OXFORD

Briefings in Bioinformatics, 2023, **24(2)**, 1–12
https://doi.org/10.1093/bib/bbad030
Advance access publication date 8 February 2023
Problem Solving Protocol

AD-Syn-Net: systematic identification of Alzheimer's disease-associated mutation and co-mutation vulnerabilities via deep learning

Xingxin Pan, Zeynep H. Coban Akdemir, Ruixuan Gao, Xiaoqian Jiang, Gloria M. Sheynkman, Erxi Wu, Jason H. Huang, Nidhi Sahni and S. Stephen Yi

Corresponding authors: S.S. Yi. E-mail: stephen.yi@austin.utexas.edu; N. Sahni. E-mail: nsahni@mdanderson.org; J.H. Huang. E-mail: jason.huang@bswhealth.org; E. Wu. E-mail: erxi.wu@bswhealth.org

Abstract

Alzheimer's disease (AD) is one of the most challenging neurodegenerative diseases because of its complicated and progressive mechanisms, and multiple risk factors. Increasing research evidence demonstrates that genetics may be a key factor responsible for the occurrence of the disease. Although previous reports identified quite a few AD-associated genes, they were mostly limited owing to patient sample size and selection bias. There is a lack of comprehensive research aimed to identify AD-associated risk mutations systematically. To address this challenge, we hereby construct a large-scale AD mutation and co-mutation framework ('AD-Syn-Net'), and propose deep learning models named Deep-SMCI and Deep-CMCI configured with fully connected layers that are capable of predicting cognitive impairment of subjects effectively based on genetic mutation and co-mutation profiles. Next, we apply the customized frameworks to data sets to evaluate the importance scores of the mutations and identified mutation effectors and co-mutation combination vulnerabilities contributing to cognitive impairment. Furthermore, we evaluate the influence of mutation pairs on the network architecture to dissect the genetic organization of AD and identify novel co-mutations that could be responsible for dementia, laying a solid foundation for proposing future targeted therapy for AD precision medicine. Our deep learning model codes are available open access here: https://github.com/Pan-Bio/AD-mutation-effectors.

Keywords: Alzheimer's disease, network models, mutations and co-mutations, genetic interactions, deep learning

Introduction

As a neurodegenerative disease with amyloid beta peptide, neuritic plaques and hyperphosphorylated tau protein deposited in Alzheimer's Disease (AD) brain tissues, AD is the key driver that causes dementia in the elderly, resulting in progressive damage to the brain and finally even death [1–3]. With increasing incidences each year, AD is rising to the sixth leading cause of death in the United States, and the number of AD dementia cases will grow to more than 100 million in 2050 if effective treatment is not found [4]. Although the biology of AD is gradually understood, currently the exact etiology of AD is still not illuminated, because the development of AD comprises a complicated plethora of progressive, interactive, devastating processes [5]. Evidence from family and twin studies shows that genetic factors may play a vital role and account for at least 80% of AD cases, indicating high heritability of the development and progression of AD [6–8]. Decade-long research works have sought to identify the molecular genetic mechanism of AD and have found multiple autosomal pathogenic germline mutations such as mutations in *APP*, *PSEN1* and *PSEN2* associated with the early onset form of AD (EOAD) and mutations in *APOE* and *TREM2* associated with the late-onset form of AD (LOAD) [9–13]. However, known mutations in these genes can only explain a small fraction of AD cases. For example, rare mutations in *APP*, *PSEN1* and *PSEN2* only account for 5–10% of EOAD cases, leaving a large number of genetically unexplained AD cases, especially for LOAD where there is a more complex genetic etiology and less explained genetic component [7, 14–19]. Therefore, understanding the genetics and the mechanism of AD would benefit early detection, prevention and treatment of AD. The community currently is in urgent need of illuminating the

Xingxin Pan is a research engineering scientist associate at Dell Medical School, the University of Texas at Austin.
Zeynep H. Coban Akdemir is an assistant professor at the University of Texas Health Science Center at Houston.
Ruixuan Gao is an assistant professor at the University of Illinois Chicago.
Xiaoqian Jiang is Associate Vice President of MedicalAI, Center Director of Secure Artificial Intelligence for Health care (SAFE), at the University of Texas Health Science Center at Houston.
Gloria M. Sheynkman is an assistant professor at the University of Virginia.
Erxi Wu is a professor and Associate Director in Neuroscience Institute at Baylor Scott & White Health.
Jason H. Huang is Chairman of Department of Neurosurgery at Baylor Scott White, and a professor of Neurosurgery at Baylor College of Medicine.
Nidhi Sahni is an associate professor at the Department of Epigenetics and Molecular Carcinogenesis at MD Anderson Cancer Center, and a regular faculty member at the QCB graduate program at Baylor College of Medicine.
S. Stephen Yi is Director of Bioinformatics at Dell Medical School, and a faculty member at Oden Institute for Computational Engineering & Sciences and Department of Biomedical Engineering, University of Texas at Austin.
Received: July 9, 2022. **Revised:** December 19, 2022. **Accepted:** January 13, 2023

genetic cause of AD development and progression [20].

With genome-wide association and meta-analysis, multiple susceptibility loci were identified related to LOAD, including *CLU*, *PICALM*, *CR1*, *BIN1* and *TOMM40* genes [21–24]. Taking advantage of linkage analyses, additional AD-associated risk genes were discovered, including *DAPK1* and *UQBLN1* on chromosome 9, *IDE* and *TFAM* on chromosome 10, *GAB2* and *SORL1* on chromosome 11 [25–30]. Using next-generation sequencing technology, researchers identified *TYROBP* and *NOTCH3* as candidates for causing AD [31–34]. By using the approximate Firth regression test for unbalanced case–control phenotypes, dozens of more genes were identified to be associated with cognitive disorders in AD, including *AMPD3*, *GBE1* and *PLD1* [35, 36].

Although the abovementioned designs and methods identified quite a few AD-associated genes and mutations successfully, there were several notable limitations to these studies. For example, most of the association analyses were carried out in a limited small number of subjects and usually were restricted to familial symptomatic and asymptomatic individuals, making identifying liable genes challenging because of sampling and diagnostic issues [25, 29, 30, 37]; a considerable portion of AD-associated genes identified by GWAS and linkage could not be confirmed in independent research as well [38–40]. A gene associated with AD only reflects a correlation relationship with AD, but does not imply that the gene is a definite deterministic risk factor involved in AD development and progression [25, 29, 30]. Also sometimes, a specific mutation or variant of AD-associated genes could not be identified owing to different populations [29, 41–43]. More importantly, most of the previous analyses did not take account into the interaction between genes, which is found more important in the development and progression of AD [44–47]. Besides, lacking a unified and comprehensive metric measuring cognition impairment may be another reason why identified AD-associated variants cannot be consistent across independent research works [40, 48, 49].

Therefore, novel integrative computational frameworks are necessary to more accurately evaluate AD variants. In addition, our previous work indicates that protein products of mutated disease genes do not function in isolation, but are part of highly interconnected signaling networks [50–53]. With systematic whole-exome sequencing approaches, AD patients are expected to harbor many more disease-liable variants [54–56]. Even in the same AD patient, several interactive AD-associated gene mutations are often at play [57–60]. Our conceptual framework is to systematically link AD ('disease phenome') [61] with a complete list of AD susceptibility mutations ('disease genome') [62–65], resulting in a global view of the 'AD diseasome'. With the advancement of artificial intelligence and high-performance computing, deep learning technologies are becoming increasingly active in biology and genome science, and achieving substantial success in a variety of biological domains, including vaccine target identification, cancer subtyping, cancer target identification, functional genomics and even drug discovery [66–74]. For example, researchers have developed a robust prediction model with a graph convolutional network to identify novel drug-target interactions, which accelerates drug development [72]. The application of deep learning in genetic data would be promising in identifying AD susceptibility mutations and illuminating the mechanism of AD development and progression.

Here we propose a deep learning-based framework to address the abovementioned problems, and meanwhile, the framework we propose presents advantages compared with previous work. To our best knowledge, this is one of the first studies to incorporate unified large-scale AD samples and data sets, taking into account population stratification, sample size and diagnostic issues. Owing to the large sample size in our study, we, therefore, could integrate a comprehensive genetic mutation landscape, and this is beneficial to measure the effect of each mutation systematically and unbiasedly on patient populations. More importantly, our framework trained on the large-scale samples could evaluate the effects of mutations on AD development and progression and even classify these mutations into two groups, including AD-promoting mutations and AD-suppressive mutations. Considering the combination of a gene such as *APOE* with another AD-associated gene such as *GAB2* or *SORL1* becomes a more useful predictor of AD risk based on previous reports [28], our study could probably become the first to integrate comprehensive mutation pairs and evaluate the effects of these co-mutation pairs systematically in gene interaction networks.

