DOI: 10.1002/alz70856_099373

BIOMARKERS

PODIUM PRESENTATION

Alzheimer's & Dementia®
THE JOURNAL OF THE ALZHEIMER'S ASSOCIATION

BIOMARKERS (NON-NEUROIMAGING)

Benchmarking the AI-based diagnostic potential of plasma proteomics for neurodegenerative disease in 17,710 people

Lijun An[1] | Yu Xiao[2] | Ines Hristovska[3] | Shinya Tasaki[4] | Bart Smets[5] | Ying Xu[6,7] | Varsha Krish[8] | Farhad Imam[8] | Erik Stomrud[9,10] | Shorena Janelidze[11] | Sebastian Palmqvist[9,10] | Alexa Pichet Binette[12,13,14] | Rik Ossenkoppele[15,16,17] | Niklas Mattsson-Carlgren[9,18] | Oskar Hansson[10,19] | Jacob W. Vogel[20]

[1]Department of Clinical Sciences Malmö, SciLifeLab, Lund Univerisity, Lund, Sweden

[2]Department of Clinical Sciences Malmö, SciLifeLab, Lund University, Lund, Sweden

[3]Clinical Memory Research Unit, Department of Clinical Sciences, Lund University, Lund, Sweden

[4]Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL, USA

[5]Janssen Pharmaceutica NV, a Johnson & Johnson company, Beerse, Belgium

[6]Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, USA

[7]NeuroGenomics and Informatics Center, Washington University School of Medicine, St. Louis, MO, USA

[8]Gates Ventures, Seattle, WA, USA

[9]Memory Clinic, Skåne University Hospital, Malmö, Skåne, Sweden

[10]Clinical Memory Research Unit, Department of Clinical Sciences Malmö, Lund University, Lund, Sweden

[11]Clinical Memory Research Unit, Department of Clinical Sciences Malmö, Faculty of Medicine, Lund University, Sweden, Lund, Sweden

[12]Department of Physiology and Pharmacology, Université de Montréal, Montréal, QC, Canada

[13]Clinical Memory Research Unit, Department of Clinical Sciences Malmö, Lund University, MOntreal, QC, Canada

Abstract

Background: Plasma proteomic biomarkers have shown significant potential for accurate and cost-effective dementia diagnosis. However, plasma proteomic has not been validated for multi-disease diagnosis in large neurodegenerative cohorts yet. This study develops AI models on data from the Global Neurodegeneration Proteomics Consortium (GNPC) to predict major forms of dementia, accounting for multiple underlying pathologies, and outputting probabilistic information.

Method: This study included 17,170 GNPC participants with SomaLogic 7K plasma proteomics, comprising controls and patients with AD, PD, FTD, ALS, or Stroke (Figure 1A). We describe 'ProtAIDe', a deep network for six clinical diagnostic categories classification using proteomics (Figure 1B). 10-fold cross-validation with built-in feature selection was used to evaluate overall performance. Leave-one-site-out scheme, with and without k-shot fine-tuning, was adopted to evaluate out-of-site generalization performance. Additionally, baseline proteomics embeddings from ProtAIDe's last layer were leveraged to predict longitudinal CDR progression (advancing from CDR 0 to >0). Predicted probabilities were validated against cognition and *APOE* genotype, and the cross-disease classification probability space was visualized using 2-dimensional t-SNE. Predictive importance of individual proteins was estimated by feature permutation.

Result: ProtAIDe achieved simultaneous balanced classification accuracy (BCA) >0.7 and AUC >0.8 across all six targets (Figure 1C) with ~200 proteins. Performance dipped when generalizing to unseen sites, though it was partially recovered by finetuning (Figure 2A). Despite being trained only on baseline data, ProtAIDe's baseline

[14]Centre de recherche de l'institut universitaire de gériatrie de Montréal (CRIUGM), Montréal, QC, Canada

[15]Amsterdam Neuroscience, Neurodegeneration., Amsterdam, Netherlands

[16]Alzheimer Center Amsterdam, Neurology, Vrije Universiteit Amsterdam, Amsterdam UMC location VUmc, Amsterdam, Netherlands

[17]Clinical Memory Research Unit, Lund University, Lund, Sweden

[18]Clinical Memory Research Unit, Lund University, Malmö, Skåne, Sweden

[19]Memory Clinic, Skåne University Hospital, Malmö, Sweden

[20]Department of Clinical Sciences Malmö, Faculty of Medicine, SciLifeLab, Lund University, Lund, Sweden

