January ADSP Science Report **01/25/2021**

Project: Ultra-scale Machine Learning to Empower Discovery in Alzheimer's Disease Biobanks (U01 AG068057-01)

Report prepared by:
Paul M Thompson (Contact), Christos Davatzikos, Heng Huang, Andrew J Saykin, Li Shen, Duygu Tosun, Sarah Zaranek, Sasha Zaranek, Timothy Hohman, Gyungah Jun, Paul Crane, and Adam M Brickman, on behalf of the AI4AD investigators

A. Introduction

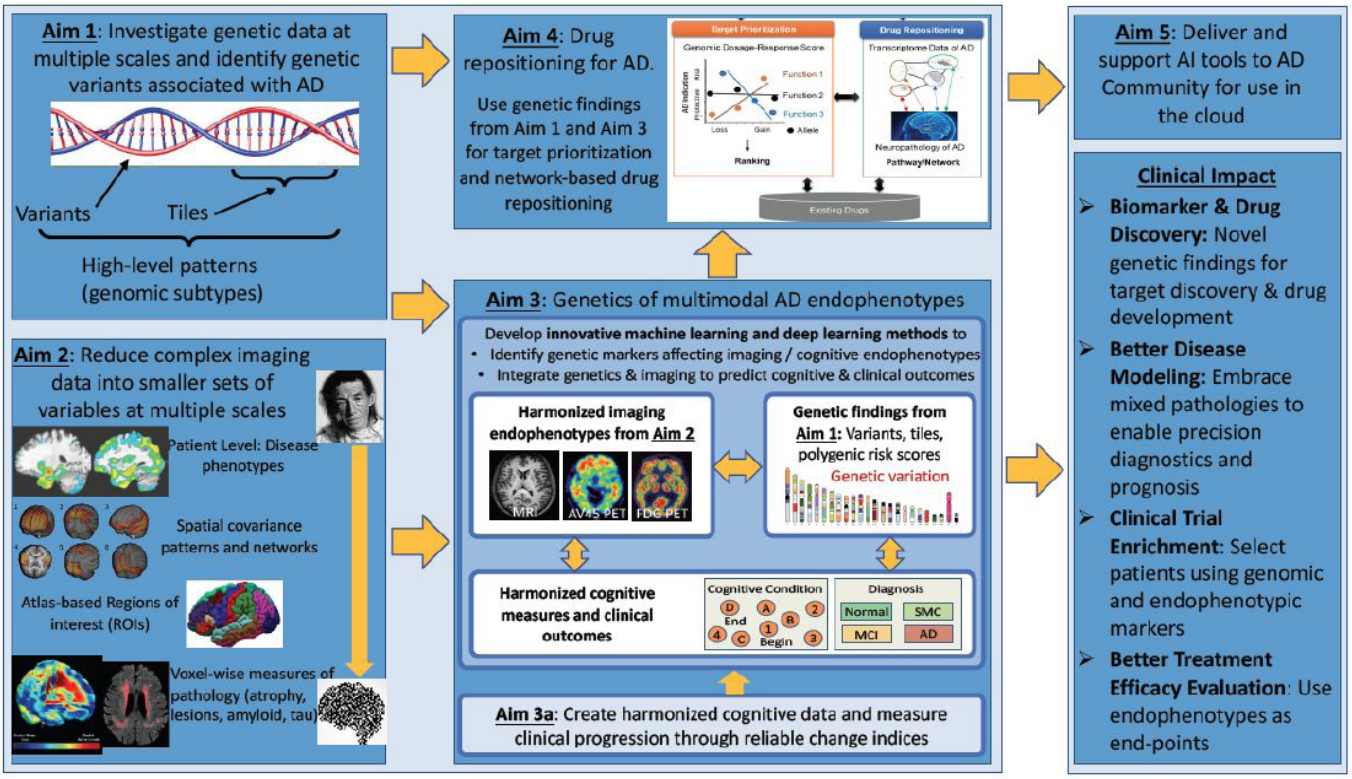


Figure 1. Goals of the AI4AD Initiative.

The overall goal of the Artificial Intelligence for Alzheimer's Disease (AI4AD) machine learning initiative is to use advances in deep learning, machine learning, and AI (artificial intelligence) to tackle key challenges in Alzheimer's disease (AD) research. The Project's overall Aims (**Figure 1**) are:
(1) **Genomics:** use AI to identify genomic motifs and features associated with AD, and clinical resilience and decline in whole genome sequence data;
(2) **Imaging Harmonization and Disease Subtyping:** use AI to merge and calibrate MRI, amyloid- and tau-PET and vascular imaging across cohorts to identify AD subtypes, and relate these subtypes to specific genomic predictors and outcomes;
(3) **Imaging and Genomic Predictors of Cognition:** use AI to identify genomic motifs that predict brain imaging signatures of AD and decline in specific cognitive domains;
(4) **Genome-Guided Drug Repurposing:** create a drug prioritization system to discover drugs to repurpose based on genomic markers discovered in the other Aims;
(5) **Train** the AD community in easy-to-use AI methods to accelerate AD research.