In this paper, we first introduce the data sets we collect from the Alzheimer Disease Neuroimaging Initiative (ADNI) and the processing pipeline of the data sets. To perform an unbiased predictive performance evaluation for our models, namely, Deep-SMCI (single mutation for cognitive impairment) and Deep-CMCI (co-mutation for cognitive impairment), we devise and adopt multiple baselines and metrics to compare with them utilizing cross-validation. Afterwards, we detail the framework of Deep-SMCI and Deep-CMCI and the optimization processes based on Bayesian applied to the model training. After model optimization, we show the robust predictive performance of Deep-SMCI and Deep-CMCI compared with control models. Next, we apply the optimized Deep-SMCI and Deep-CMCI to the data sets and rank single mutations and co-mutations based on their effect size on the cognitive impairment using permutation tests, and further identify AD-promoting and AD suppressive effector mutations based on statistical analysis. To furthermore extend our findings, we also apply mutual information (MI) and Lime to the data sets, by comparing the results from these two methods. In addition to the overlapping effectors, novel effectors are found as well, further expanding our mutation findings [75]. Based on these AD-promoting effectors and suppressors (mutations and mutation pairs), we would not only gain new insights into the mechanisms of action underlying AD progression but also provide potential novel diagnostic and therapeutic targets to treat cognitive impairment in AD.

Materials and methods

Data sets

Genetic mutation data sets (Illumina Omni 2.5 M, WGS platform) were collected from the ADNI, and these data sets covered subjects with available cognitive evaluation reports that were recruited from multiple cohorts, including ADNI-1–3 and ADNI-GO. Detailed information on the data acquisition and processing of ADNI WGS could be found on the website (https://adni.loni.usc.edu/data-samples/data-types/). For characterizing the extent of cognitive dysfunction for subjects, the Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS-Cog) was obtained from ADNI. More specifically, we adopted ADAS-Cog-11 to measure cognitive impairment for subjects. For consistency, we kept the ADAS-Cog-11 test scores of subjects at the baseline visit code in the ADNI project. Finally, we filtered out and obtained 794 subjects with available WGS and ADAS-Cog-11 data sets at baseline time, including 275 ADNI-1, 395 ADNI-2 and 124 ADNI-GO subjects. The detailed clinical information of our AD subjects is shown in Supplementary Table 1.

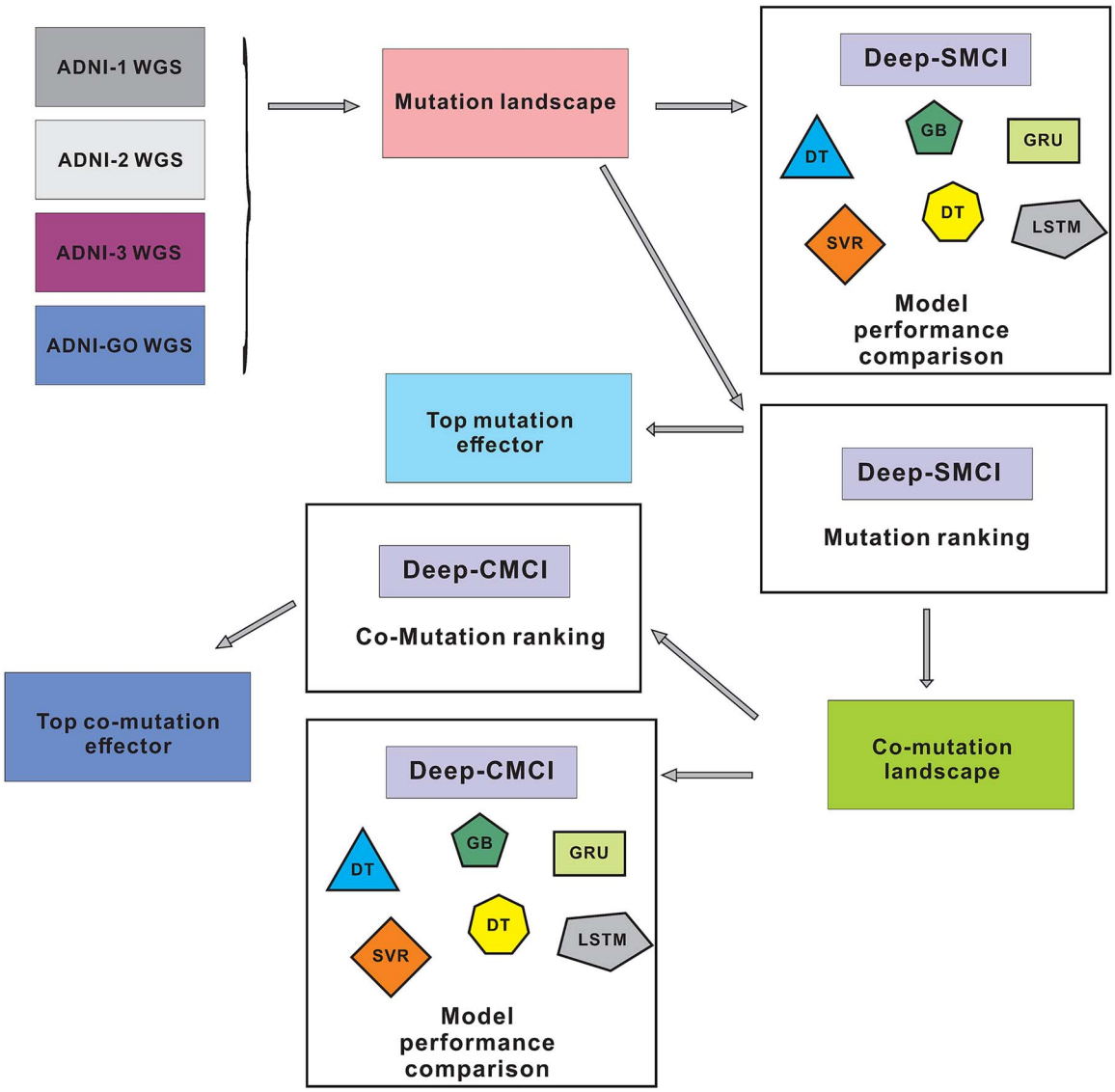

Figure 1. The workflow of the cognitive impairment prediction and effector identification.

Data processing

Based on mutation profiles in the VCF format obtained from ADNI, we merged all mutation features across 794 subjects and generated a huge matrix where each row represents a subject and each column represents a mutation feature. To facilitate the cognitive impairment prediction and AD mutation effectors identification, we kept the mutations arising from those proven AD-associated genes, including 307 GenAge genes and 550 LongevityMap genes, and finally got 2 671 687 unique mutation landscapes [76, 77]. If one subject carries one mutation, then the entry value for the subject would be 1, otherwise, it would be 0.

Predictive performance evaluation

To carry out a reliable performance evaluation, we devised multiple traditional machine learning models as baselines, including decision tree (DT), gradient boosting (GB), elastic net (EN) and support vector regression (SVR) (Figure 1). We carried out a grid search scheme for optimizing these four baseline models. In addition, we constructed two deep learning models based on gated recurrent units (GRU) and long short-term memory (LSTM), which are known to be effective and robust in biological sequence data, and optimized these two models based on Bayesian optimization and then compared their predictive performance with other methods in the aspect of mean absolute error (MAE) and root mean squared error (RMSE) [78, 79]. We adopted repeats 10-fold cross-validation five times with different randomization in each repetition, making the evaluation more unbiased. Here, the definitions of MAE and RMSE for evaluating the predictive performance are as follows:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} | y_i - \hat{y}_i |,$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2},$$

where N is the number of samples tests; y_i is the ADAS-Cog-11 score for sample i; $\hat{y}_i$ is the predicted ADAS-Cog-11 score for sample i.