Correspondence
Lijun An, Department of Clinical Sciences Malmö, SciLifeLab, Lund Univerisity, Lund, Sweden.
Email: lijun.an@med.lu.se

embeddings predicted longitudinal CDR progression with an AUC of 0.76±0.09 (N = 2052; Figure 2B). AD classification probability showed a strong anti-correlation (R = -0.63) with MMSE score (Figure 2C) and reflected the protective and risk qualities of the *APOE* ε2 and ε4 alleles, respectively (Figure 2D). t-SNE visualization of probability spaces (Figure 2E) revealed unique (multiple PD and FTD clusters) and shared (middle regions shared by AD, Control, FTD, and Stroke) clustering patterns among neurodegenerative diseases, suggesting potential comorbidities. Key proteins emerged as major contributors to single- or multi-disease classifications (Figure 3).

Conclusion: Our results demonstrate that ~200 key proteins in blood can differentiate major forms of dementia at the patient-level. ProtAIDe's probabilistic information highlights its potential for identifying co-pathologies and determining proteins driving symptoms at the individual-level.

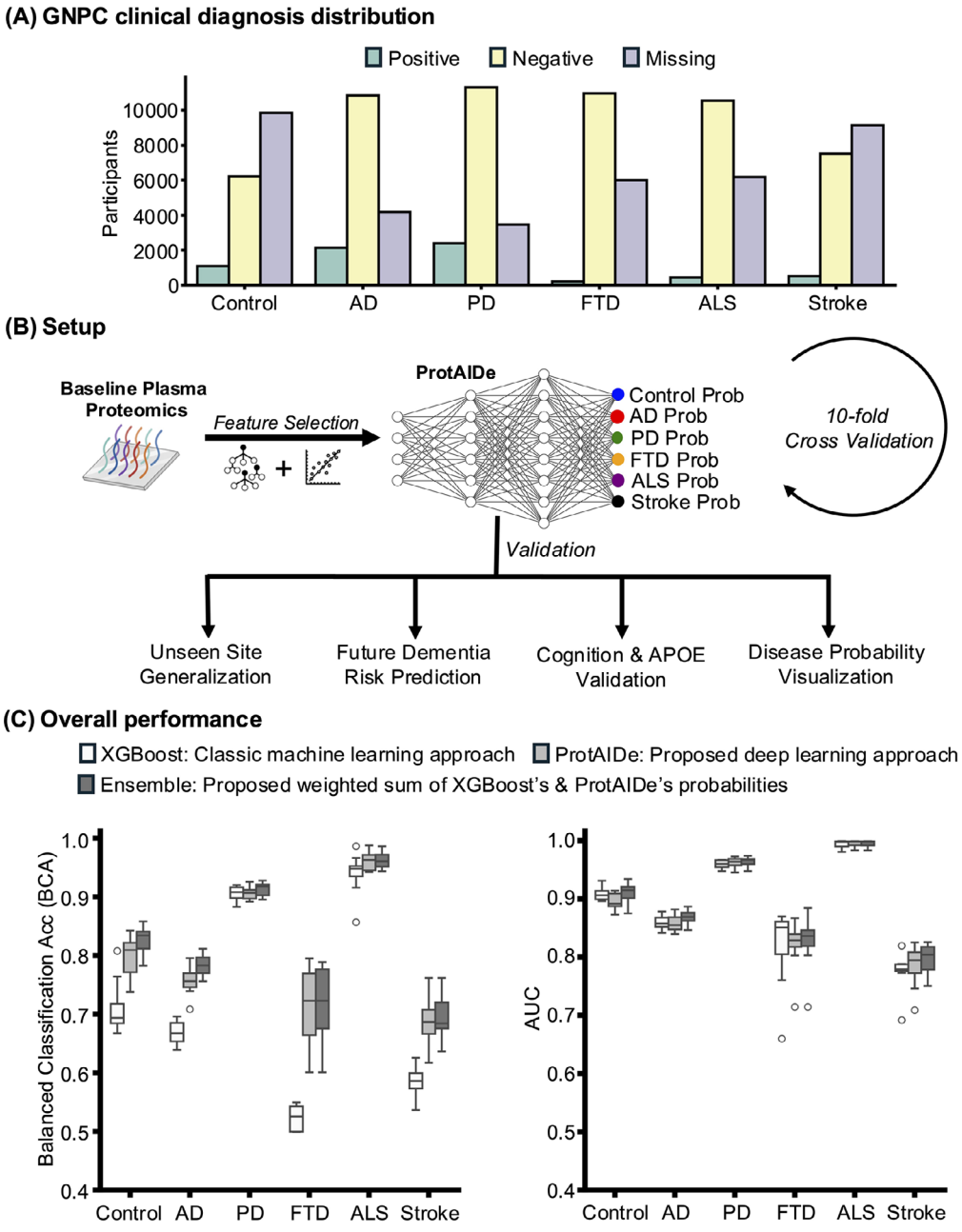

Figure 1. Clinical diagnosis distribution, project setup and overall performance. (**A**) Clinical diagnosis distributions of GNPC participants. (**B**) Baseline SomaLogic 7k proteomics (N=17,770) underwent feature selection using XGBoost and GLM before input into ProtAIDe models, a deep network for multi-disease classification. ProtAIDe outputs probabilities for six targets, following a 10-fold cross-validation scheme. ProtAIDe was further validated by generalizing to unseen sites, predicting longitudinal neurodegeneration risk, correlating biomarkers with output probabilities, and clustering patient on probability spaces. **(C)** ProtAIDe performance: Box plots show Balanced Classification Accuracy (BCA, Y axis of left panel) and Area Under the Curve (AUC, Y axis of right panel) for 10 test folds across six targets (X axis).