B. Progress to Date

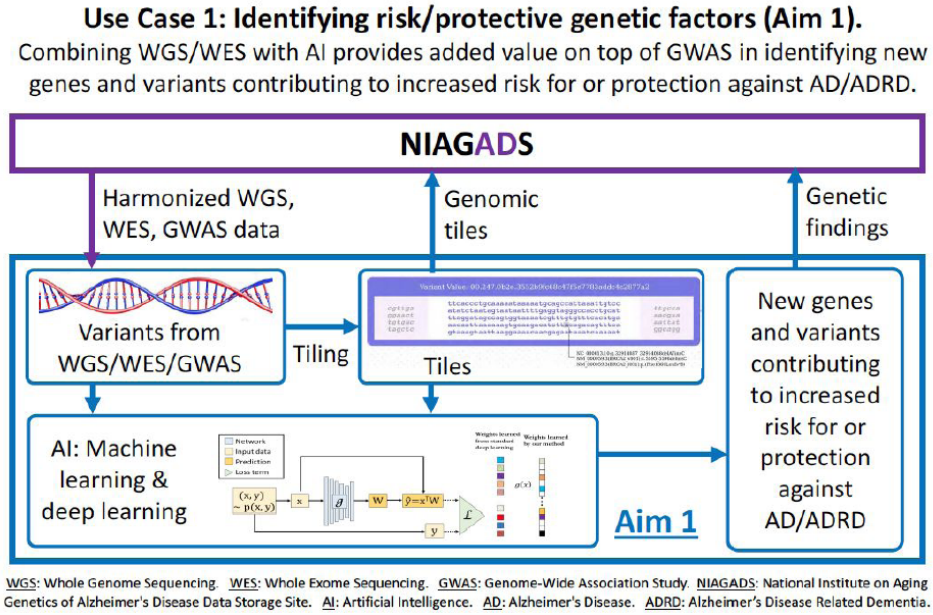


Figure 2. Data Flow for Identifying Risk/Protective Factors in the Genome, using AI.

1. AI applied to Whole Genome Sequences. Drs Sarah and Sasha Zaranek (Curii and Harvard University) have pioneered a method to represent whole genome sequences using a standardized and curated library of 'tiles' that can then be read into **AI and machine learning methods to identify AD risk and protective factors (Figure 2)**. This greatly reduces the massive-scale WGS data into tractable inputs for feature detection and integration. Tiling is a way to break up the genome into shorter sequences called tiles, which are defined by a set of two tags (24-mers). Pilot work on whole genome tiling shows that unsupervised sparse regression methods - such as adaptive LASSO - can discover genomic predictors of AD, and can combine them into weighted risk scores that merge the predictive effects of multiple genetic variants. In 4,000+ tiled genomes from ADSP and AD Neuroimaging Initiative (ADNI), the best fit model so far (GLM Adaptive Lasso) has identified 411 tile variants that help to predict AD status. Encouragingly, the top two coefficients were phase 0 and phase 1 of the APOE $\varepsilon 4$ variant (rs429358); ongoing work includes generating a ranked variant list and comparing discovered loci to those from standard mass-univariate GWAS.

Extending classical machine learning methods to handle whole genome sequences, Dr Sarah Zaranek presented to the AI4AD and ADSP Analysis Working Groups and gave tutorials on how to use whole-genome data with AI algorithms, by first representing it using tiles. She is now creating a GitHub tool repository to decompose whole genomes into tiles, so that computer scientists across AI4AD can work with WGS, starting with the ADNI and WHICAP datasets. This toolset will make it much easier to input enormous genetic sequencing datasets into novel AI methods, and will offer head-to-head testing of new AI methods against a suite of well-known traditional machine learning algorithms. Whole genome tiling is now underway for the ADNI and WHICAP datasets, for initial tests that will use deep learning to predict AD diagnosis, and prognosis (clinical decline). Tiled and curated WGS data will be returned to NIAGADS for wide public use. Parallel work by the ADSP Harmonization Core (Hohman, Saykin, and Crane labs) has co-calibrated 4 cognitive domains across ADNI and ROSMAP datasets, to allow deep learning methods to predict future decline in individual patients, from this same whole genome data. Benchmark tests already suggest that tiled whole genome data predict AD status better than standard regression methods (such as PRS) applied to univariate markers. This work is now being extended to the more ethnically diverse WHICAP cohort (1,700 individuals), with help from WHICAP Co-PI, Adam Brickman (Columbia Univ.).

