

<https://doi.org/10.1038/s44400-025-00053-9>

Plasma proteomic signatures associate with near-future Alzheimer's disease dementia conversion in mild cognitive impairment patients

Check for updates

Jingyue Xi¹, Karen Y. He¹, Liping Hou¹, Bart Smets², Silke Miller³, Ziad S. Saad⁶, Christopher D. Whelan⁴, Ruiz Laza Agustin⁵, Raquel Puerta⁶, Amanda Cano^{6,7}, Cabrera-Socorro Alfredo⁸, Shuwei Li¹ & Yanfei Zhang¹ ✉

Predicting Alzheimer's Disease (AD) dementia conversion from mild cognitive impairment (MCI) is crucial for therapeutic strategies. Plasma proteomics offers a powerful approach for biomarker identification and predictive modeling. Using SomaScan plasma proteomics in F.ACE cohort, we identified 77 somamers significantly associated with dementia conversion, encompassing immune, inflammatory and neurological processes. Thirteen were replicated in EMIF-AD MBD validation cohorts, including SMOC1. Leveraging machine learning techniques, we developed an optimal model, integrating demographics and 48 proteins, to predict near-future dementia conversion from MCI. It achieved a concordance index of 0.69 and a median time-dependent AUC of 0.75, with comparative performance in A+/T+ subgroups. The 48-protein panel also showed comparative performance with CSF pTau-181. Model can effectively stratify patients by inferred risk. Independent validation evaluated its generalizability, with variant performance decline. These findings demonstrate plasma proteomics' potential for biomarker discovery and risk prediction, facilitating precision strategies for drug development.

Dementia is a symptom of a spectrum of neurological conditions that progressively impair cognitive function and daily activities, affecting approximately 55 million people globally¹. Alzheimer's disease (AD) is the most prevalent causes of dementia, accounting for 60% to 80% of cases².

AD is characterized by abnormal extracellular beta-amyloid (A β) plaques and neurofibrillary tangles in the brain³. Recent research, particularly genome-wide association studies utilizing biobanks and disease cohorts, has significantly broadened our understanding of AD, identifying additional mechanisms that contribute to its risk and progression^{4–6}, reinforcing the multifactorial nature of this complex disease. This view is further supported by proteogenomic studies and molecular subtyping efforts^{7,8,9}.

The mechanistic heterogeneities of AD lead to substantial phenotypic variability. Alzheimer's Association Workgroup has revised diagnostic and staging criteria for AD by integrating both clinical and biological assessment address such variabilities¹⁰. The new framework introduced biological stages—initial, early, intermediate and advanced—based on presence of AD core biomarkers. The clinical staging remains the same: preclinical stage is when people have AD pathology but remain cognitively unimpaired, then progress to mild cognitive impairment (MCI) stage defined by cognitive impairment without significant impact on daily function. Eventually, patients progress to mild, moderate and severe dementia where functional impairment is significant. The new framework allows for phenotypic variabilities such that a person may have early biological stage AD but more

¹Population Analytics & Insights, AI/ML, Data Science & Digital Health, Johnson & Johnson, Spring House, PA, USA. ²Neuroscience Data Science, Johnson & Johnson, Beerse, Belgium. ³Neuroscience Biomarker, Johnson & Johnson, La Jolla, CA, USA. ⁴RCSI University of Medicine and Health Sciences, Dublin, Ireland. ⁵Glenn Biggs Institute for Alzheimer's & Neurodegenerative Diseases and Department of Microbiology, Immunology and Molecular Genetics, University of Texas Health Science Center, San Antonio, TX, USA. ⁶ACE Alzheimer Center Barcelona, Universitat Internacional de Catalunya, Barcelona, Spain. ⁷Centro de Investigación Biomédica en Red de Enfermedades Neurodegenerativas (CIBERNED), Instituto de Salud Carlos III, Madrid, Spain. ⁸Neuroscience Discovery, Johnson & Johnson, Beersa, Belgium. ✉e-mail: yzhan378@its.jnj.com

advanced clinical stage than expected, as in cases with mixed pathology. Alternatively, a person could have more advanced biological stage AD but have less clinical impairment than expected, potentially carrying some resilience to the AD pathophysiology.

Annual conversion rates from MCI due to AD to AD dementia are estimated at 10–15%^{11–13}. Over a six-year follow-up period, the conversion rate rises to about 80%, although some individuals may remain stable^{14–16}. This variability in phenotypes and progression rates underscores the necessity for identifying novel biomarkers that capture additional pathways to predict disease progression.

Blood plasma serves as an accessible and dynamic source of biomarkers that reflects the physiological and pathological states of individuals¹⁷. The successful development of plasma pTau-217 for detecting AD pathology in patients with MCI, with consistent and high performance^{18–20}, has made plasma biomarkers for neurological diseases increasingly promising. Plasma proteomics, which can measure thousands of proteins simultaneously, has emerged as a powerful tool for biomarker discovery²¹. Several studies have utilized biobank or cohort-based plasma proteomics to identify plasma signatures associated with the future onset of AD or dementia^{22–24}. However, due to sampling limitations, these investigations have primarily focused on identifying risk proteins in middle-aged adults and often employed a cross-sectional design or did not incorporate temporal factors in their prediction modeling. While these studies have pinpointed proteins that change decades before AD onset, their role in the progression from MCI to dementia remains unclear, limiting their translational relevance.

In this study, we leveraged proteomics data from two independent at-risk AD cohorts and investigated the MCI-to-dementia conversion. Most conversion cases were due to AD, although a few were attributed to other types of dementia. We leveraged machine learning techniques to (1) identify and characterize the plasma proteins associated with near-term MCI-to-dementia conversion, and (2) develop a parsimonious prognostic model to predict MCI to dementia conversion while accounting for temporal dynamics. Our findings support progression-associated biomarker discovery and show potential to facilitate AD drug discovery and development.

Results

Characteristics of the study cohorts

The overall workflow is illustrated in Fig. 1. The study included 925 individuals from three datasets across two cohorts: Fundació ACE (F.ACE) as the discovery cohort and two datasets from the EMIF-AD MBD cohort as the validation cohort.