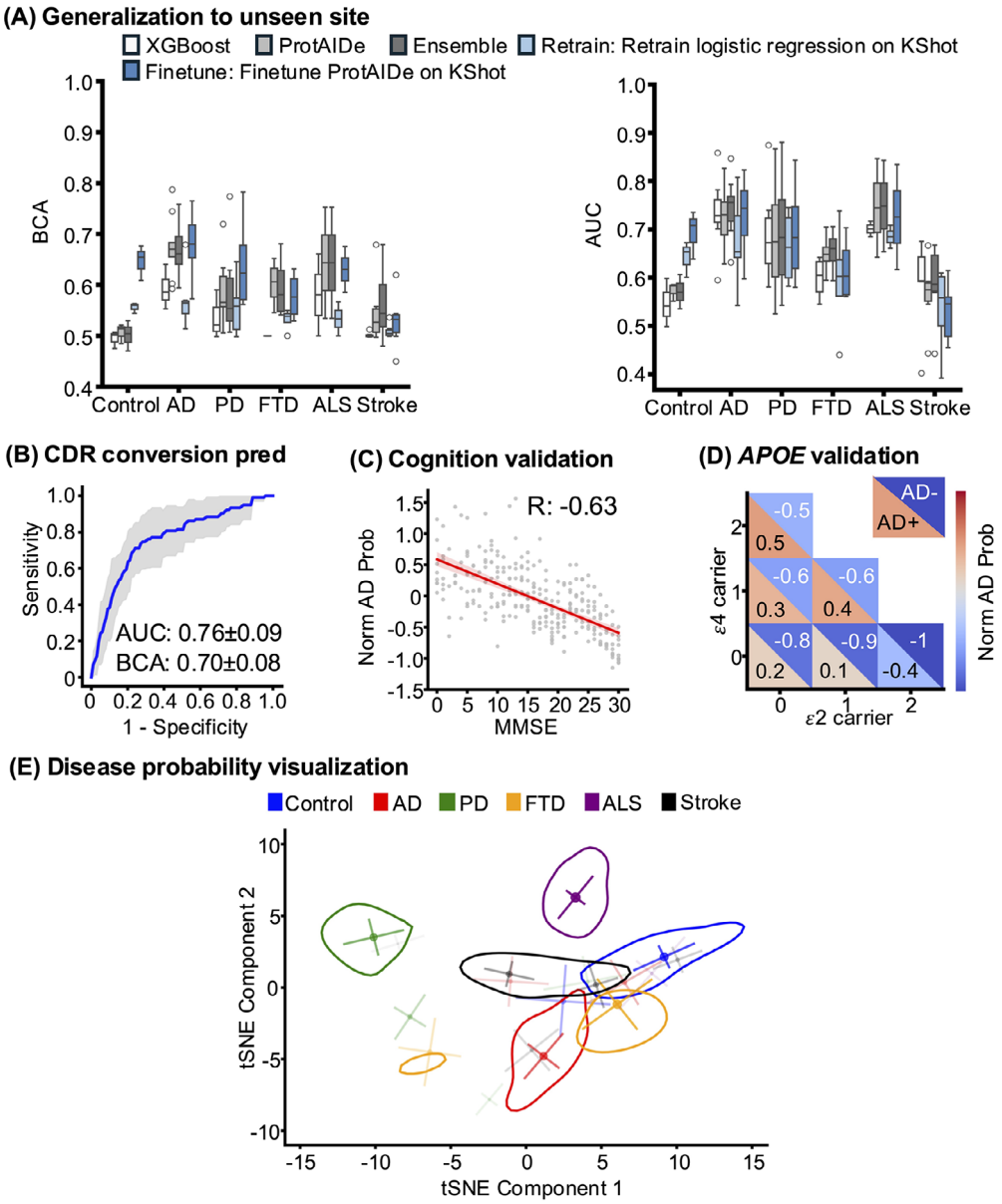

Figure 2. Validation of ProtAIDe models. (**A**) Generalization performance on unseen sites using leave-one-site-out scheme. Y-axis: metric scores (BCA for Y axis of left panel, AUC for Y axis of right panel); X-axis: results for each target (dots in each box represent unseen site results). (**B**) AUC for predicting progression from baseline CDR 0 (N=2,052) to CDR 1/2/3 (N=107) or non-progression (N=1,945), using ProtAIDe's baseline protein embeddings as input. (**C**) Correlation of normalized AD probabilities with MMSE scores. Y-axis: log of AD probabilities divided by thresholds; X-axis: MMSE scores. Dots represent mean normalized AD probabilities for participants with same MMSE score from same test fold. (**D**) Normalized AD probabilities by *APOE* $\varepsilon 2/\varepsilon 4$ carriers. Y-axis: $\varepsilon 4$ carriers; X-axis: $\varepsilon 2$ carriers. Lower triangles: mean normalized AD probabilities for AD-positive