2. AI applied to Brain Imaging for Predictions in Individual Patients. We developed state-of-the-art convolutional neural networks that can predict a person's age and whether they have Alzheimer's disease based on a standard T1-weighted brain MRI scan. We have used two complementary approaches to this problem. The first is summarized in **Figure 3** (Lam 2021, Zhang 2021, Stripelis 2021).

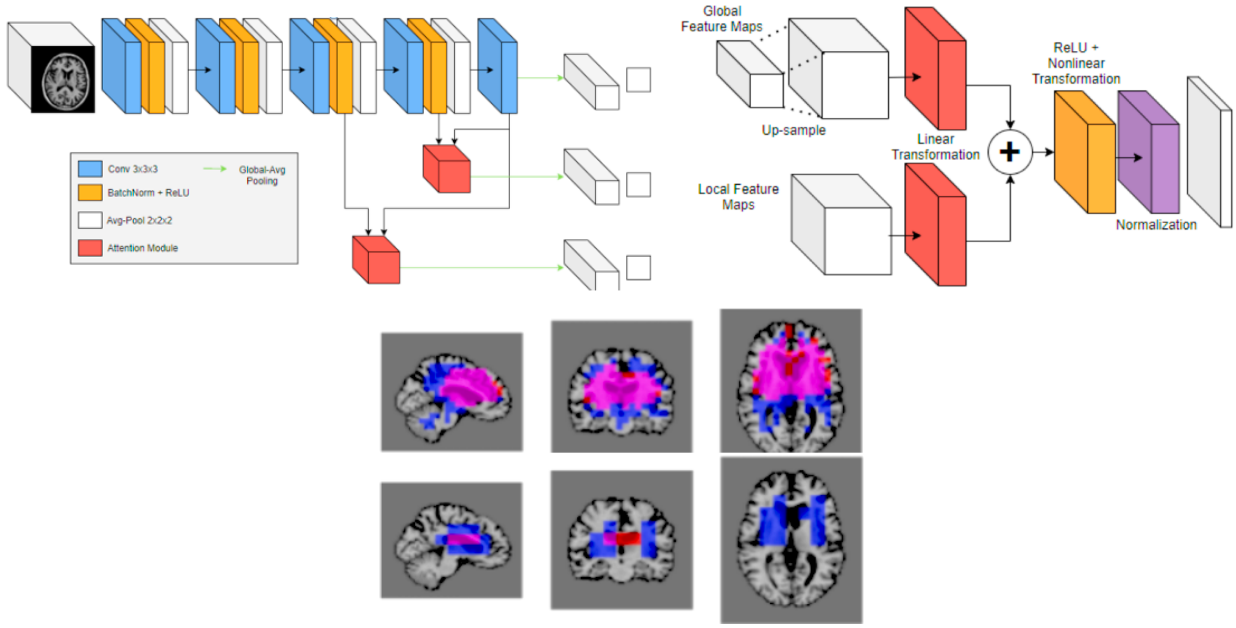
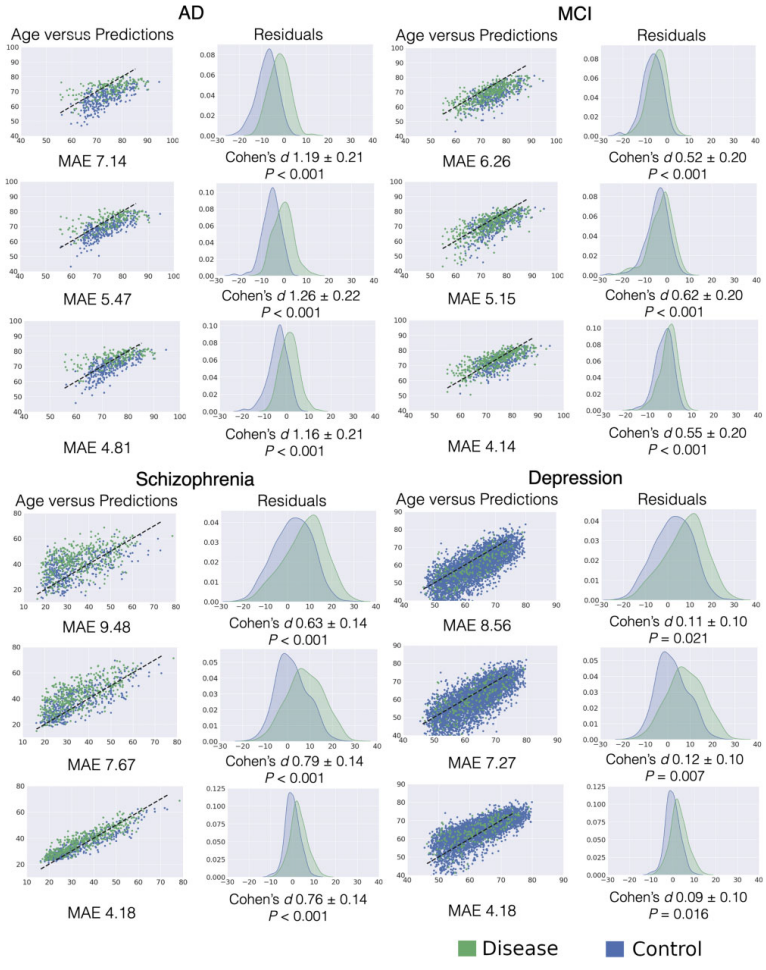