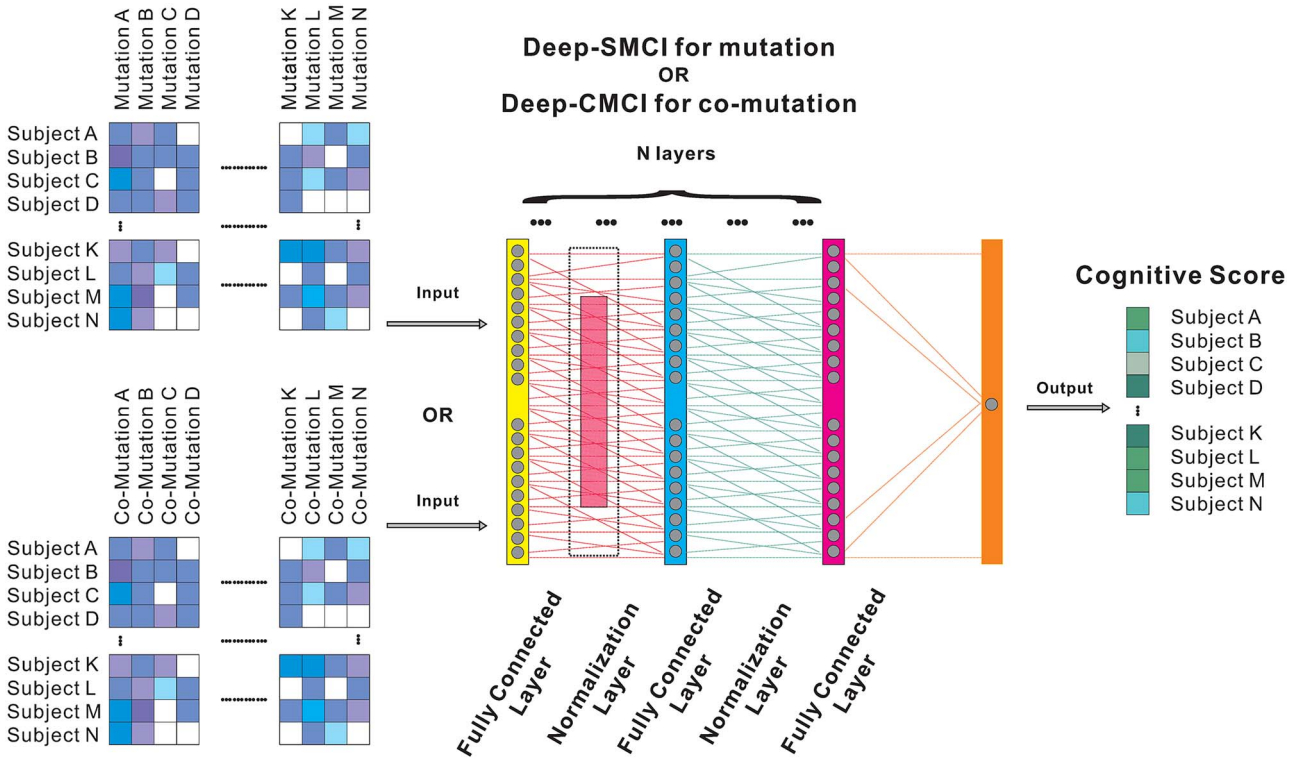

Figure 2. The deep learning model for cognitive impairment prediction based on mutations and co-mutations.

Construction of deep-SMCI and deep-CMCI for cognitive impairment predictions

As shown in Figure 2, both Deep-SMCI and Deep-CMCI consist of a multiple-layer neural network. The input for this model was the 1-D dimensional mutation profiles for each subject, and the output was the ADAS-Cog-11 for the subject. Basically, the model was constructed by a fully connected (FC) network. In the network, a hidden unit j in the FC layer k takes the sum of the weighted outputs plus the bias from the previous layer $k-1$ as the input and generates an output o_j^k:

$$o_j^k = f\left(\sum_{i=1}^{U} w_{i,j}^{k-1} o_i^{k-1} + b_j^{k-1}\right),$$

where U is the number of hidden neural units; $\left\{w_{i,j}^{k-1}, b_j^{k-1}\right\}_{i=i=1}^{U}$ represents the weights and the bias term of unit j in the layer k to be optimized; and f is a nonlinear activation function.

To boost the predictive efficacy, we placed a batch normalization layer between FC layers, which could significantly reduce the problem of coordinating updates across layers based on reparameterization [80]. After being transformed multiple times, the input information was fed into the output layer consisting of one unit to predict ADAS-Cog-11 of the input subject. We took advantage of multiple optimization techniques to facilitate model training and make the model optimize effectively. We applied dropout to the intermediate layer between FC layers, which could average multiple layers and avoid overfitting during the model training [81]. Early stopping was used to regularize the model training, and we set the following criteria: monitor aim is validation loss and training patience is five epochs [82]. Lastly, to achieve a better local minimum of parameters efficiently, Adam method was adopted for gradient descent, which was proved to be invariant to diagonal rescaling of the gradients and well suited for the ADAS-Cog-11 prediction where there was high-dimensional input data [83]. The loss function for training Deep-SMCI and Deep-CMCI was the mean squared error, namely,

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2,$$

where N was the number of samples; y_i was the ADAS-Cog-11 score for sample i; $\hat{y}_i$ was the predicted ADAS-Cog-11 score for sample i.

Deep-SMCI and deep-CMCI optimization

Considering the input of Deep-SMCI and Deep-CMCI was high-dimensional features, we took into account the hyperparameters well suited for cognitive score prediction. Therefore, we defined an enormous hyperparameter space to tune the model: the number of FC layers ranged from 2 to 8 with a step size is 1; the number of hidden units for each FC layer varied from 16 to 256 in a double-fold way; a group of dropout rates was evaluated, ranging from 0 to 0.5, with a step size of 0.1; the batch size varied from 8 to 64, with a 2-fold way; L1 and L2 regularization penalty ranged from 1e−5 to 1e−1 in the output FC layers with a 10-fold step size; the initialized value of each parameter in the FC layers was sampled randomly from the set of ones, random uniform and random normal distribution; each FC layer applies the exponential linear unit, rectified linear activation or hyperbolic tangent (Tanh) randomly.

To facilitate the process of hyperparameter search in the high-dimensional space, we adopted Bayesian optimization during model training, which could automatically explore hyperparameter combinations wisely and at a low computation cost [39]. Here, we defined 500 initiation points in the hyperparameters space and a maximum of 10 iterations for models with the number of FC layers from 2 to 8, respectively, when setting up Bayesian optimization. During the training process, as iteration grew, and the posterior distribution of the model's cost function improved, the Bayesian optimization algorithm could

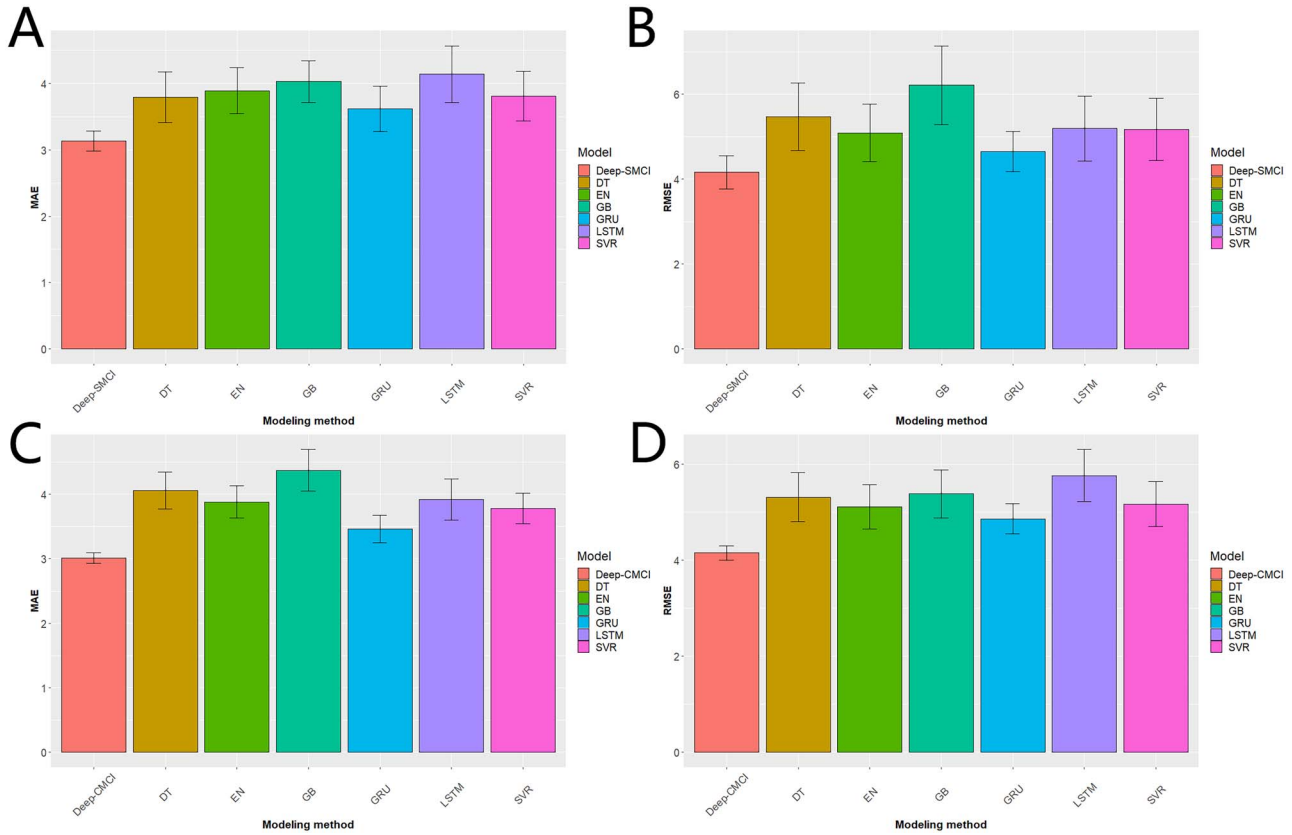

Figure 3. Evaluation of predictive performance for cognitive impairment prediction. (**A**) MAE predictive performance comparison for Deep-SMCI and baselines based on mutation landscape. (**B**) RMSE predictive performance comparison for Deep-SMCI and baselines based on mutation landscape. (**C**) MAE predictive performance comparison for Deep-CMCI and baselines based on co-mutation landscape. (**D**) RMSE predictive performance comparison for Deep-CMCI and baselines based on co-mutation landscape.

further explore hyperparameter space that was worth exploring automatically and sought the best hyperparameter combination efficiently. For each iteration, it took ~40 h to finish the model training and evaluation on AMD EPYC 7642 48-core processor. We finally chose the models constructed with five FC layers as basic components for Deep-SMCI and Deep-CMCI based on performance comparison with models with different numbers of FC layers in the aspect of MAE and RMSE. Deep-SMCI and Deep-CMCI were implemented based on TensorFlow2 and Keras [84, 85]. Model evaluation optimization was achieved by scikit-learn and Bayesian Optimization Python library [86, 87].