Table 1 lists the characteristics of the study cohorts. The discovery cohort is comprised of 755 individuals with MCI (mean age 72 years, 46% male, follow-up duration range: [0.1, 7.6] years), of whom 394 subjects (52.19%) were amyloid positive at baseline (Supplementary Table 1A), and 368 (48.7%) progressed to dementia during the follow-up period

(Supplementary Fig. 1A, 1B). Kaplan-Meier (KM) curves shows that 50% of participants in F.ACE are expected to experience conversion at 4 years post-baseline (Supplementary Fig. 1A). The EMIF-AD MBD Dataset 1 includes 60 individuals with follow-up durations ranging from 0.4 to 7.0 years. Of these, 57 subjects (95%) were amyloid positive at baseline (Supplementary Table 1A), and 34 individuals (56.7%) experienced MCI-to-dementia conversion (Table 1, Supplementary Fig. 1C–D). EMIF-AD MBD Dataset 2 includes 110 individuals and follow-up durations ranged from 0.4 to 4.9 years. Baseline amyloid was observed in 68 participants (61.8%), and 17 individuals (15.4%) progressed to dementia. KM curves for both datasets are presented in Supplementary Fig. 1B, showing that 50% of participants in Dataset 1 are statistically expected to experience dementia conversion at 2.0 years post-baseline, while 50% of participants in Dataset 2 are expected to experience conversion at 2.9 years post-baseline. In both cohorts, the majority of dementia conversion cases were due to AD (Supplementary Table 1B).

Plasma proteins associate with MCI-to-dementia conversion

To identify proteins associated with the MCI-to-dementia conversion, in the F.ACE discovery cohort, we performed bootstrapped univariate Cox proportional hazards (PH) analyses on the 3006 somamers corresponding to human proteins that passed quality control in F.ACE cohort and were measured in both discovery and validation cohorts. This approach incorporated subsampling of events and censoring to ensure a robust set of candidate protein features for predictive modeling (Supplementary Table 2, see details in “Materials & Methods”).

We identified 77 somamers corresponding to 74 unique proteins significantly associated with MCI-to-dementia conversion (Supplementary Table 3). Of these, 67 somamers were linked to an increased risk of conversion, while 10 somamers were associated with a decreased risk of conversion (Fig. 2A, Supplementary Fig. 2A). Some proteins are known to be involved in immune and inflammatory signaling (e.g., CXCL9, CXCL11, TNFSF11B, SERPINA3, SERPINA1, S100A2, SLP1, etc.) as well as neuronal functions (e.g., CBLN1, LRRN1, NTN1, CHRDL1). Notably, several proteins have been previously reported in association with AD or other forms of dementia, including GDF15^{22,23}, SERPINA11²², and CA6^{25,26}. Eight proteins (PTN, SMOC1, TNFRSF1A, COL18A1, HNF1A, NTN1, SERPINA3, TNFRSF11B) were nominated as potential AD drug targets by the AMP-AD consortium²⁷. Five proteins (SMOC1, TREM1, LRRN1, SERPINA11, TIMP4, Supplementary Table 4) were suggested to be associated with AD disease risk by previous studies utilizing Mendelian randomization (MR), a widely-used approach for causal inference, or protein-wide association studies (PWAS), using protein quantitative trait loci (pQTLs) identified from the CSF⁷ and plasma²⁸. The dynamic expression pattern of TREM1 in CSF and plasma was also reported as being associated with AD progression^{29,30}.

Fig. 1 | Study workflow. The F.ACE longitudinal cohort served as the discovery cohort, where people diagnosed with MCI were followed up to monitor dementia conversion. Plasma proteomics were generated from the collected plasma samples at baseline. Proteomic data from the F.ACE discovery cohort underwent quality control (QC), normalization, and marginal screening within a predefined training set. A separate testing set was held out for evaluating selected features. Feature selection was performed using Lasso, RSF minimal depth, and recursive feature elimination (RFE), followed by model training using survival machine learning methods: survival gradient boosting (SGB), random survival forests (RSF), and survival support vector machines (sSVM). External validation was performed using two datasets from the EMIF-AD MBD cohort.