Figure 3. Gated-Attention Convolutional Neural Networks can make Individualized Patient Predictions from Brain MRI Scans. In Lam (2021), we supplemented a standard convolutional neural network (CNN; *upper left*) with an attention mechanism (*upper right*) to identify imaging regions and features within them that can make predictions for an individual patient (here estimating their age, and AD diagnosis). Trained on diverse AD biobanks, automated AI-driven feature discovery is also now being tested for identifying brain signatures associated with AD genomic risk signatures.



The second approach, developed by Dr. Davatzikos and his laboratory (Bashyam 2020) was trained on a large cohort of the adult lifespan, and was tested in various cross-validation paradigms. Critically, that study revealed that both very tight and very loose brain age models produce suboptimal clinical value, when it comes to using brain age as a disease related biomarker (tested on MCI, AD, schizophrenia, and depression; see **Figure 4**).

Figure 4. Results from Bashyam et al., *Brain, 2020.* Modestly-fitting brain age models provide the best diagnostic value in MCI, AD, schizophrenia, and depression, indicating that brain age models that try to very accurately predict an individual's age might produce clinically less-informative brain age residuals,

by virtue of focusing on disease-unrelated imaging features. These results will further guide some of our developments of deep learning diagnostic and predictive models, in this project.

Gated-attention mechanisms identify regions of the brain image that are especially useful for predicting a person's diagnosis, and are now being extended to identify brain regions where atrophy, microstructure, or amyloid-/tau- deposition are associated with diagnosis of AD and prognosis. The same approach is being tested for predicting amyloid load and CSF amyloid markers from MRI.

The Tosun Lab (UCSF) developed a novel machine learning approach to identify patterns in MRI and amyloid-sensitive PET data that predict neuropathologically confirmed cellular and molecular markers of AD and other related neurodegenerative diseases. This information is normally only obtainable at autopsy, and in a limited number of subjects - the ADNI and National Alzheimer's Coordinating Center (NACC) datasets have cellular histopathology in a subset of individuals. These paired datasets were used to train a machine learning method to predict specific cellular histologic markers; in living subjects, the inferred pathology was useful in predicting a proportion of the observed variance in clinical decline, over a 1-year follow up.

Dr Tosun recently presented this proposed model at the Alzheimer's Therapeutic Research Institute (ATRI) of USC as a precision medicine approach to clinical trials in sporadic AD. This subtyping of dementia is likely to be extremely valuable for drug trials. Drugs may work better in specific subtypes of patients identifiable using AI - either to guide enrolment or for later stratification. AI will be used to cluster patients by ATN(V) categorization for precision medicine (stratified testing of interventions).