Ranking of AD effectors

To identify AD promoters and suppressors, including mutations and mutation pairs, we adopted the permutation test method based on ELI5 to compute feature importance for the prediction model by measuring how scores decreased when a mutation or mutation pair was not available. This method randomly shuffled each input feature and computed the change in the model's predictive performance. Therefore, we could get AD-promoting and suppressive mutations or co-mutations that were related to cognitive impairment. To further expand our findings, we also applied MI strategy with scikit-learn and Lime to rank mutation importance scores [86, 88].

Results

Deep-SMCI outperforms existing algorithms based on mutatome landscapes

By taking advantage of Bayesian optimization, we got customized and optimized Deep-SMCI for predicting the cognitive impairment of subjects based on mutation landscapes. To evaluate the predictive performance of the Deep-SMCI, we compared the model to the abovementioned control algorithms, including DT, GB, SVR and EN after grid search optimization and GRU, LSTM after Bayesian optimization in the aspect of MAE and RMSE based on 10-fold cross-validation with five repeats from randomization. The detailed hyperparameter space of baseline modeling methods and optimized hyperparameter configurations of Deep-SMCI based on mutations are shown in Supplementary Table 2.

As shown in Figure 3A, Deep-SMCI achieved the best MAE and the least standard deviation (SD) of MAE compared with the other four models. Deep-SMCI achieved MAE at 3.13 ± 0.151 (SD) when predicting ADAS-Cog-11 based on mutation landscape; however, SVR got 3.81 ± 0.371 (SD), EN got 3.89 ± 0.347 (SD), DT got 3.79 ± 0.382 (SD), GB got 4.03 ± 0.314 (SD), GRU got 3.62 ± 0.341 (SD) and LSTM got 4.14 ± 0.428 (SD). Deep-SMCI not only predicted ADAS-Cog-11 with a less MAE but also performs stable performance. To note, although GRU and LSTM were applied to sequence and time series data successfully, they did not perform well in the task. The reason might be our mutation landscape was not 'real' sequence data, and the order of the mutations was not important, making Deep-SMCI with FC networks a better candidate compared with the models with GRU and LSTM layers [78, 79].

Furthermore, Deep-SMCI performed well in the aspect of RMSE. Deep-SMCI achieved the best RMSE and the least SD of RMSE compared with the other four models (Figure 3B). For example, Deep-SMCI achieved RMSE at 4.16 ± 0.389 (SD) when predicting ADAS-Cog-11 based on mutation landscape; however, SVR got 5.17 ± 0.730 (SD), EN got 5.09 ± 0.672 (SD), DT got 5.47 ± 0.793 (SD), GB got 6.21 ± 0.931 (SD), GRU got 4.65 ± 0.470 (SD) and LSTM

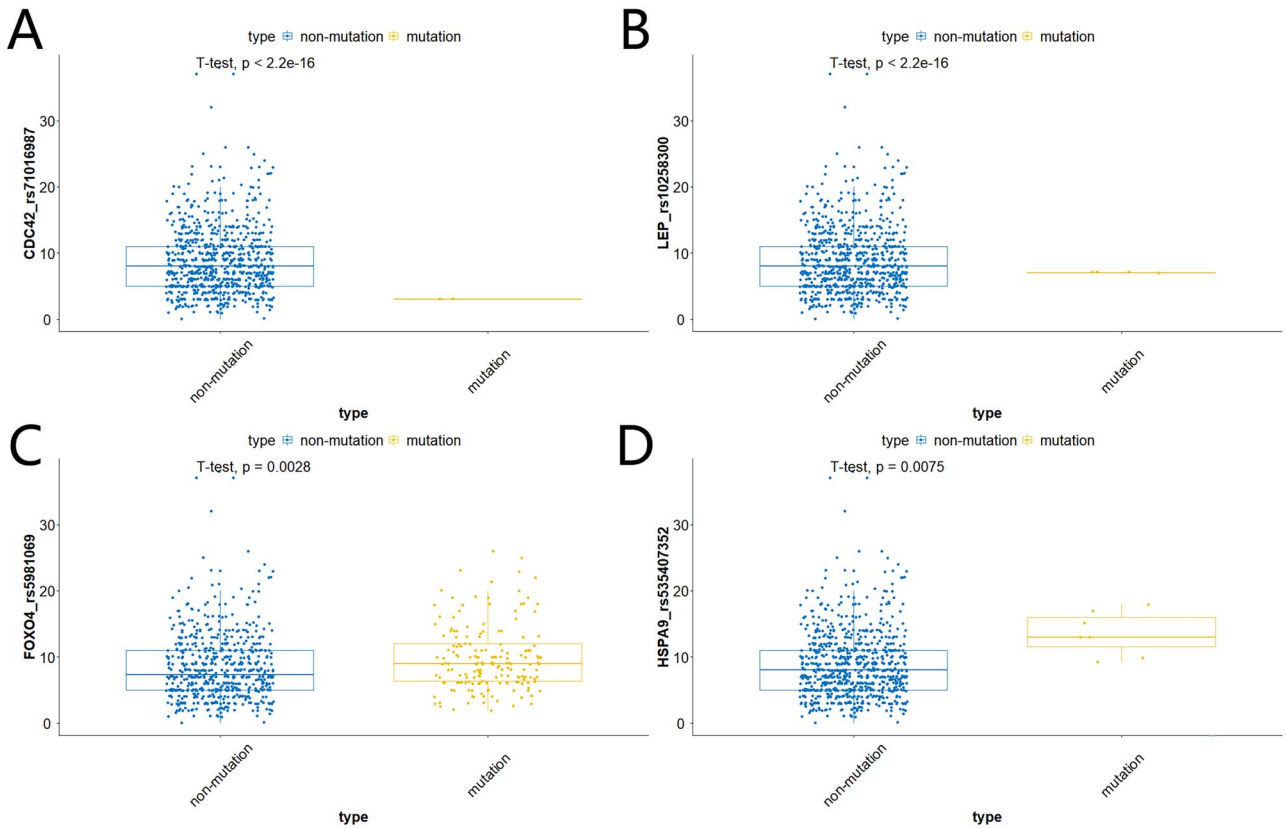

Figure 4. Examples of mutation effectors in cognitive impairment. (**A**) *CDC42* rs71016987 is an AD-suppressive mutation (*t*-test, *P*-value < 2.2e−16). (**B**) *LEP* rs10258300 is an AD-suppressive mutation (*t*-test, *P*-value < 2.2e−16). (C) *FOXO4* rs5981069 is an AD-promoting mutation (*t*-test, *P*-value = 0.0028). (**D**) *HSPA9* rs535407352 is an AD-promoting mutation (*t*-test, *P*-value = 0.0075).

got 5.19 ± 0.767 (SD). All these evaluations suggest the Bayesian optimization managed to optimize the Deep-SMCI to the maximum extent and made the model perform best, even if the control algorithms performed reasonably well after being optimized. The detailed predictive performance in this part is shown in Supplementary Table 3.

Deep-SMCI ranks AD-associated mutations based on custom effect scores

Considering that Deep-SMCI achieved prospective MAE and RMSE compared with baselines, it was reasonable to apply the model to dive into the mutations that might contribute to cognitive impairment. Although there were high-dimension mutation profiles (2 671 687) in the subjects, it was critical to know which mutations might promote cognitive impairment and which mutations might suppress cognitive impairment. To find these potential mutation effectors, we applied the customized Deep-SMCI to the mutation landscapes. Based on the permutation importance module in ELI5, the Deep-SMCI computed the contribution of each mutation to cognitive impairment and evaluated the mutations based on their effect importance.

Of 2 671 687 unique mutations across the subjects, 1 353 199 (50.65%) mutations were assigned to 0 importance score, meaning these mutations could not be directly relevant to the cognitive performance of the subjects. For example, as a candidate gene for AD susceptibility, multiple mutations of *TP73* were found to be irreverent with cognitive impairment, including rs6697769, rs10752739 and rs2181484 [89]. In addition, quite a few mutations are assigned to high-importance scores. For example, *MTOR* rs4845987 and *JUN* rs2760501 are classified as high impact mutations, and it had been reported that *MTOR* was a culprit and crucial in both Aβ and tau pathology and JNK/c-JUN cascade was activated in neurons of the AD brain [90, 91]. These results could likely help illuminate and understand the potential cause of cognitive disorders.

At the gene level, we summarized the mutation effects for each gene and found the mutations of some genes are likely to drive cognitive development, and the top 20 genes include *LRP1B*, *WWOX*, *MACROD2*, *FHIT*, *CTNNA3*, *CNTN5*, *PTPRN2*, *NRXN3*, *CTNNA2*, *ERBB4*, *GPC6*, *CSMD3*, *GRM7*, *FRMD4A*, *NRG1*, *CAMTA1*, *ASIC2*, *CTNND2*, *CDH4* and *THSD7B*. All these genes could enable follow-up functional studies in the future given that the effect size of mutations in these genes was dominant. In marked contrast, *LY6G6F*, *PPP1CA*, *HOXB7*, *LAMTOR2*, *GPX1*, *DDIT3*, *AGER*, *TAS2R9*, *TNF*, *METTL1*, *CERS1*, *MIF*, *APEX1*, *H2AFX*, *PIGC*, *TERC*, *PMCH*, *HTRA2*, *PINLYP* and *ATP5J2-PTCD1* would be less important in terms of their mutations contributing to AD progression. The detailed effect score for each gene is shown in Supplementary Table 4.

