoz, I. M., Martisova, E., Solas, M., Howlett, D. R., Francis, P. T., et al. (2017). Increased levels of brain adrenomedullin in the neuropathology of Alzheimer's disease. *Mol. Neurobiol.* 55, 5177–5183. doi: 10.1007/s12035-017-0700-6
- Gatz, M., Reynolds, C. A., Fratiglioni, L., Johansson, B., Mortimer, J. A., Berg, S., et al. (2006). Role of genes and environments for explaining Alzheimer disease. *Arch. Gen. Psychiatr.* 63:168. doi: 10.1001/archpsyc.63.2.168
- Goldstein, B. A., Polley, E. C., and Briggs, F. B. S. (2011). Random forests for genetic association studies. *Stat. Appl. Genet. Mol.* 10:32. doi: 10.2202/1544-6115.1691
- Guerreiro, R. J., and Hardy, J. (2012). TOMM40 association with Alzheimer disease: tales of APOE and linkage disequilibrium. *Arch. Neurol.* 69, 1243–1244. doi: 10.1001/archneurol.2012.1935
- Gustaw-Rothenberg, K., Lerner, A., Bonda, D. J., Lee, H., Zhu, X., Perry, G., et al. (2010). Biomarkers in Alzheimer's disease: past, present and future. *Biomark. Med.* 4, 15–26. doi: 10.2217/bmm.09.86
- Hampel, H., O'Bryant, S. E., Molinuevo, J. L., Zetterberg, H., Masters, C. L., Lista, S., et al. (2018). Blood-based biomarkers for Alzheimer disease: mapping the road to the clinic. *Nat. Rev. Neurol.* 14, 639–652. doi: 10.1038/s41582-018-0079-7
- Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M. L., et al. (2009). Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat. Genet.* 41, 1088–1093. doi: 10.1038/ng.440
- Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., et al. (2018). NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimers Dement.* 14, 535–562. doi: 10.1016/j.jalz.2018.02.018
- Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., et al. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* 51, 404–413. doi: 10.1038/s41588-018-0311-9
- Jeyifous, O., Lin, E. I., Chen, X., Antinone, S. E., Mastro, R., Drisdell, R., et al. (2016). Palmitoylation regulates glutamate receptor distributions in postsynaptic densities through control of PSD95 conformation and orientation. *Proc. Natl. Acad. Sci. U.S.A.* 113, E8482–E8491. doi: 10.1073/pnas.1612963113
- Jolliffe, I. T. (1986). *Principal Component Analysis*. Springer Series in Statistics. New York, NY: Springer, 115–128.
- Karch, C. M., and Goate, A. M. (2014). Alzheimer's disease risk genes and mechanisms of disease pathogenesis. *Biol. Psychiatr.* 77, 43–51. doi: 10.1016/j.biopsych.2014.05.006
- Koen, J. D., Barrett, F. S., Harlow, I. M., and Yonelinas, A. P. (2016). The ROC toolbox: a toolbox for analyzing receiver-operating characteristics derived from confidence ratings. *Behav. Res. Methods* 49, 1399–1406. doi: 10.3758/s13428-016-0796-z
- Koffie, R. M., Hashimoto, T., Tai, H.-C., Kay, K. R., Serrano-Pozo, A., Joyner, D., et al. (2012). Apolipoprotein E4 effects in Alzheimer's disease are mediated by synaptotoxic oligomeric amyloid- $\beta$ . *Brain J. Neurol.* 135, 2155–2168. doi: 10.1093/brain/aws127
- Kunkle, B. W., Grenier-Boley, B., Sims, R., Bis, J. C., Damotte, V., Naj, A. C., et al. (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates A $\beta$ , tau, immunity and lipid processing. *Nat. Genet.* 51, 414–430. doi: 10.1038/s41588-019-0358-2
- Lambert, J.-C., Heath, S., Even, G., Campion, D., Sleegers, K., Hiltunen, M., et al. (2009). Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat. Genet.* 41, 1094–1099. doi: 10.1038/ng.439
- Lander, E., and Schork, N. (1994). Genetic dissection of complex traits. *Science* 265, 2037–2048. doi: 10.1126/science.8091226
- Larrayoz, I. M., Ferrero, H., Martisova, E., Gil-Bea, F. J., Ramírez, M. J., and Martínez, A. (2017). Adrenomedullin contributes to age-related memory loss in mice and is elevated in aging human brains. *Front. Mol. Neurosci.* 10:384. doi: 10.3389/fnmol.2017.00384
- Ma, Y., Jun, G. R., Zhang, X., Chung, J., Naj, A. C., Chen, Y., et al. (2019). Analysis of Whole-exome sequencing data for Alzheimer disease stratified by APOE genotype. *JAMA Neurol.* 76, 1099–1108. doi: 10.1001/jamaneurol.2019.1456
- Mars, N., Koskela, J. T., Ripatti, P., Kiiskinen, T. T. J., Havulinna, A. S., Lindbohm, J. V., et al. (2020). Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat. Med.* 26, 549–557. doi: 10.1038/s41591-020-0800-0
- Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* 51, 584–591. doi: 10.1038/s41588-019-0379-x
- Mathworks (2020a). *Multilayer Shallow Neural Network Architecture*. Mathworks: Natick, MA.