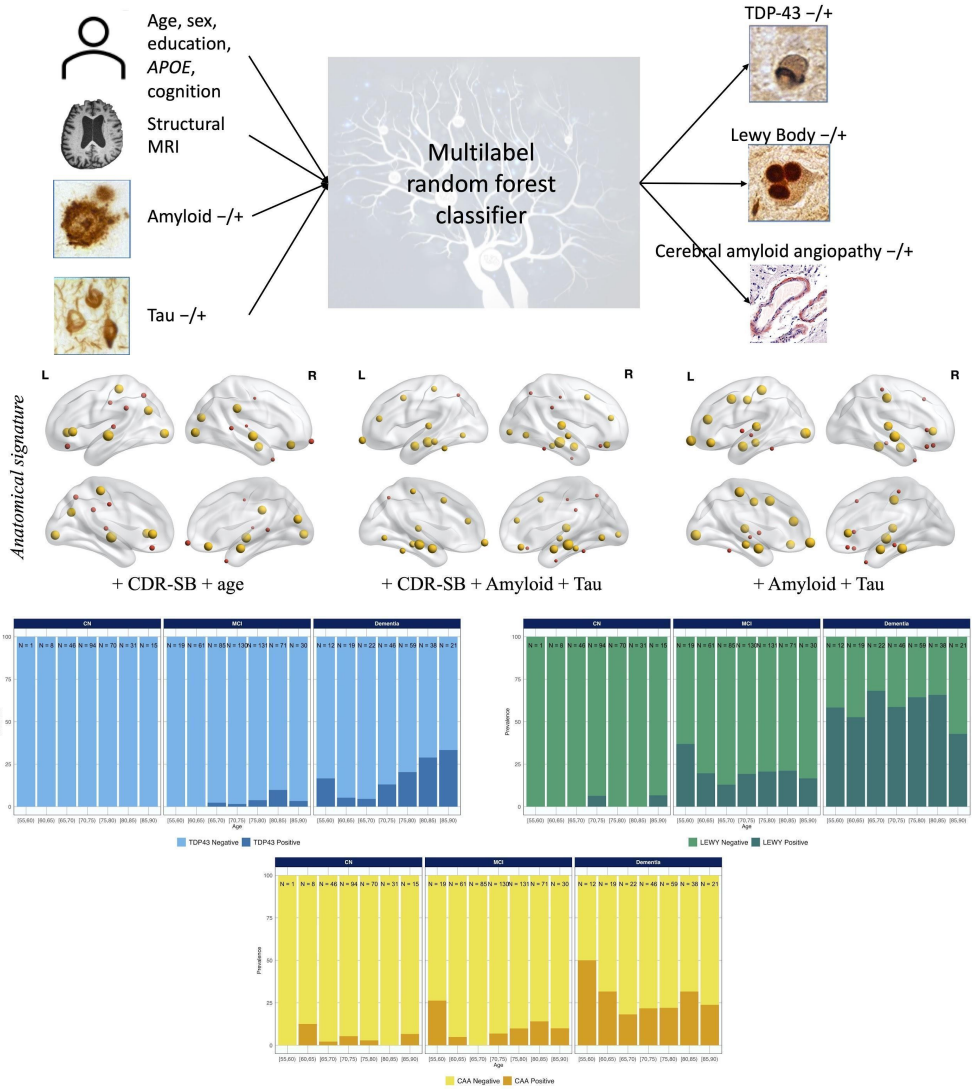


Figure 5. A multilabel classifier to identify individuals with non-AD co-pathologies. As presence of AD and comorbid non-AD pathologies (i.e., TDP-43, Lewy Body, and cerebral amyloid angiopathy) are not mutually exclusive, we developed a multilabel random-forest classifier to jointly predict presence of non-AD pathologies from demographic information (age, sex, education), *APOE* genotype, clinical measures, brain morphometry from structural MRI, and presence of AD pathologies of amyloid and tau. Trained on diverse AD autopsy cohorts, including ADNI and NACC, automated ML-driven anatomical brain signatures recapitulate brain regions affected by non-AD neuropathologies independent of AD pathologies. When applied to *in vivo* data, the ML model allows us to assess age and disease stage specific prevalence of mixed neuropathologies and their contribution to cognitive decline in living humans.

The Davatzikos lab (U Penn) and the Tosun lab (UCSF) developed and trained other team members in the use of AI methods to harmonize MRI and PET imaging data across scanners. AI methods including CycleGANs have revolutionized AI - making it easier to extend predictive models to data collected with different resolutions and contrasts. We are using them in AI4AD to harmonize data from multiple scanners, by mapping them to a canonical space of uniform image appearance. An example is shown in the figure below.

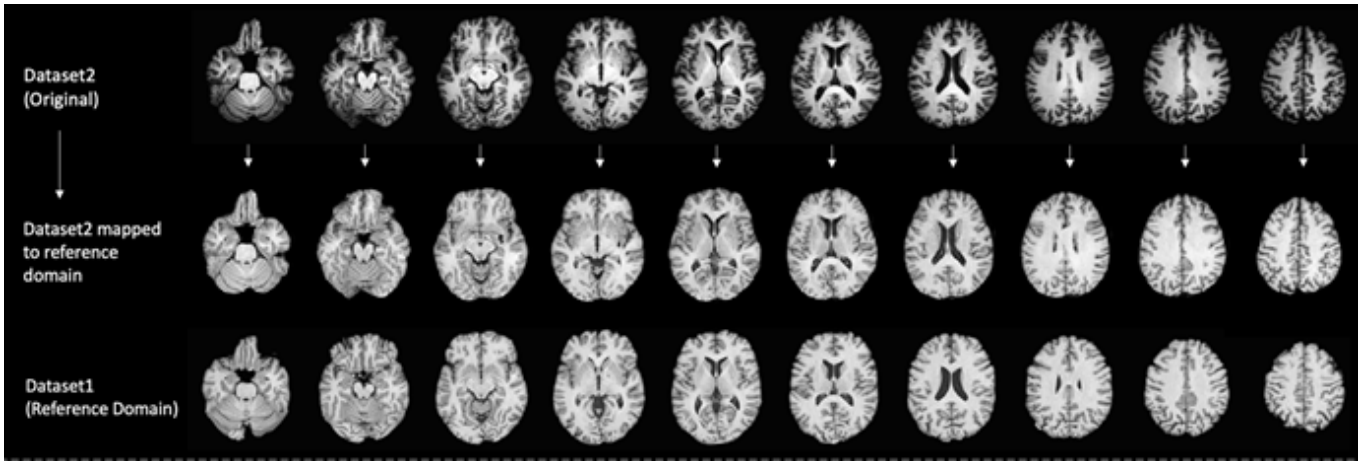


Figure 6. Top row shows an MRI scan from Scanner 2, and bottom row shows an MRI scan from Scanner 1, which is used as reference. The middle row shows the transformed scan from Scanner 2 to the reference appearance. This method was based on deep learning models that are now used in a variety of applications to change the stylistic appearance of an image without changing relevant aspects of its content. We will use these methodologies to reduce confounding variations due to scanners and MR acquisition protocols. Results from (Bashyam et al., *in review*).