Deep-SMCI identifies promising AD effector mutations

To facilitate the potential therapeutic target identification, we focused on the top 20 000 potential AD effectors based on the importance scores for cognitive impairment. Based on statistical tests, we identified 527 AD-promoting mutations and 1459 AD-suppressive mutations.

As shown in Figure 4A, *CDC42* rs71016987 was negatively associated with cognitive dysfunction significantly (*t*-test, *P*-value < 2.2e−16). *CDC42* had been found to be highly expressed in AD patients compared with age-matched controls, and our mutation finding could help elucidate the relationship between this mutant's differential expression and cognitive development [92]. It had been reported that the low brain leptin signaling would

worsen hippocampal function and decrease neuroprotective effects such as tau and Aβ. Consistent with this observation, we found *LEP* rs10258300 might suppress neurodegenerative dysfunction (Figure 4B, *t*-test, *P*-value < 2.2e−16) [93]. In addition, we found that *FOXO4* rs5981069 and *HSPA9* rs535407352 would promote cognitive dysfunction (Figure 4C and D, *t*-test, *P*-value = 0.0028 and 0.0075, respectively). To note, there was evidence showing that these two genes were associated with AD. For example, FOXO protein participated in protecting neurons against Aβ-induced toxicity, and HSPA9 mRNA abnormal expression was associated with AD patients [94, 95].

Moreover, we observed three types of AD genes. Although there was evidence showing NRG1/ErbB4 complex could induce neuroinflammation and reduce memory formation and accumulation in neuritic plaques, we found the different mutations of *NRG1* could provide converse effects [96]. *NRG1* rs6468112 might promote cognitive impairment (Supplementary Figure 1A, *t*-test, *P*-value = 0.015), and *NRG1* rs1554540779, however, seemed to impede cognitive decline (Supplementary Figure 1B, *t*-test, *P*-value = 0.0039). As for *ADCY5*, we found all mutations of this gene were likely to be AD suppressive such as rs188403297 (Supplementary Figure 1C, *t*-test, *P*-value = 0.046). However, in another gene *RAD52*, all mutation effects for this gene were likely to promote cognitive impairment (Supplementary Figure 1D, *t*-test, *P*-value = 0.032). The complete list of all AD mutation effectors is shown in Supplementary Table 5.

Deep-CMCI outperforms other computational methods based on co-mutation landscapes

Furthermore, we also customized and optimized Deep-CMCI for predicting cognitive impairment of subjects based on co-mutation landscapes. Similarly, we compared the model to the abovementioned DT, GB, SVR and EN after grid search optimization and GRU, and LSTM after Bayesian optimization in the aspect of MAE and RMSE based on 10-fold cross-validation with five repeats from randomization. The detailed information for the optimization hyperparameter of Deep-CMCI and grid hyperparameter space of those baselines is shown in Supplementary Table 2. As shown in Figure 3C, Deep-CMCI achieved the best MAE and the least SD of MAE compared with the other four models. In this co-mutation landscape, Deep-CMCI achieved MAE at 3.01 ± 0.080 (SD) when predicting ADAS-Cog-11; however, SVR only got 3.78 ± 0.235 (SD), EN got 3.88 ± 0.247 (SD), DT got 4.06 ± 0.286 (SD), GB got 4.37 ± 0.321 (SD), GRU got 3.46 ± 0.212 (SD) and LSTM got 3.92 ± 0.321 (SD).

In addition, Deep-CMCI achieved the best RMSE and the least SD of RMSE compared with the other four models (Figure 3D). In the aspect of RMSE, Deep-CMCI achieved 4.15 ± 0.149 (SD) when predicting ADAS-Cog-11 based on the co-mutation landscape; however, SVR got 5.17 ± 0.467 (SD), EN got 5.11 ± 0.465 (SD), DT got 5.31 ± 0.512 (SD), GB got 5.38 ± 0.495 (SD), GRU got 4.86 ± 0.311 (SD), LSTM got 5.76 ± 0.542 (SD). The detailed predictive performance in this part is shown in Supplementary Table 3.

Compared with the predictive performance of the single mutation landscape, Deep-CMCI performed even better based on the co-mutation landscape, suggesting the co-mutation landscape we used was more comprehensive and capable of representing the cognitive status of AD subjects. To note, it was found most of the deep learning models in the performance evaluation part generally performed better than other traditional machine learning methods in predicting AD impairment scores based on mutation landscapes and co-mutation landscapes. A possible explanation for this is that deep learning models possess much richer representability, and are capable of capturing complex features such as the interactions between mutations and mutations or cognitive scores, making learning intrinsic characteristics much easier [97].

Deep-CMCI identifies potential AD-liable co-mutation effectors

Given that Deep-CMCI achieved intriguing MAE and RMSE compared with baselines, it was necessary to apply the model to dive into the co-mutation contributing to cognitive impairment as well.

To find the potential co-mutation effectors, we applied the Deep-CMCI to the co-mutation landscapes. Based on permutation tests, the model evaluated the contribution of each co-mutation to cognitive impairment and ranked the co-mutations based on their effect score. Here we picked out the top 20 000 potential AD co-mutation effectors based on the importance scores for cognitive impairment and finally identified 1366 AD co-promoting mutations and 1187 AD co-suppressive mutations.

As shown in Figure 5, *PARP1*_rs2027440-*ERCC4*_rs3784872 and *TP63*_rs4234613-*RAD51B*_rs117921607 co-mutation pairs were negatively associated with cognitive impairment significantly (Figure 5A and B, *t*-tests, *P*-value = 0.022, 0.0012, respectively). However, *FAT4*_rs10025859-*CSMD3*_rs11783840 and *CSMD3*_rs10109993-*FAT4*_rs10025859 were positively associated with cognitive decline (Figure 5C and D, *t*-tests, *P*-value = 0.041, 0.031, respectively). Interestingly, there was evidence showing that these four gene pairs exhibited corresponding interactions in known interactome networks. Based on the gene network from HumanNet v3, of 2553 co-mutation effectors, there were existing gene networks for 60 co-promoting mutations and 48 co-suppressive mutations we identified [98]. The list of all AD co-mutation effectors is shown in Supplementary Table 6.

AD mutation and co-mutation effectors discovery by other interpretability strategies

To further find more potential mutation and co-mutation effectors, we also adopted MI and Lime to rank and identify mutations and co-mutations. Similarly, we picked out the top 20 000 mutations and co-mutations based on the importance scores, to identify potential effectors based on statistical tests.

For AD mutations, we identified 624 and 611 AD-promoting effectors by Lime and MI, respectively, and there were four shared AD-promoting effectors by the three methods, including rs4951685, rs7825574, rs11988948, rs33983420 (Supplementary Figure 2A); we identified 1362 and 1465 AD suppressive effectors by Lime and MI, respectively, and there were five shared AD suppressive effectors by the three methods, including rs1286756, rs1540876, rs2005768, rs75655297 and rs231305 (Supplementary Figure 2B). For AD co-mutations, we identified 1130 and 1140 AD co-promoting effectors by Lime and MI, respectively, and there were 137 shared AD co-promoting effectors by the three methods (Supplementary Figure 2C); we identified 1181 and 1269 AD co-suppressive effectors by Lime and MI, respectively, and there were 50 shared AD co-suppressive effectors by the three methods (Supplementary Figure 2D). The list of AD mutation and co-mutation effectors are shown in Supplementary Table 7.