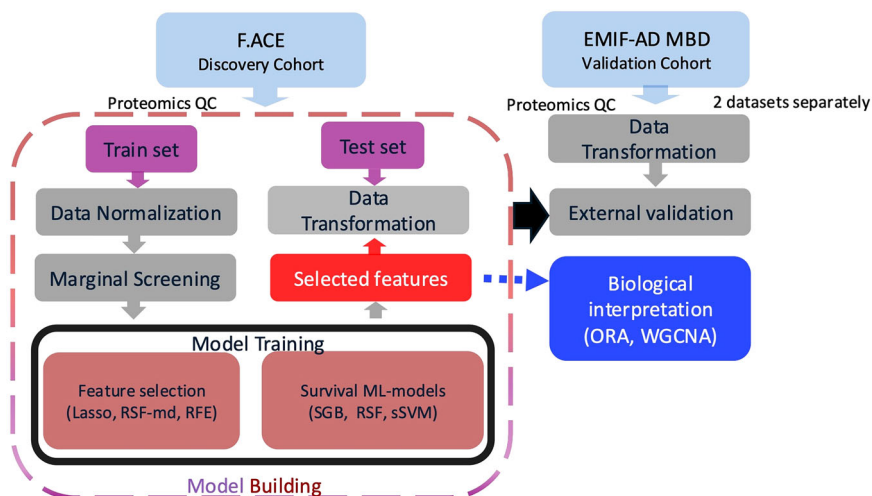


Table 1 | Characteristics of the study cohorts

| COHORT | COHORT CHARACTERISTICS | | | |
|-----------------------|---|------------------|------------------|-----------------------|
| | | Overall | MCI to dementia | MCI at last follow-up |
| FACE | <i>n</i> | 755 | 368 | 387 |
| | Age, mean (s.d.) | 72.02 (8.94) | 74.68 (7.56) | 69.51 (9.40) |
| | Sex, male (%) | 347 (46.0) | 163 (44.3) | 184 (47.6) |
| | APOE4 carrier, E4 carrier (%) | 263 (35.6) | 158 (43.4) | 105 (28.1) |
| | Baseline MMSE, mean (s.d.) | 25.61 (3.22) | 24.55 (3.30) | 26.64 (2.79) |
| | CSF A β 42 /A β 40, mean (s.d.) | 0.07 (0.02) | 0.05 (0.02) | 0.08 (0.02) |
| | CSF pTau-181, mean (s.d.) | 69.43 (41.29) | 85.72 (45.91) | 53.94 (28.84) |
| | CSF total Tau, mean (s.d.) | 442.41 (282.44) | 548.44 (316.79) | 341.59 (198.75) |
| | Total follow-up years, median (IQR) | 2.52 (1.18–4.14) | 1.57 (1.02–2.64) | 3.68(2.35–4.81) |
| EMIF-AD MBD, Dataset1 | <i>n</i> | 60 | 34 | 26 |
| | Age, mean(s.d.) | 71.30 (7.65) | 71.13 (8.22) | 71.53 (6.84) |
| | Sex, male (%) | 27 (45.0) | 17 (50.0) | 10 (38.5) |
| | APOE4 carrier, E4 carrier (%) | 40 (66.7%) | 24 (70.6) | 16 (61.5) |
| | Baseline MMSE, mean (s.d.) | 25.05 (2.66) | 24.59 (2.84) | 25.65 (2.31) |
| | CSF A β 42 /A β 40, mean (s.d.) | 0.04 (0.01) | 0.04 (0.01) | 0.04 (0.02) |
| | CSF pTau, mean (s.d.) | 83.51 (37.75) | 90.44 (35.59) | 74.44 (39.25) |
| | CSF total Tau, mean(s.d.) | 606.47 (365.64) | 694.82 (422.97) | 490.95 (234.21) |
| | Total follow-up years, median (IQR) | 1.92 (1.03–2.00) | 1.52 (1.03–1.97) | 1.97 (1.53–2.02) |
| EMIF-AD MBD, Dataset2 | <i>n</i> | 110 | 17 | 93 |
| | Age, mean (s.d.) | 71.46 (7.40) | 74.97 (7.41) | 70.82 (7.21) |
| | Sex, male (%) | 47 (42.72) | 5 (29.4) | 42 (45.2) |
| | APOE4 carrier, E4 carrier (%) | 49 (44.5) | 9 (52.9) | 40 (43.0) |
| | Baseline MMSE, mean (s.d.) | 25.53 (2.94) | 25.53 (2.94) | 26.76 (2.55) |
| | CSF A β 42 /A β 40, mean (s.d.) | 0.06 (0.02) | 0.05 (0.02) | 0.06 (0.02) |
| | CSF pTau, mean (s.d.) | 68.18 (35.86) | 87.74 (38.66) | 64.60 (34.35) |
| | CSF total Tau, mean(s.d.) | 460.20 (311.07) | 670.52 (363.98) | 421.76 (286.25) |
| | Total follow-up years, median (IQR) | 1.98 (1.45–2.02) | 2.12 (1.05–2.36) | 1.98 (1.46–2.01) |

To gain a deeper understanding of the biological significance of the identified proteins, we conducted pathway enrichment and protein-protein interaction analyses, and tested their association with CSF biomarkers, including CSF pTau-181, total Tau (tTau) and A β 42. The 74 unique proteins were predominantly enriched in GO terms related to the external encapsulating structure, extracellular matrix (ECM), and ligand-receptor activity (FDR < 0.05; Fig. 2B, Supplementary Table 7), suggesting a role of ECM in the MCI to dementia progression. Protein-protein interaction analysis on the 74 proteins revealed clusters of serine protease inhibitor and neutrophil chemotaxis, suggesting inflammation, tissue remodeling processes during dementia progression (Supplementary Fig. 2B). Fifty-five proteins were significantly associated with AD CSF biomarkers, while 19 were not (Fig. 2C, Supplementary Table 5). The 19 proteins not associated with A β and Tau are enriched in terms related to regulation of endocrine process, primarily driven by INHBA/INHBB and RETN (Supplementary Fig. 2C, Supplementary Table 6). In addition, some proteins are implicated in inflammation (TNFSF1A, SLPI, C9, CCL23) and tissue remodeling (MMP12 and COL18A1), suggesting other processes underlying dementia conversion. These results suggest that plasma proteomics can capture additional biological signals beyond the A β and Tau pathways.

We further validated the associations of the 74 proteins (77 somamers) in the two EMIF-AD MBD datasets (Supplementary Table 8). In EMIF-AD MBD Dataset 1, 10 proteins (11 somamers) were significantly associated with time to MCI-to-dementia conversion with the same directions of effect as in the discovery cohort, including SMOCI(seq.13118.5) and DNAJB12, etc. In EMIF-AD MBD Dataset 2, only three proteins (RNF148, QPCTL,

MCCD1) were replicated. No protein was replicated in both datasets of EMIF-AD MBD, which may be due to the cohort differences and small sample size.

Protein co-expression signatures associated with AD-type dementia

Next, we employed weighted gene co-expression network analysis (WGCNA; see Materials and Methods)³¹, a data-driven method for identifying protein modules based on co-expression patterns. To ensure unbiased estimation, we included the 7k panel proteomics data from all individuals including those who did not have MCI at baseline in the F.ACE cohort. We assessed the association of these modules with clinical traits, including CSF biomarkers (pTau, tTau and A β 42), APOE4 genotypes, MCI-to-dementia conversion status, time to conversion and demographic factors (Supplementary Fig. 3A, Supplementary Table 9).

We identified 12 protein co-expression modules (Supplementary Fig. 3A). A large portion of proteins (3771, 64.11%) lacking a clear co-expression pattern were assigned to the grey module. A large grey module is not uncommon in proteomic studies^{32–34}. The diverse sources of proteins in plasma, the indirect tissue relevance to neurological diseases and the non-complete coverage of targeted proteomics technology all contribute to the weak or absent co-expression signals for many proteins.