The Huang lab (U Pittsburgh) developed a mathematically novel deep learning method (Zhang 2021a) that can make predictions in a smaller dataset using much larger auxiliary datasets for training, an approach known as transfer learning. The new deep neural network model is designed to use transferable batch normalization to eliminate the distribution difference of different brain datasets. They formulated the brain outcome and phenotype prediction task as a semi-supervised problem and the virtual adversarial training was introduced as the regularization. The experiments were conducted on ADNI and NACC T1-weighted brain MRI data to predict brain outcomes with good performance. In a second study (Zhang 2021b), they developed a new multimodal brain data integration method (the multimodal brain data could be multi-view connectomes or multimodal imaging genetic data) to predict cognitive status. They proposed to learn a unified representation for multimodal brain data using a novel variational graph autoencoder (VGAE). As an example, they applied this method to unify multi-layer brain networks computed from different commonly used fiber tractography algorithms. The new autoencoders create a merged “multi-view” generative representation, based on several complementary fiber tracking methods.

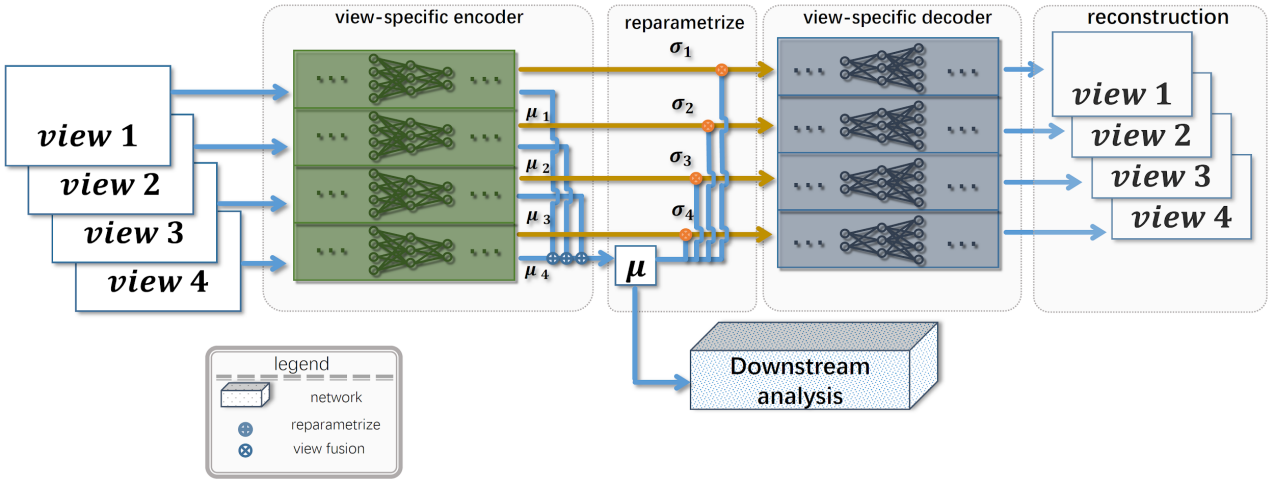


Figure 7. The structure of the proposed multimodal data integration deep neural network. For each view, an independent variational graph auto-encoder (VGAE) is used. We enforce the multi-view VGAE to learn a unified representation, and allow the *sigma* parameter to be different. The reparameterization is based on the shared *mu* and view-wise sigma. *mu* is used for the downstream analysis.

In a third study, Zhang et al. (2021c) addressed the 'incomplete multimodal brain data problem' - whereby a predictive machine learning model must be trained on diverse datasets that each may lack some data subtypes. They proposed to learn the modal-wise representations and synthesize the targets accordingly. A surrogate sampler was derived to generate target representations from incomplete observations; these were then used to design an interpretable attention-redistribution network. Experiments were conducted on ADNI MRI and PET data. The new method can handle different missing data scenarios and outperforms prior methods consistently.