patients; upper triangles: AD-negative patients' mean normalized AD probabilities. (**E**) Two-dimensional summary of classification probabilities across all diseases, which distributes and clusters patients based on the protein-based likelihood of having each disease. Dots and crosses indicate KMeans cluster centers and standard deviations running within each target; dot sizes and color shades reflect cluster size.

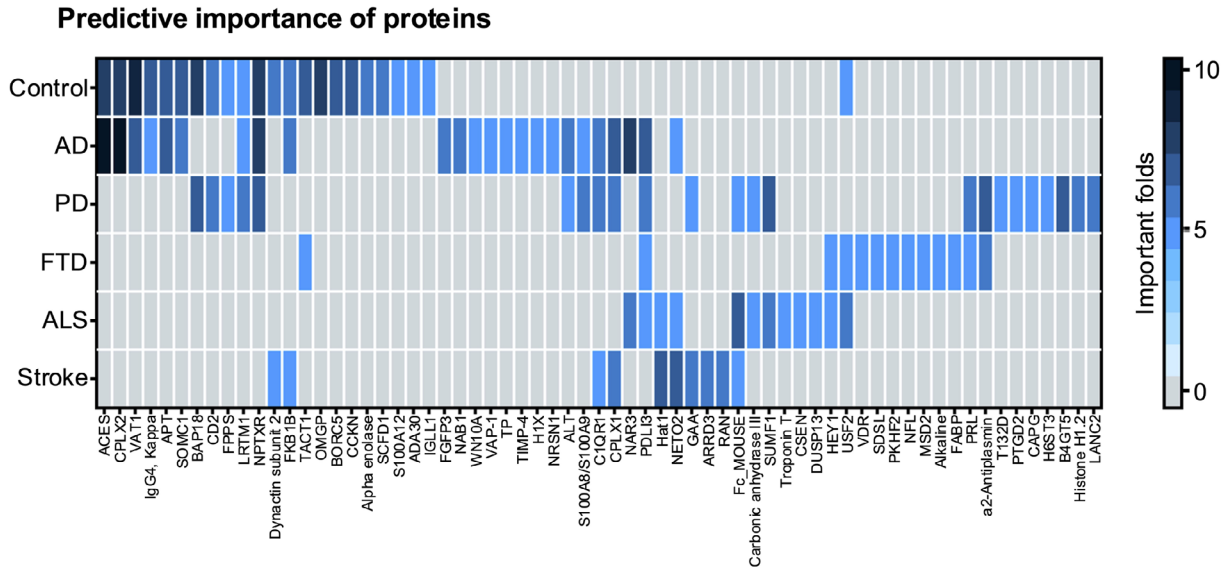

Figure 3. Feature importance of ProtAIDe models. Features importances of ensemble models by random feature permutations. If an input protein is randomly permuted and then significantly affects model performance, this protein is important. The procedure was repeated across 10 folds. Row of heatmap indicates targets to predict, columns indicate the input proteins, color represents in how many folds a protein is important for predicting one target. Only proteins important in at least 5 folds were visualized.