To understand the biological signals captured by these modules, we performed eigen-protein association analyses with clinical traits and pathway enrichment analyses. Five modules showed significant correlations with AD CSF biomarkers (Fig. 3A). Of the five, two modules—blue and magenta

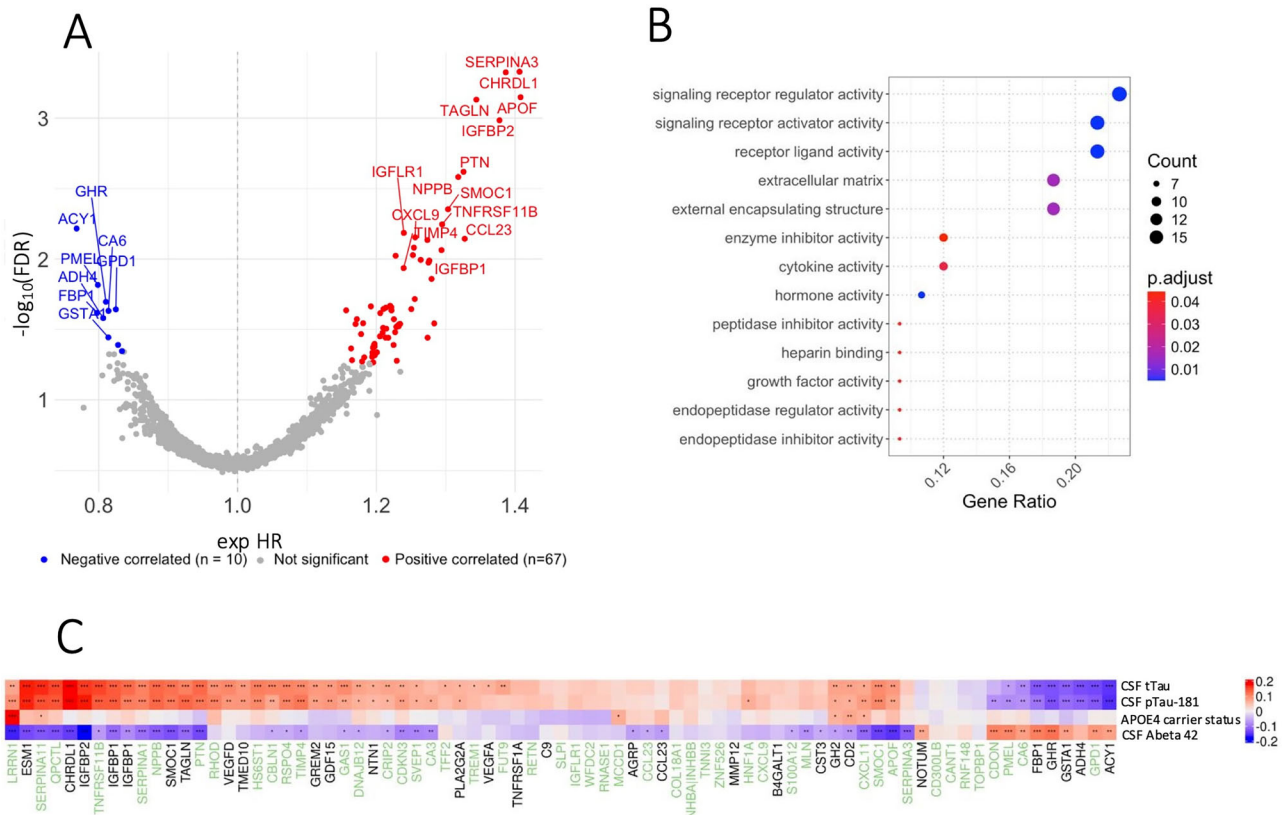


Fig. 2 | Plasma proteins associated with time to dementia conversion. **A** Volcano plot of Cox PH regression showing exponential hazard ratio (HR) and the overall *p*-value for 3006 proteins in F.ACE cohort. Each dot indicates a protein somamer. Red indicates the protein somamer is positively correlated with MCI-dementia conversion risk; blue indicates reverse correlation, and grey indicates no significant correlation. The top proteins are labeled. **B** Enriched Cellular Component Gene

ontology (GO) terms identified from the overrepresentation analysis (ORA) of 74 marginal screened proteins. **C** Heatmap showing the Spearman's correlation between 77 somamers (corresponding to 74 proteins) and AD CSF biomarkers (CSF tTau, pTau-181, APOE4 carrier status, CSF Aβeta 42) as well as APOE4 carrier status for F.ACE discovery cohort. Protein names highlighted in green represent members of the 48-protein panel.

—are identified as being associated with MCI-to-dementia conversion. The blue module, also referred to as the neuronal module, consists of 191 proteins and is significantly enriched in pathways related to axon development, neuron projection guidance, and Wnt/Ephrin signaling (Supplementary Table 10 and Supplementary Fig. 3B). This module is associated with a higher risk of progressing from MCI to dementia (*p*-value = 0.001, Fig. 3A). Conversely, the magenta module, defined as the metabolic module and comprising 32 proteins, is primarily enriched for amino-acid metabolism, extracellular vesicle and exosome pathways, and oxidoreductase activity, which represent metabolic and detoxification processes (Supplementary Table 10, Supplementary Fig. 3B). The metabolic module is related to a lower risk of conversion (*p* = 2e-04, Fig. 3A). The 77 MCI-to-dementia conversion associated somamers appeared only in grey (54, 70%), blue (18, 23%) and magenta (5, 7%) modules (Fig. 3A, Supplementary Table 12), further highlighting the importance of blue and magenta modules in MCI-to-dementia conversion. The red module, showing inverse correlations with CSF tTau, is enriched for RNA metabolism, nuclear processes and transcription elongation, implying transcriptional and RNA regulatory activity involved in dementia progression. Only the purple module significantly associated with copies of APOE4 allele (*p* = 5e-26). No significant pathway enrichment was observed for the black or purple module.

Additionally, we identified 150 driver proteins with high module membership (MM; see Materials and Methods) across the five significant modules (Supplementary Table 11). These proteins have strong connectivity within each module and may represent key proteins in the co-expression structure driving the association with AD CSF biomarkers and MCI-to-dementia conversion. Indeed, 7 out of the 9 driver proteins identified from the 77 somamers showed significant correlations with AD CSF biomarkers,

and two with APOE4 genotypes, and all nine were significantly correlated with MCI-to-dementia conversion (Fig. 3B-F). The driver protein that most strongly correlated with MCI-to-dementia conversion and CSF biomarkers was SMOC1 in the blue module, underscoring its central role in mediating MCI-to-dementia conversion and AD pathology.

Identifying the “best model” predicting MCI-to-dementia conversion

To evaluate the potential applicability of the identified proteins for predicting the near-future MCI-to-dementia conversion, we built and evaluated three models utilizing different algorithms and feature selection methods via a 5-fold cross-validation (CV) approach in the F.ACE discovery cohort (see Materials and Methods). Three prediction models with different feature sets were constructed: demographics model (model 0); proteomics model (model 1); demographics and proteomics model (model 2).