3. AI to Discover Imaging Features that Mediate the Effect of Alzheimer's Disease Polygenic Risk on Clinical Decline. The Shen lab (U Penn) pioneered the application of a causal mediation model for AD imaging genetics studies, which can detect features in brain scans that mediate the effect of polygenic AD risk on clinical outcomes. In Eng et al. 2020, they performed a polygenic mediation analysis in an amyloid imaging genetic study of AD, and identified multiple imaging mediators linking genetic variants and polygenic risk scores (PRS) to AD outcomes. In the AI4AD initiative, the same strategy will be used to prioritize and identify imaging biomarkers (e.g., multimodal MRI and PET measures) that can best explain the association between sets of genetic variants and standard measures of clinical disease burden. This method will be integrated with deep learning to detect biomarker patterns in brain scans that mediate (explain) effects of specific sets of genomic features on clinical disease burden, enabling disease subtyping and genome-to-brain mapping to understand previously unknown effects of risk genes and deconvolute mechanistic complexity of the AD phenotypic outcomes.

4. Drug Repurposing. The Jun Lab (Boston University) is building a bioinformatics system that encodes and learns from results from genomics (ADSP and other GWAS data) and multi-omics studies of AD risk as well as imaging and fluid biomarkers to determine relations between druggable targets, genomic patterns, and signature profiles - including transcriptomic, proteomic, methylomic, and metabolomic data - and multimodal brain imaging. Leading our AI4AD Drug Repurposing Core, which is coordinated with ADSP Functional Genomics efforts including AMP-AD, Dr Jun's team has been adding existing data from drug informatics databases to her PSBO system (relating *predictors, signatures, biomarkers, and outcomes*; see **Figure 8**) to identify and prioritize drug candidates by storing information on their known effects on biomarkers.

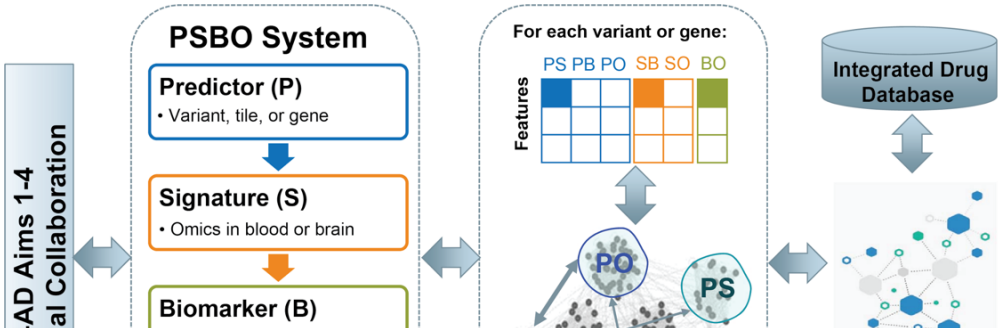


Figure 8. Concept of the PSBO target prioritization system and network-based drug repurposing. Association results in the PSBO system are shown as

relationships: PO: predictor-outcome; PB: predictor-biomarker; PS: predictor-signature; SB: signature-biomarker; SO: signature-outcome; BO: biomarker-outcome.

In silico trials are also possible by determining downstream effects of a change in a biomarker, or effects in a population subgroup, such as those with a protective allele. The PSBO system is aiding Dr Jun's work on APOE-mediated drug candidates by discovering modulators of the tau and complement cascade pathways (Patnich 2021; Jun 2021). Dr Jun's team is working together with other members of AI4AD to determine AD gene-centric networks enriched with AD related signatures, biomarkers, and clinical/ neuropathological outcomes to match to existing drugs applicable in subgroups of AD patients influenced by APOE genotypes. Given our AI methods for patient subtyping based on imaging, the networks in the PSBO system will be used to search for omics signatures that are characteristic of specific imaging-defined AD subtypes, and in patients with decline in specific cognitive domains. We will soon be adding our brain biomarker GWAS summary statistics (for MRI, PET, and white matter hyperintensities) as well as metabolomic analysis results from ADNI to the PSBO system.

C. Conclusions

Although active for only 4 months, our AI4AD Consortium has made progress in AI-assisted diagnosis of AD based on neuroimaging, and is starting work on AD subtyping and clustering, with methods that can already predict neuropathology and cellular hallmarks of AD in the ADNI and NACC cohorts. The efforts to use brain imaging along with AI to predict clinical outcomes will be assisted by the coming year's work on co-calibration of common cognitive domains across ADSP cohorts, in partnership with the ADSP Harmonization Core. On the whole-genome side, work has begun on decomposing the genome into units (tiles) that can be readily input into AI algorithms for diagnosis, prognosis, and subtyping.

Subtyping the heterogeneity of AD is the basis of the now widely-accepted ATN(V) system; AI methods are being developed to identify and better define clusters of patients with distinct pathology and outcomes, who then become more homogeneous groups in whom to assess interventions, and for clinical trial stratification. For AI methods to work well on unseen data, we are testing generative adversarial networks (GANs), transfer learning, domain adaptation, and related innovations in AI, to transfer successful predictive algorithms to datasets collected with different protocols.