The primary reason why there were only a small portion of effectors overlapping might be the distinct characteristics of the three strategies: feature importance was measured by examining how much the score decreases when a feature was removed in the permutation test; MI was calculated between two features

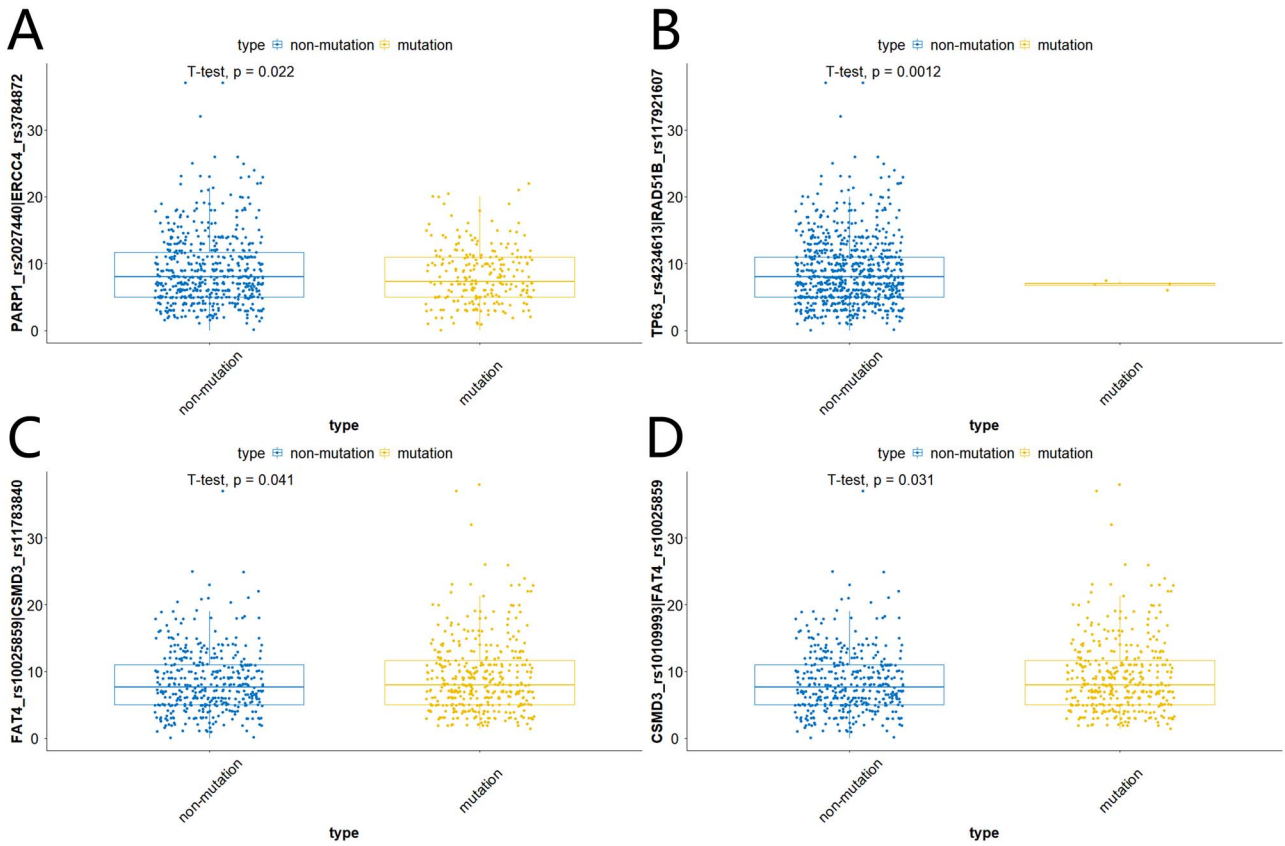

Figure 5. Examples of co-mutation effectors in cognitive impairment. (**A**) *PARP1*_rs2027440-*ERCC4*_rs3784872 is an AD-suppressive mutation pair (t-test, *P*-value = 0.022). (**B**) *TP63*_rs4234613-*RAD51B*_rs117921607 is an AD-suppressive mutation pair (t-test, *P*-value = 0.0012). (**C**) *FAT4*_rs10025859-*CSMD3*_rs11783840 is an AD-promoting mutation pair (t-test, *P*-value = 0.041). (**D**) *CSMD3*_rs10109993-*FAT4*_rs10025859 is an AD-promoting mutation pair (t-test, *P*-value = 0.031).

and measured the reduction in uncertainty for one feature given a known value of the other feature; in the case of Lime, it perturbed the data sample of interest, and learned a sparse linear model around it as an explanation. Another reason might be that we only prioritized top 20 000 mutations and co-mutations from the three algorithms. If we were interested in those mutations and co-mutations of lower ranking, there would be a larger overlap of mutation effectors among the three methods.

Discussion

As one of the most prevalent and difficult diseases to treat, AD is typically characterized by dementia, which often begins with a slow decline in cognitive function especially the memory of subjects, and gradually becomes more severe and incapacitating. Research reports have shown that genetic factors may play a vital role in AD development and progression [7, 14–19, 99, 100], therefore understanding the genetic causes of AD may benefit its early detection, prevention and even therapy. Although there were plenty of works focused on AD genetics in the past decades, the molecular mechanisms of AD development remain enigmatic, and the genetic etiology of most AD cases is still restricted and lagging. There are quite a few restrictions in the previous AD works: most works are focused on several genes or variants, making their findings low throughput; small sample size and even being restricted to familial individuals or populations make identifying rare variants challenging and even cannot be confirmed across independent research works; failure to take account into the interaction between genes and variants and lacking a unified and comprehensive metric characterizing cognition impairment of subjects are also important reasons why genetic findings are not interpretable enough for AD cases and fluctuating across existing AD works [25, 29, 30, 38–49].

The method we propose not only can predict the cognitive impairment of subjects robustly but also overcome the limitations of previous research works and provide deep insights into AD development and progression. By using optimized Deep-SMCI and Deep-CMCI, researchers can evaluate the cognitive status of potential subjects based on available mutation profiles. Besides, based on whole-genome mutation profiles of ~1000 AD-associated genes, our method systemically evaluates the effects of those mutations on cognitive development and identifies AD mutation effectors in a comprehensive and high-throughput way, which can offer useful resources and facilitate other relevant AD research works. Network medicine has been proven to be more effective and increasingly appreciated for dissecting the genetic organization of complicated diseases such as AD, and our method integrates the interactions between mutations in AD-associated genes and derives comprehensive co-mutation effectors in AD cases, laying a solid co-mutation landscape of AD. Our results further support that co-mutations may play an important role in AD development, if not more than single mutations.

To further illuminate the mechanism of AD systematically, there are several directions as future perspectives. As AD is a known age-dependent disorder, more evidence shows there is a progressive process during AD development and dynamic

changes in mutational signatures occur for the same AD subject [101–103]. For example, the effect of APOE $\varepsilon 4$ varies depending on the stage of AD [103]. Therefore, it will be worth exploring and tracking the mutation profiles of the same AD subjects longitudinally if data sets are available, and this will depict a better mutation landscape change model as AD progresses and benefit understanding the mechanism of AD further. In addition, aneuploidy, copy number variation, as well as transposable element mobilization are found to be involved in AD development [104–107]. For instance, APP copy number gains are enriched in the majority of AD neurons compared with control neurons [105]. As the next-generation sequencing technologies advance, integrating this information will expand the knowledge of AD. Moreover, genetic interactions existing in brain tissues are far more diverse and complicated, therefore how to characterize the complex interactions and incorporate them into AD development will be a long-standing topic [108]. Last but not least, several types of neurodegenerative diseases including Parkinson's disease, amyotrophic lateral sclerosis and Huntington's disease are found to share clinical and pathologic features with AD, and integrating those overlapping genetic factors and exclusive genetic factors may be helpful in understanding AD mechanisms [14, 109–111]. For example, these common features of the abovementioned diseases might be vital to the abnormal protein aggregates in the nervous system. Together, our findings here provide a valuable resource for further AD research in the scientific community, and also could possibly facilitate the development of more effective diagnosis and treatment for AD.

Key Points

- A robust deep learning model for liable mutations and co-mutations responsible in Alzheimer's disease.
- Accurate assessment of mutation prediction is critical with significant clinical implication, and our prediction models could robustly predict drug targets and combinations with high confidence.
- Our deep learning model could identify landmark signatures associated with neurological disease development, and provide novel and effective therapeutic targets for hard-to-treat disease.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgements

The authors acknowledge the Biomedical Research Computing Facility at UT Austin, and Texas Advanced Computing Center (TACC) for high-performance computing assistance.

Funding

The National Institutes of Health (grants R35GM133658 to S.S.Y., R01AG066749 to X.J., R35GM137836 to N.S.). S.S.Y. was also supported by Scialog Award# 28706 sponsored jointly by Chan Zuckerberg Initiative, Research Corporation for Science Advancement (RCSA) and the Cottrell Foundation.

Data availability

Raw data and clinical data are available at https://adni.loni.usc.edu. Processed data and codes are available at https://github.com/Pan-Bio/AD-mutation-effectors.