We first identified the optimal model for each of the three feature sets separately using fivefold cross validation (Fig. 4A) and then evaluated them on the testing set using C-index and median dynamic AUCs (Table 2). Evaluation metrics of the three models with different feature sets using three feature selection techniques and three ML algorithms were compared to identify the “best” model. Full CV results of feature selection and ML algorithms are provided in Supplementary Fig. 4A. Overall, among all model combinations, the demographics and proteomics model (model 2), which includes 48 proteins, age and sex, achieved the highest C-index and median dynamic AUCs, with a C-index of 0.68 and median dynamic AUCs of 0.75 in the training set, and corresponding metrics of 0.69 and 0.75 in the testing set (Fig. 4A and Table 2). Bootstrapped Wilcoxon test (*p*-value < 0.05) indicates model 2 has statistically significant better performance than

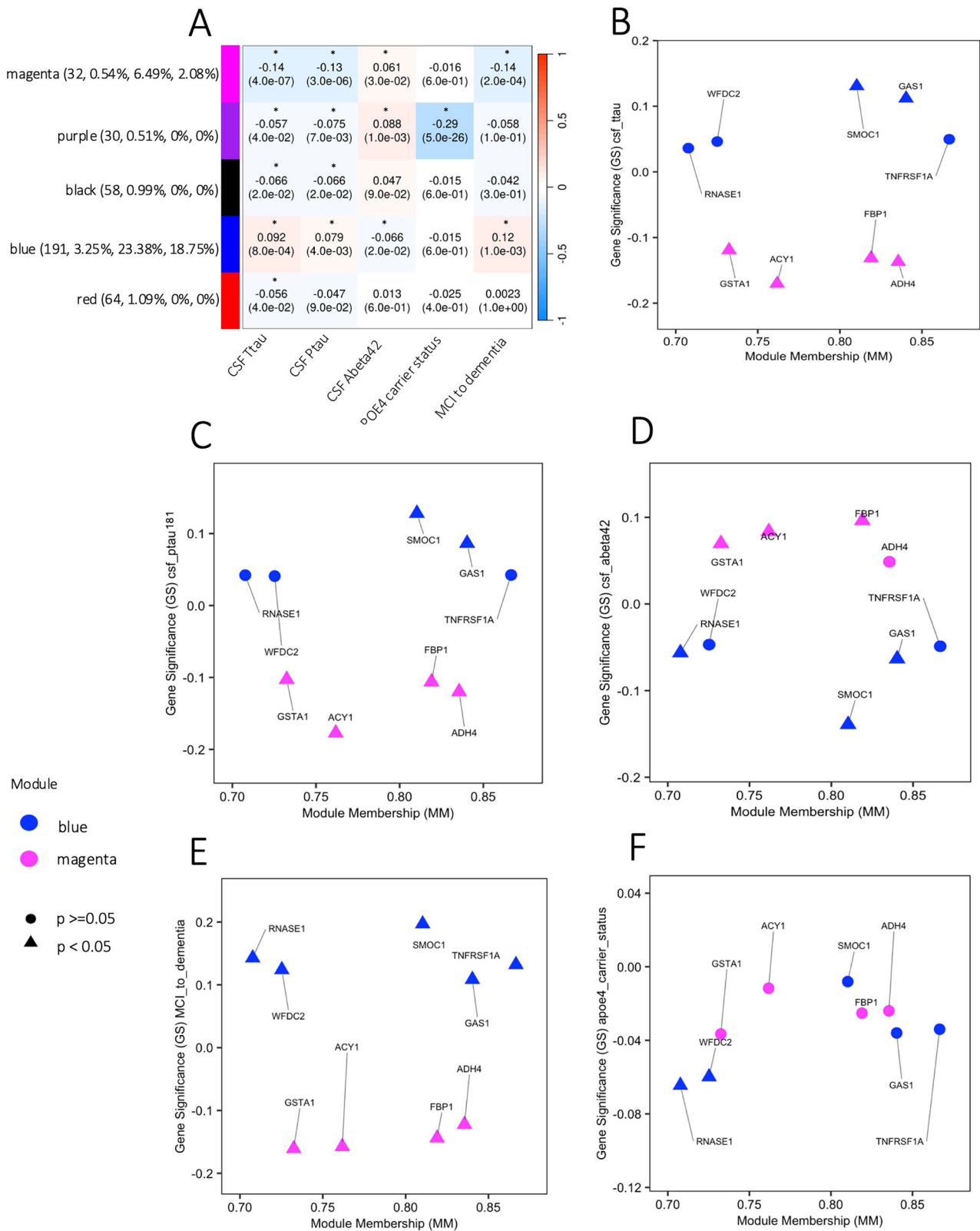


Fig. 3 | Protein co-expression modules associated with traits of interest.

A Heatmap showing the module-trait correlation and corresponding *p*-value of each module with AD CSF biomarkers (CSF Ttau, CSF pTau-181, CSF Aβ 42), MCI to dementia conversion status (0 = MCI at the last follow-up, 1=dementia), and APOE4 carrier status (0 = non-carrier, 1= one copy of APOEe4, 2 = APOE e4/e4). Only the five modules with a *p*-value < 0.05 are shown. The stars are used to flag levels of significance: *: 0.01 < *p*-value <= 0.05; **: 0.001 < *p*-value <= 0.01; ***: *p* <= 0.001. For each module, the numbers in parenthesis indicate number of proteins in

the module, percentage of all proteins in network, percentage of the 77 somamers in the network, and percentage of the 48 proteins in the network. **B–F** Gene significance and module membership of driver proteins (absolute module membership (MM) > 0.70, *p*-value < 0.05) from the set of 77 somamers (74 proteins) for the five modules (blue, black, magenta, red, purple). Notably, such proteins exist exclusively within the blue and magenta module. Proteins with significant (*p* < 0.05) gene significance (GS) coefficient in the co-expression network are indicated by a triangle shape.