D. Diversity, bias, and inclusion

Diversity, bias and inclusion are crucial to consider when developing AI methods that must work for all individuals, regardless of sex and race/ethnicity: methods trained on one cohort may not generalize well to others. Our most powerful genome-wide screen of 70,000 brain MRIs with Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) and Enhancing Neuro Imaging Genetics through Meta-Analysis (ENIGMA) consortia recently yielded a brain-related polygenic score that predicts ~4% of the variation in hippocampal morphometry in Whites, but only 2-3% in non-Whites, mandating greater inclusion of non-Whites. As ADSP is finding ancestry-dependent effects of AD genetic risk factors such as *APOE*, we will start to adapt and test our AI methods using the racially and ethnically diverse WHICAP dataset, expanding to more diverse cohorts in future years, in close partnership with other ADSP efforts.

Bibliography

Bashyam VM, Erus G, Doshi J, Habes M, Nasralah I, Truelove-Hill M, Srinivasan D, Mamourian L, Pomponio R, Fan Y, Launer LJ, Masters CL, Maruff P, Zhuo C, Völzke H, Johnson SC, Fripp J, Koutsouleris N, Satterthwaite TD, Wolf D, Gur RE, Gur RC, Morris J, Albert MS, Grabe HJ, Resnick S, Bryan RN, Wolk DA, Shou H, Davatzikos C. MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. **Brain**. 2020 Jul 1;143(7):2312-2324. doi: 10.1093/brain/awaa160.

Bashyam VM, Doshi J, Erus G, Srinivasan D, Abdulkadir A, Habes M, Fan Y, Masters CL, Maruff P, Zhuo C, Völzke H, Sterling C. Johnson SC, Jurgen Fripp J, Nikolaos Koutsouleris N, Theodore D. Satterthwaite TD, Wolf D, Gur RE, Gur RC, Morris J, Albert MS, Grabe HJ, Resnick S, Bryan RN, Wolk DA, Shou H, Ilya M. Nasrallah IM, Davatzikos C. Medical Image Harmonization Using Deep Learning Based Canonical Mapping: Toward Robust and Generalizable Learning in Imaging. submitted, 2021. Preprint available at ***arXiv***: https://arxiv.org/abs/2010.05355.

Eng Y, Yao X, Liu K, Risacher SL, Saykin AJ, Long Q, Zhao Y, Shen L, for the ADNI. Polygenic mediation analysis of Alzheimer's disease implicated intermediate amyloid imaging phenotypes. **AMIA 2020 Annual Symposium**, pp 422-431, Virtual Conference, November 14-18, 2020.

Ficiarà E, Crespi V, Gadewar SP, Thomopoulos SI, Boyd J, Thompson PM, Jahanshad N, Pizzagalli F, and the Alzheimer's Disease Neuroimaging Initiative. Predicting Progression from Mild Cognitive Impairment to Alzheimer's Disease using MRI-based Cortical Features and a Two-State Markov Model, **ISBI 2021**, accepted.

Gupta U, Pradeep K. Lam, Greg Ver Steeg, Paul M. Thompson. Improved Brain Age Estimation With Slice-based Set Networks, **ISBI 2021**, accepted.

Jun GR, You Y, Zhu C, Meng G, Chung J, Panitch R, Hu J, Xia W, the Alzheimer's Disease Genetics Consortium, Bennett DA, Foroud TM, Wang L-S, Haines JL, Mayeux R, Pericak-Vance MA, Shellenberg GD, Au R, Lunetta KL, Ikezu T, Stein TD, Farrer LA. Protein phosphatase 2A, complement component 4, and APOE genotype linked to Alzheimer's disease using a systems biology approach. ***medRxiv*** 2020.11.20.20235051.

Lam P, Alyssa H. Zhu, Iyad Ba Gari, Neda Jahanshad, Paul M. Thompson. 3D Grid-attention Networks For Interpretable Age And Alzheimer's Disease Prediction From Structural MRI, submitted to **ISBI 2021**.

Panitch R, Hu J, Chung J, Zhu C, Meng G, Xia W, Bennett DA, Lunetta KL, Ikezu T, Au R, Stein TD, Farrer LA, Jun GR. Integrative brain transcriptome analysis links complement component 4 and HSPA2 to the APOE ε2 protective effect in Alzheimer disease. **medRxiv** 2020.11.23.20235762.

Stripelis D, José Luis Ambite, Pradeep Lam, Paul Thompson. 2021 [submitted]

Zhang Y, Zhan L, Dodge H, Thompson PM, Huang H. 2021 [submitted].

Zhang Y, Zhan L, Paul Thompson PM, Huang H. 2021 [submitted].

Zhang Y, Zhan L, Thompson PM, Huang. 2021 [submitted].