References

1. Jakob-Roetne R, Jacobsen H. Alzheimer's disease: from pathology to therapeutic approaches. *Angew Chem Int Ed* 2009;**48**:3030–59.
2. Shaw LM, Vanderstichele H, Knapik-Czajka M, *et al.* Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. *Ann Neurol* 2009;**65**:403–13.
3. Maccioni RB, Rojo LE, Fernandez JA, *et al.* The role of neuroimmunomodulation in Alzheimer's disease. *Ann N Y Acad Sci* 2009;**1153**:240–6.
4. Alzheimer's Association. 2019 Alzheimer's disease facts and figures. *Alzheimers Dement* 2019;**15**:321–87.
5. Trejo-Lopez JA, Yachnis AT, Prokop S. Neuropathology of Alzheimer's disease. *Neurotherapeutics* 2022;**19**:173–85.
6. Dosunmu R, Wu J, Basha MR, *et al.* Environmental and dietary risk factors in Alzheimer's disease. *Expert Rev Neurother* 2007;**7**:887–900.
7. Wingo TS, Lah JJ, Levey AI, *et al.* Autosomal recessive causes likely in early-onset Alzheimer disease. *Arch Neurol* 2012;**69**:59–64.
8. Gatz M, Reynolds CA, Fratiglioni L, *et al.* Role of genes and environments for explaining Alzheimer disease. *Arch Gen Psychiatry* 2006;**63**:168–74.
9. Goate A, Chartier-Harlin M-C, Mullan M, *et al.* Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature* 1991;**349**:704–6.
10. Sherrington R, Rogaev E, Ya L, *et al.* Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature* 1995;**375**:754–60.
11. Levy-Lahad E, Wasco W, Poorkaj P, *et al.* Candidate gene for the chromosome 1 familial Alzheimer's disease locus. *Science* 1995;**269**:973–7.
12. Kim J, Basak JM, Holtzman DM. The role of apolipoprotein E in Alzheimer's disease. *Neuron* 2009;**63**:287–303.
13. Jonsson T, Stefansson H, Steinberg S, *et al.* Variant of TREM2 associated with the risk of Alzheimer's disease. *N Engl J Med* 2013;**368**:107–16.
14. Giau VV, Bagyinszky E, An SSA, *et al.* Clinical genetic strategies for early onset neurodegenerative diseases. *Mol Cell Toxicol* 2018;**14**:123–42.
15. Cacace R, Sleegers K, Van Broeckhoven C. Molecular genetics of early-onset Alzheimer's disease revisited. *Alzheimers Dement* 2016;**12**:733–48.
16. Karch CM, Goate AM. Alzheimer's disease risk genes and mechanisms of disease pathogenesis. *Biol Psychiatry* 2015;**77**:43–51.
17. Brouwers N, Sleegers K, Van Broeckhoven C. Molecular genetics of Alzheimer's disease: an update. *Ann Med* 2008;**40**:562–83.
18. Guerreiro RJ, Gustafson DR, Hardy J. The genetic architecture of Alzheimer's disease: beyond APP, PSENs and APOE. *Neurobiol Aging* 2012;**33**:437–56.
19. Ridge PG, Hoyt KB, Boehme K, *et al.* Assessment of the genetic variance of late-onset Alzheimer's disease. *Neurobiol Aging* 2016;**41**:200. e213–20.
20. Folch J, Petrov D, Ettcheto M, *et al.* Current research therapeutic strategies for Alzheimer's disease treatment. *Neural Plast* 2016;**2016**:8501693.

21. Kamboh MI. Molecular genetics of late-onset Alzheimer's disease. *Ann Hum Genet* 2004;**68**:381–404.
22. Lambert JC. F1–01–01: meta-analysis in more than 74,000 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Alzheimers Dement* 2013;**9**:P123–3.
23. Avila J, Gómez-Ramos A, Bolós M. AD genetic risk factors and tau spreading. *Front Media SA* 2015;**7**:99.
24. Harold D, Abraham R, Hollingworth P, *et al.* Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat Genet* 2009;**41**:1088–93.
25. Li Y, Grupe A, Rowland C, *et al.* DAPK1 variants are associated with Alzheimer's disease and allele-specific expression. *Hum Mol Genet* 2006;**15**:2560–8.
26. Bertram L, Tanzi RE. The genetic epidemiology of neurodegenerative disease. *J Clin Invest* 2005;**115**:1449–57.
27. Bertram L, Blacker D, Mullin K, *et al.* Evidence for genetic linkage of Alzheimer's disease to chromosome 10q. *Science* 2000;**290**:2302–3.
28. Izzo G, Forlenza OV, Santos BD, *et al.* Single-nucleotide polymorphisms of GSK3B, GAB2 and SORL1 in late-onset Alzheimer's disease: interactions with the APOE genotype. *Clinics* 2013;**68**:277–80.
29. Rogaeva E, Meng Y, Lee JH, *et al.* The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer disease. *Nat Genet* 2007;**39**:168–77.
30. Günther C, Von Hadeln K, Müller-Thomsen T, *et al.* Possible association of mitochondrial transcription factor A (TFAM) genotype with sporadic Alzheimer disease. *Neurosci Lett* 2004;**369**:219–23.
31. Pottier C, Ravenscroft TA, Brown PH, *et al.* TYROBP genetic variants in early-onset Alzheimer's disease. *Neurobiol Aging* 2016;**48**:e229–2, 222. e215.
32. Guerreiro RJ, Lohmann E, Kinsella E, *et al.* Exome sequencing reveals an unexpected genetic cause of disease: NOTCH3 mutation in a Turkish family with Alzheimer's disease. *Neurobiol Aging* 2012;**33**:e1017–08, 1008. e1023.
33. Pottier C, Hannequin D, Coutant S, *et al.* High frequency of potentially pathogenic SORL1 mutations in autosomal dominant early-onset Alzheimer disease. *Mol Psychiatry* 2012;**17**:875–9.
34. Nicolas G, Charbonnier C, Wallon D, *et al.* SORL1 rare variants: a major risk factor for familial early-onset Alzheimer's disease. *Mol Psychiatry* 2016;**21**:831–6.
35. Backman JD, Li AH, Marcketta A, *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* 2021;**599**:628–34.
36. Mbatchou J, Barnard L, Backman J, *et al. Computationally Efficient Whole-Genome Regression for Quantitative and Binary Traits*. Nature Genetics 2021;**53**:1097–1103.
37. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996;**273**:1516–7.
38. Slifer MA, Martin ER, Haines JL, *et al.* The ubiquilin 1 gene and Alzheimer's disease. *N Engl J Med* 2005;**352**:2752–3.
39. Li Y, Rowland C, Catanese J, *et al.* SORL1 variants and risk of late-onset Alzheimer's disease. *Neurobiol Dis* 2008;**29**:293–6.
40. Webster JA, Myers AJ, Pearson JV, *et al.* Sorl1 as an Alzheimer's disease predisposition gene? *Neurodegener Dis* 2008;**5**:60–4.
41. Logue MW, Schu M, Vardarajan BN, *et al.* Two rare AKAP9 variants are associated with Alzheimer's disease in African Americans. *Alzheimers Dement* 2014;**10**:609–618. e611.
42. van der Lee SJ, Wolters FJ, Ikram MK, *et al.* The effect of APOE and other common genetic variants on the onset of Alzheimer's disease and dementia: a community-based cohort study. *Lancet Neurol* 2018;**17**:434–44.
43. Nicolas G, Acuña-Hidalgo R, Keogh MJ, *et al.* Somatic variants in autosomal dominant genes are a rare cause of sporadic Alzheimer's disease. *Alzheimers Dement* 2018;**14**:1632–9.
44. Choi KY, Lee JJ, Gunasekaran TI, *et al.* APOE promoter polymorphism-219T/G is an effect modifier of the influence of APOE ε4 on Alzheimer's disease risk in a multiracial sample. *J Clin Med* 2019;**8**:1236.
45. Cechova K, Andel R, Angelucci F, *et al.* Impact of APOE and BDNF Val66Met gene polymorphisms on cognitive functions in patients with amnestic mild cognitive impairment. *J Alzheimers Dis* 2020;**73**:247–57.
46. Brown DT, Vickers JC, Stuart KE, *et al.* The BDNF Val66Met polymorphism modulates resilience of neurological functioning to brain ageing and dementia: a narrative review. *Brain Sci* 2020;**10**:195.
47. Laczó J, Cechova K, Parizkova M, *et al.* The combined effect of APOE and BDNF Val66Met polymorphisms on spatial navigation in older adults. *J Alzheimers Dis* 2020;**78**:1473–92.
48. Gómez-Tortosa E, Ruggiero M, Sainz MJ, *et al.* SORL1 variants in familial Alzheimer's disease. *J Alzheimers Dis* 2018;**61**:1275–81.
49. Korpioja A, Krüger J, Koivuluoma S, *et al.* Novel rare SORL1 variants in early-onset dementia. *J Alzheimers Dis* 2021;**82**:761–70.
50. Fuxman Bass JI, Sahni N, Shrestha S, *et al.* Human gene-centered transcription factor networks for enhancers and disease variants. *Cell* 2015;**161**:661–73.
51. Sahni N, Yi S, Taipale M, *et al.* Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* 2015;**161**:647–60.
52. Rolland T, Tasan M, Charloteaux B, *et al.* A proteome-scale map of the human interactome network. *Cell* 2014;**159**:1212–26.
53. Sahni N, Yi S, Zhong Q, *et al.* Edgotype: a fundamental link between genotype and phenotype. *Curr Opin Genet Dev* 2013;**23**:649–57.
54. Bis JC, Jian X, Kunkle BW, *et al.* Whole exome sequencing study identifies novel rare and common Alzheimer's-associated variants involved in immune response and transcriptional regulation. *Mol Psychiatry* 2020;**25**:1859–75.
55. Raghavan NS, Brickman AM, Andrews H, *et al.* Whole-exome sequencing in 20,197 persons for rare variants in Alzheimer's disease. *Ann Clin Transl Neurol* 2018;**5**:832–42.
56. Zhang X, Farrell JJ, Tong T, *et al.* Association of mitochondrial variants and haplogroups identified by whole exome sequencing with Alzheimer's disease. *Alzheimers Dement* 2022;**18**:294–306.
57. Cheng F, Zhao J, Wang Y, *et al.* Comprehensive characterization of protein-protein interactions perturbed by disease mutations. *Nat Genet* 2021;**53**:342–53.
58. Greene JA, Loscalzo J. Putting the patient back together - social medicine, network medicine, and the limits of reductionism. *N Engl J Med* 2017;**377**:2493–9.
59. Silbersweig D, Loscalzo J. Precision psychiatry meets network medicine: network psychiatry. *JAMA Psychiat* 2017;**74**:665–6.
60. Li Y, McGrail DJ, Latysheva N, *et al.* Pathway perturbations in signaling networks: linking genotype to phenotype. *Semin Cell Dev Biol* 2020;**99**:3–11.
61. Goh KI, Cusick ME, Valle D, *et al.* The human disease network. *Proc Natl Acad Sci U S A* 2007;**104**:8685–90.
62. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* 2011;**12**:56–68.