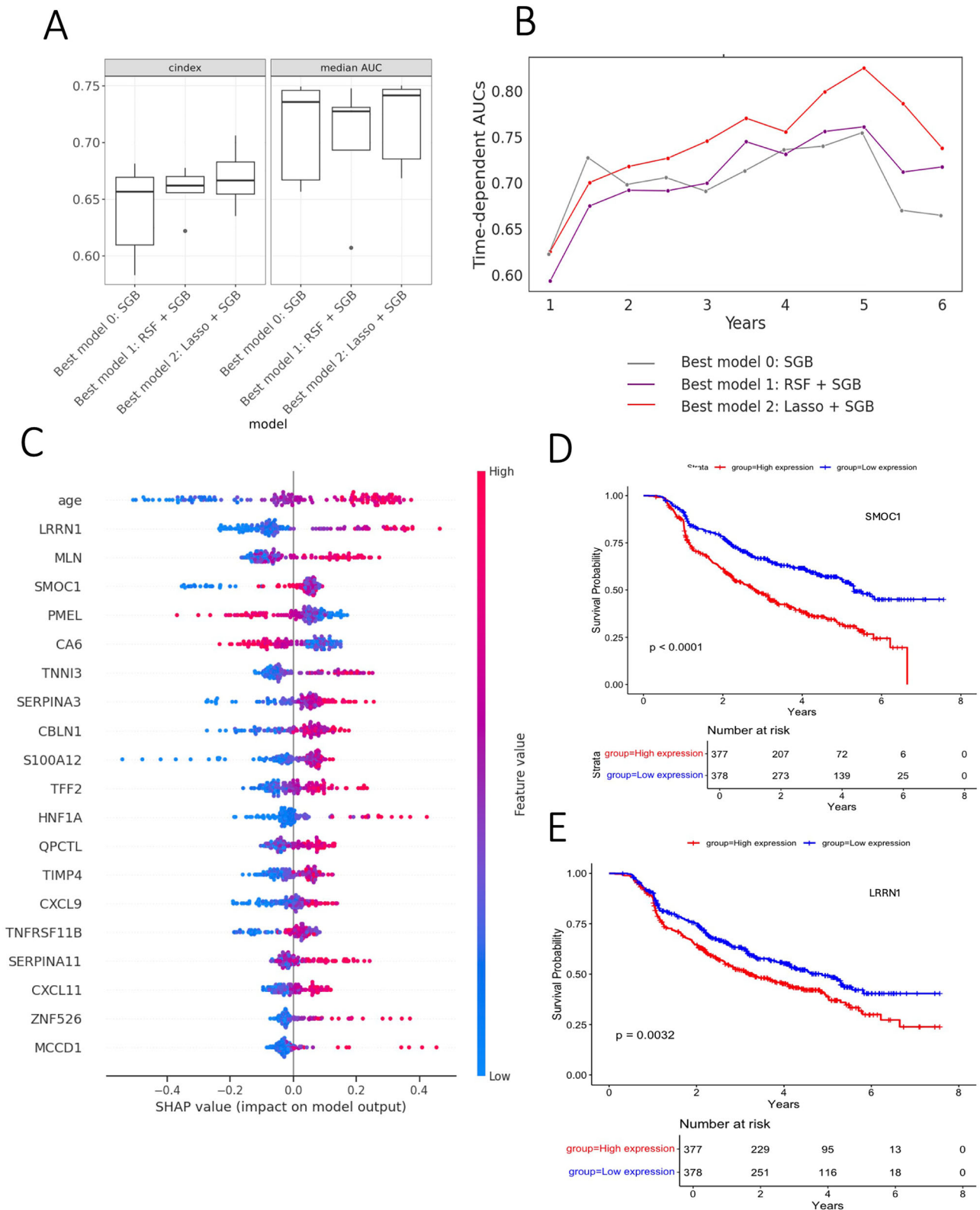


Fig. 4 | Predictive model for MCI-dementia conversion. **A** The concordance index and median AUCs (5-fold cross-validation) from three optimal models with three feature sets (demographics only, proteomics only, demographics and proteomics). The box indicates 1.5*IQR (interquartile range). The middle bar indicates the median value of c-index and AUC. **B** Dynamic AUCs from year 1 to year 6 applying the “best model” for three feature sets on F.ACE test set. **C** SHAP beeswarm plot showing importance of the top 20 features (19 protein features plus age) in the “best”

model in descending order. The plot illustrates each feature’s effect based on its magnitude and the direction of influence. The width of the range of horizontal bars is the contribution to the prediction of dementia progression. The magnitude of the features was coded in a gradient way: blue (low) to right (high) and the direction on axis. Kaplan-Meier survival curves with corresponding risk tables colored by plasma protein SMOC1 (**D**) and LRRN1 (**E**) expression in F.ACE cohort. Expression strata were determined by the median abundance value.

Table 2 | Evaluation metrics of three optimal models on F.ACE testing set

| Model | Evaluation metrics on testing set | |
|--|-----------------------------------|-------------|
| | C-index | Median AUCs |
| Demographics model (model 0): SGB | 0.68 | 0.71 |
| Proteomics model (model 1): RSF + SGB | 0.65 | 0.71 |
| Demographics + Proteomics model (model 2): Lasso + SGB | 0.69 | 0.75 |

the model 0 and model 1. Therefore, model 2 with 48 proteins, age and sex as feature sets was deemed the “best model”.

The time-dependent AUC plots on testing set (Fig. 5B) reveal that all three models exhibit strong performance between 1.5 and 5 years, with demographics and proteomics model (model 2) being the best. There is a prediction decline after 5 years. This reduction may be attributed to the lower number of events during early follow-ups and increased rate of censoring near the end of the follow-up period (Supplementary Fig. 6B).

The 48-protein panel collectively show enrichment in signaling-receptor activities, extracellular matrix (ECM) remodeling and peptidase regulation (Supplementary Fig. 5A), corroborating with the PPI networks, which highlighted three functional clusters—neutrophil chemotaxis, serine protease inhibitor and RETN-NPPB-TNNI3 (Supplementary Fig. 5B). These pathways and functional protein clusters converge on the inflammation regulation, underscoring the role of inflammation in AD progression.

To enhance the interpretability of the prediction model, we used Shapley Additive Explanations (SHAP)³⁵ beesworn plot to visualize the impact of each feature on the likelihood of MCI-to-dementia conversion. The SHAP beesworn plot for the top 20 features is shown in Fig. 4C, while the full panel SHAP beesworn plot is provided in Supplementary Fig. 4C. Among all features, age exhibited the widest range and ranked as the most important predictor, highlighting its significant predictive power. LRRN1 emerged as the most important protein feature among the model-selected 48-protein panel. Participants with elevated levels of LRRN1 showed a higher likelihood of converting from MCI to dementia, while those with lower levels of LRRN1 were more likely to remain stable in their MCI diagnosis. This relationship between LRRN1 and dementia progression shown in the beesworn plot aligns with the trends observed in the KM curve of LRRN1 (Fig. 4D). A similar interpretation applies to SMOC1, as reflected in Fig. 4C, E.