- Mathworks (2020b). *Receiver Operating Characteristic (ROC) Curve or Other Performance Curve for Classifier Output*. Mathworks: Natick, MA.
- Mendez, M. F. (2017). Early-Onset Alzheimer disease. *Neurol. Clin.* 35, 263–281. doi: 10.1016/j.ncl.2017.01.005
- Moller, M. F. (1993). A scaled conjugate-gradient algorithm for fast supervised learning. *Neural Netw.* 6, 525–533. doi: 10.1016/s0893-6080(05)80056-5
- Molnar, C. (2021). *Interpretable Machine Learning*. Available online at: <https://www.lulu.com/shop/christoph-molnar/interpretable-machine-learning/paperback/product-24036234.html> (accessed February 24, 2021).
- Naj, A. C., Lin, H., Vardarajan, B. N., White, S., Lancour, D., Ma, Y., et al. (2018). Quality control and integration of genotypes from two calling pipelines for whole genome sequence data in the Alzheimer's disease sequencing project. *Genomics* 111, 808–818. doi: 10.1016/j.ygeno.2018.05.004
- Patel, D., Mez, J., Vardarajan, B. N., Staley, L., Chung, J., Zhang, X., et al. (2019). Association of rare coding mutations with Alzheimer disease and other dementias among adults of European ancestry. *JAMA Netw. Open* 2:e191350. doi: 10.1001/jamanetworkopen.2019.1350
- Penner, G., Lecocq, S., Chopin, A., Vedoya, X., Lista, S., Vergallo, A., et al. (2019). Blood-based diagnostics of Alzheimer's disease. *Expert Rev. Mol. Diagn.* 19, 613–621. doi: 10.1080/14737159.2019.1626719
- Raghavan, N. S., Brickman, A. M., Andrews, H., Manly, J. J., Schupf, N., Lantigua, R., et al. (2018). Whole-exome sequencing in 20,197 persons for rare variants in Alzheimer's disease. *Ann. Clin. Trans. Neurol.* 5, 832–842.

- Saunders, A. M., Strittmatter, W. J., Schmechel, D., George-Hyslop, P. H. S., Pericak-Vance, M. A., Joo, S. H., et al. (1993). Association of apolipoprotein E allele 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology* 43, 1467–1472. doi: 10.1212/wnl.43.8.1467
- Sejnowski, T. J. (2020). The unreasonable effectiveness of deep learning in artificial intelligence. *Proc. Natl. Acad. Sci. U.S.A.* 117:201907373. doi: 10.1073/pnas.1907373117
- Selzam, S., Coleman, J. R. I., Caspi, A., Moffitt, T. E., and Plomin, R. (2018). A polygenic p factor for major psychiatric disorders. *Transl. Psychiat.* 8:205. doi: 10.1038/s41398-018-0217-4
- Shi, Y., Yamada, K., Liddel, S. A., Smith, S. T., Zhao, L., Luo, W., et al. (2017). ApoE4 markedly exacerbates tau-mediated neurodegeneration in a mouse model of tauopathy. *Nature* 549, 523–527. doi: 10.1038/nature24016
- Sisodia, S. S., and George-Hyslop, P. H. S. (2002).  $\gamma$ -Secretase, notch, A $\beta$  and alzheimer's disease: where do the presenilins fit in? *Nat. Rev. Neurosci.* 3, 281–290. doi: 10.1038/nrn785
- Slatkin, M. (2007). Inbreeding coefficients and coalescence times. *Genet. Res.* 89, 479–487. doi: 10.1017/s0016672308009671
- Yokoi, N., Fukata, Y., Sekiya, A., Murakami, T., Kobayashi, K., and Fukata, M. (2016). Identification of PSD-95 depalmitoylating enzymes. *J. Neurosci.* 36, 6431–6444. doi: 10.1523/jneurosci.0419-16.2016
- Yu, C.-E., Seltman, H., Peskind, E. R., Galloway, N., Zhou, P. X., Rosenthal, E., et al. (2007). Comprehensive analysis of APOE and selected proximate markers for late-onset Alzheimer's disease: patterns of linkage disequilibrium and disease/marker association. *Genomics* 89, 655–665. doi: 10.1016/j.ygeno.2007.02.002
- Zhang, Q., Sidorenko, J., Couvy-Duchesne, B., Marioni, R. E., Wright, M. J., Goate, A. M., et al. (2020). Risk prediction of late-onset Alzheimer's disease implies an oligogenic architecture. *Nat. Commun.* 11:4799. doi: 10.1038/s41467-020-18534-1
- Zhang, X., Zhu, C., Beecham, G., Vardarajan, B. N., Lancour, D., Farrell, J. J., et al. (2019). A rare missense variant of CASP7 is associated with familial late-onset Alzheimer's disease. *Alzheimers Dement.* 15, 441–452. doi: 10.1016/j.jalz.2018.10.005

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Monk, Rajkovic, Petrus, Rajkovic, Gaasterland and Malinow. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.