63. Silverman EK, Loscalzo J. Network medicine approaches to the genetics of complex diseases. *Discov Med* 2012;**14**:143–52.
64. Li Y, Burgman B, Khatri IS, *et al.* E-MutPath: computational modeling reveals the functional landscape of genetic mutations rewiring interactome networks. *Nucleic Acids Res* 2021;**49**:e2.
65. Li Y, McGrail DJ, Xu J, *et al.* Gene regulatory network perturbation by genetic and epigenetic variation. *Trends Biochem Sci* 2018;**43**:576–92.
66. Pan X, Burgman B, Wu E, *et al.* I-modern: integrated multi-omics network model identifies potential therapeutic targets in glioma by deep learning with interpretability. *Comput Struct Biotechnol J* 2022;**20**:3511–21.
67. Pan X, Liu B, Wen X, *et al.* D-GPM: a deep learning method for gene promoter methylation inference. *Gen* 2019;**10**:807.
68. Pan X, Huang LF. Multi-omics to characterize the functional relationships of R-loops with epigenetic modifications, RNAPII transcription and gene expression. *Brief Bioinform* 2022;**23**:bbac238.
69. Liu B, Liu Y, Pan X, *et al.* DNA methylation markers for pan-cancer prediction by deep learning. *Gen* 2019;**10**:778.
70. Zhao T, Hu Y, Valsdottir LR, *et al.* Identifying drug–target interactions based on graph convolutional network and deep neural network. *Brief Bioinform* 2021;**22**:2141–50.
71. Bukhari SNH, Jain A, Haq E, *et al.* Ensemble machine learning model to predict SARS-CoV-2 T-cell epitopes as potential vaccine targets. *Diagnostics* 2021;**11**:1990.
72. Bukhari SNH, Webber J, Mehbodniya A. Decision tree based ensemble machine learning model for the prediction of Zika virus T-cell epitopes as potential vaccine candidates. *Sci Rep* 2022;**12**:1–11.
73. Wang S, Jiang Y, Li S. PStrain: an iterative microbial strains profiling algorithm for shotgun metagenomic sequencing data. *Bioinformatics* 2021;**36**:5499–506.
74. Pan X, Burgman B, Sahni N, *et al.* Deep learning based on multi-omics integration identifies potential therapeutic targets in breast cancer. bioRxiv 2022.
75. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 2017;**30**:4765–74.
76. de Magalhaes JP, Toussaint O. GenAge: a genomic and proteomic network map of human ageing. *FEBS Lett* 2004;**571**:243–7.
77. Budovsky A, Craig T, Wang J, *et al.* LongevityMap: a database of human genetic variants associated with longevity. *Trends Genet* 2013;**29**:559–60.
78. Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Phys D Nonlinear Phenom* 2020;**404**:132306.
79. Shen Z, Bao W, Huang D-S. Recurrent neural network for predicting transcription factor binding sites. *Sci Rep* 2018;**8**:1–10.
80. Ioffe S, Szegedy C: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 448–56. PMLR.
81. Tinto V. Dropout from higher education: a theoretical synthesis of recent research. *Rev Educ Res* 1975;**45**:89–125.
82. Prechelt L. Automatic early stopping using cross validation: quantifying the criteria. *Neural Netw* 1998;**11**:761–7.
83. Kingma DP, Ba J. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. 2014.
84. Fao C. *Keras*, 2015. https://github.com/fchollet/keras.
85. Abadi M, Agarwal A, Barham P, *et al. TensorFlow: Large-scale machine learning on heterogeneous systems*. In: *12th USENIX Symposium on Operating Systems Design and Implementation*, 2015, pp. 265–83.
86. Buitinck L, Louppe G, Blondel M, *et al.* API design for machine learning software: experiences from the scikit-learn project. arXiv preprint arXiv:1309.0238. 2013.
87. Nogueira F. *Bayesian optimization: open source constrained global optimization tool for Python*, 2014. https://github.com/fmfn/BayesianOptimization.
88. Ribeiro MT, Singh S, Guestrin C: 'Why should i trust you?' Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, New York, United States. 2016, pp. 1135–44.
89. Li Q, Athan ES, Wei M, *et al.* TP73 allelic expression in human brain and allele frequencies in Alzheimer's disease. *BMC Med Genet* 2004;**5**:1–9.
90. Mueed Z, Tandon P, Maurya SK, *et al.* Poddar NK: tau and mTOR: the hotspots for multifarious diseases in Alzheimer's development. *Front Neurosci* 2019;**12**:1017.
91. Lin A. Activation of the JNK signaling pathway: breaking the brake on apoptosis. *Bioessays* 2003;**25**:17–24.
92. Saraceno C, Catania M, Paterlini A, *et al.* Altered expression of circulating Cdc42 in frontotemporal lobar degeneration. *J Alzheimers Dis* 2018;**61**:1477–83.
93. McGuire MJ, Ishii M. Leptin dysfunction and Alzheimer's disease: evidence from cellular, animal, and human studies. *Cell Mol Neurobiol* 2016;**36**:203–17.
94. Du S, Zheng H. Role of FoxO transcription factors in aging and age-related metabolic and neurodegenerative diseases. *Cell Biosci* 2021;**11**:1–17.
95. Silva PN, Furuya TK, Braga IL, *et al.* Analysis of HSPA8 and HSPA9 mRNA expression and promoter methylation in the brain and blood of Alzheimer's disease patients. *J Alzheimers Dis* 2014;**38**:165–70.
96. Mouton-Liger F, Dumurgier J, Cognat E, *et al.* CSF levels of the BACE1 substrate NRG1 correlate with cognition in Alzheimer's disease. *Alzheimers Res Ther* 2020;**12**:1–10.
97. Glorot X, Bengio Y: Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–56. -PMLR.
98. Kim CY, Baek S, Cha J, *et al.* HumanNet v3: an improved database of human gene networks for disease research. *Nucleic Acids Res* 2022;**50**:D632–9.
99. Busche MA, Hyman BT. Synergy between amyloid-β and tau in Alzheimer's disease. *Nat Neurosci* 2020;**23**:1183–93.
100. Ittner A, Ittner LM. Dendritic tau in Alzheimer's disease. *Neuron* 2018;**99**:13–27.
101. Lodato MA, Rodin RE, Bohrson CL, *et al.* Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* 2018;**359**:555–9.
102. Park JS, Lee J, Jung ES, *et al.* Brain somatic mutations observed in Alzheimer's disease associated with aging and dysregulation of tau phosphorylation. *Nat Commun* 2019;**10**:1–12.
103. Liu Y, Tan L, Wang H-F, *et al.* Multiple effect of APOE genotype on clinical and neuroimaging biomarkers across Alzheimer's disease spectrum. *Mol Neurobiol* 2016;**53**:4539–47.
104. Head E, Lott IT, Wilcock DM, *et al.* Aging in down syndrome and the development of Alzheimer's disease neuropathology. *Curr Alzheimer Res* 2016;**13**:18–29.
105. Bushman DM, Kaeser GE, Siddoway B, *et al.* Genomic mosaicism with increased amyloid precursor protein (APP) gene copy number in single neurons from sporadic Alzheimer's disease brains. *Elife* 2015;**4**:e05116.

106. Thomas EO, Zuniga G, Sun W, *et al.* Awakening the dark side: retrotransposon activation in neurodegenerative disorders. *Curr Opin Neurobiol* 2020;**61**: 65–72.
107. Niu Y, Teng X, Zhou H, *et al.* Characterizing mobile element insertions in 5675 genomes. *Nucleic Acids Res* 2022;**50**: 2493–508.
108. Moon SW, Dinov ID, Kim J, *et al.* Initiative AsDN: structural neuroimaging genetics interactions in Alzheimer's disease. *J Alzheimers Dis* 2015;**48**:1051–63.
109. Santiago JA, Bottero V, Potashkin JA. Dissecting the molecular mechanisms of neurodegenerative diseases through network biology. *Front Aging Neurosci* 2017;**9**:166.
110. Lill CM, Bertram L. Towards unveiling the genetics of neurodegenerative diseases. In: *Seminars in Neurology*. Thieme Medical Publishers, 2011, 531–41.
111. Prince M, Bryce R, Albanese E, *et al.* The global prevalence of dementia: a systematic review and metaanalysis. *Alzheimers Dement*. The Journal for The Alzheimer's Association, 2013;**9**:63–75.e2. https://doi.org/10.1016/j.jalz.2012.11.007.