Next, we investigated the assignments of these proteins within the co-expression network. The module distribution of the 48 proteins closely mirrors that of the 77 significant somamers (74 proteins): 79.2% don't involve in any specific process (the grey module), 18.7% involve in the neuronal processes (the blue module), and the rest 2.1% involve in the metabolic processes (the magenta module, Fig. 3A). Notably, proteins such as SMOC1, TREM1, RNASE1, GAS1, DNAJB12, WFDC2, COL18A1, CRIP2, and IGF1R1, assigned to the blue module, were significantly associated with AD CSF biomarkers and MCI-to-dementia conversion (Figs. 2C and 3A), further emphasizing the blue module's predictive role in MCI-to-dementia conversion.

The “best model” shows comparable prediction performance in A+/T+ subgroups

As part of a sensitivity analysis, we evaluated the model's performance in the amyloid-positive (A+) and Tau-positive (T+) subgroups within the testing set of the F.ACE discovery cohort. As shown in Fig. 5A and Supplementary Table 13, the “best model” achieved a slightly higher C-index in the T+ subgroup compared to the A+ subgroup (0.68 vs. 0.66). The dynamic AUCs were consistent across both groups up to 4.5 years of follow-up. However, beyond 4.5 years, the censoring rate in the A+ group dropped to 0.1, followed by a sharp decline at 5.5 years (Supplementary Fig. 4B). This low

censoring rate introduces potential bias and increases variability in the time-dependent AUCs. Notably, the C-index and dynamic AUCs in the A+/T+ subgroups were comparable to those of the full testing set, indicating that the model generalizes well from general dementia to AD subgroups.

The “best model” can discriminate patients with higher and lower risk of conversion

To simulate a real-world clinical application for stratifying patient risk, we calculated MCI-to-dementia conversion risk using the best predictive model. Higher risk is defined as a probability greater than 0.5 for MCI-to-dementia conversion at a specified timepoint, while a probability less than or equal to 0.5 indicates lower risk. We applied this model to the F.ACE testing set, focusing on assessing the risk of dementia conversion at 1-year intervals, from 1 year to 5 years post-baseline visit. The KM curve illustrates the observed probabilities, color-coded by inferred risk strata (Supplementary Fig. 6). Overall, the “best model” performed well in discriminating between high- and low-risk patients from years 2 to 5, as evidenced by a log-rank test *p*-value of less than 0.05. For example, 98 MCI patients were predicted to have a higher probability of having MCI-to-dementia conversion within 4 years of their initial visit, while 53 patients were predicted to have a lower probability. The log-rank test revealed a statistically significant difference between the survival curves of the two inferred risk groups (Fig. 5B, *p* = 2e-06), suggesting that our model can effectively differentiate individuals at higher risk of near-term MCI-to-dementia conversion from those at lower risk.

Validation of the “best model” in two independent datasets

We validated our findings using the EMIF AD-MBD cohort. After performing quality control on samples and flagging low-quality proteins (See Materials and Methods, Supplementary Fig. 7), the validation dataset consists of 3006 human protein somamers from two independent datasets: 60 participants from the Dataset 1 and 110 individuals from the Dataset 2 (Table 1). The two EMIF-AD datasets vary in follow-up duration, conversion rates, age distribution (Table 1), amyloid status (Supplementary Table 1B), and exhibit batch effects in proteomics data (Supplementary Fig. 1E). Combining them would introduce significant heterogeneity and potential bias, compromising the validity of the analysis. Therefore, we opted to validate the two datasets separately.

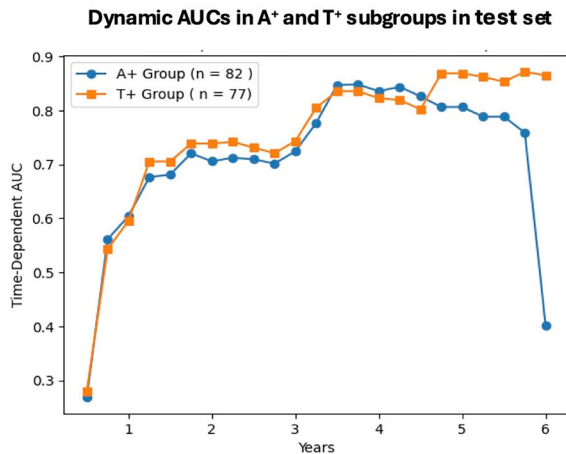
We applied our “best model” to the two datasets of the EMIF-AD MBD cohorts separately and evaluated their C-index and time-dependent AUCs (Table 3).

Overall, Dataset 2 demonstrated superior performance compared to Dataset 1, achieving a C-index of 0.61 and median AUCs of 0.60. The dynamic AUCs for Dataset 2 were relatively stable around 0.60 (Supplementary Fig. 8A). We used the same criteria for risk stratification as in the F.ACE testing set. This predictive model was applied to both datasets to infer the risk strata for dementia progression from 1 to 4 years post-baseline visit, at 1-year intervals (Supplementary Fig. 8C–J). For example, the KM curve of the observed survival probabilities, colored by inferred risk strata at year 2 post-baseline visit, shows that the model significantly distinguishes between inferred high- and low-risk groups for Dataset 2 (log-rank *p*-value = 0.01; Fig. 5D). In contrast, no significant difference in inferred risk strata was observed in Dataset 1 (log-rank *p*-value = 0.73; Fig. 5C).

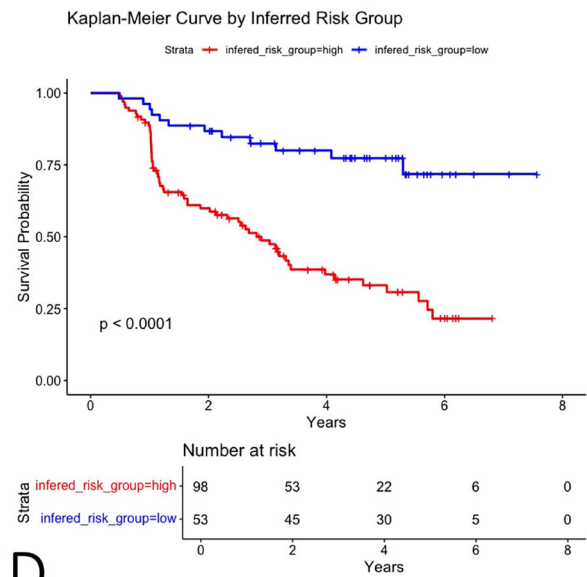
The 48-protein panel achieves comparable performance to CSF pTau-181

Plasma pTau-217 has been validated as a good predictor for AD dementia progression in MCI patients^{36,37}. While many of these studies did not include temporal factors in their models, directly comparing the performance of the 48-protein panel to plasma pTau-217 could provide valuable insights. However, CSF pTau-181 is the only measured phosphorylated Tau in F.ACE cohort. Given that plasma pTau-217 and pTau-181 exhibit high accuracy in predicting amyloid-PET with minimal difference (despite pTau-217 has demonstrated slightly better performance)^{38,39}, and that plasma pTau-181 show moderate correlation with CSF pTau-181 ($r \approx 0.66^{39}$), we

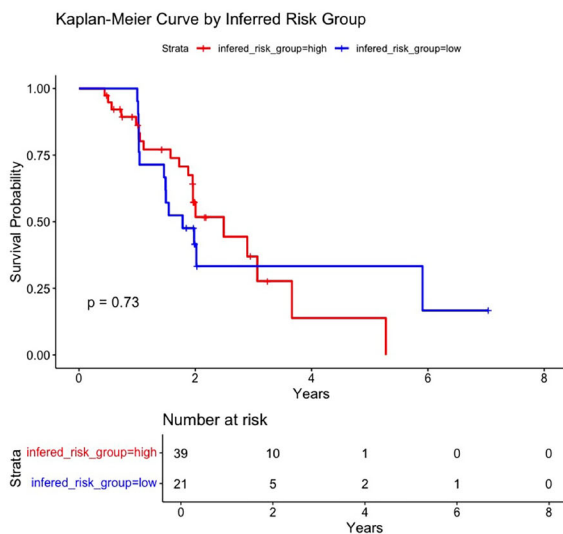
A



B



C



D

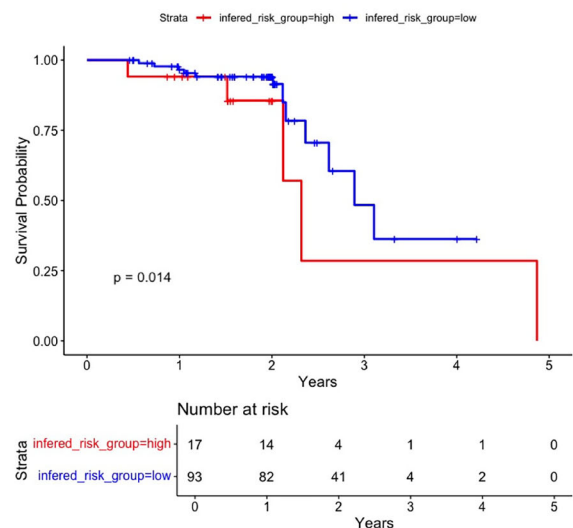


Fig. 5 | Performance of “best model” in A+ / T+ subgroup and in EMIF-AD MBD cohort. A Dynamic AUCs of the “best” model across follow-up years within amyloid-positive (A+) and Tau-positive (T+) subgroups in the F.A.C.E. test set. Kaplan-Meier survival curves with corresponding risk tables illustrating risk strata at

year 4 post-baseline in F.A.C.E. test set (B), year 2 post-baseline in EMIF-AD dataset 1 (C) and dataset 2 (D) based on predicted survival probabilities for mild cognitive impairment (MCI) stability, generated by the “best model”. Blue indicates the low-risk group, and red indicates the high-risk group.

Table 3 | Evaluation metrics of “best model” on EMIF-AD MBD dataset

| EMIF-AD MBD | Evaluation metrics on the validation set | |
|-------------|--|-------------|
| | C-index | Median AUCs |
| Dataset 1 | 0.51 | 0.56 |
| Dataset 2 | 0.61 | 0.60 |

have used the CSF pTau-181 as a surrogate for comparison. In the F.A.C.E. cohort, 439 of 755 participants had CSF pTau-181 measurements. For a fair comparison, we retrained the best model (48 proteins + age + sex) using this subset and applied 5-fold cross-validation to identify the optimal configuration and compared to the predictive model using CSF pTau-181,

age, and sex. The 48-protein panel achieved a C-index of 0.73 and median AUC of 0.78 in the testing set, which are very comparable to the performance of CSF pTau-181 (C-index = 0.73; median AUC = 0.79) (Supplementary Fig. 4D). The comparable performance suggests that the plasma-based 48-protein panel could offer a non-invasive alternative to CSF pTau-181 for predicting AD dementia progression.

Discussion

Understanding what drives the high heterogeneities in the clinical progression of AD, aside from Tau and amyloid, would further our understanding of dementia progression and enable more efficient trial design. Identifying plasma proteomic signatures associated with MCI-to-dementia conversion in AD context provides critical insights into underlying biological mechanisms and can facilitate biomarker and drug target discovery.

Leveraging plasma proteomics and longitudinal clinical data from the F.A.C.E cohort, we identified 74 proteins (corresponding to 77 somamers) associated with near-future MCI-to-dementia conversion. Of these, 13 proteins were validated in at least one independent replication dataset from the EMIF AD-MBD cohort. Seventeen proteins (22%) had been previously reported as associated with dementia risk in middle-aged adults, such as GDF15, SERPINA3, MMP12, CA6, IGFBP2, etc.^{22,23}. The relatively low overlap (22%) underscores both shared and distinct protein signatures between near-term (3–5 years) dementia conversion in MCI patients and dementia development later in life among otherwise healthy middle-aged adults. These findings highlight the necessity and importance of studying near-term MCI-to-dementia progression to uncover unique mechanisms and actionable biomarkers.

Not all significant proteins were associated with AD CSF biomarkers of A β and Tau, suggesting that plasma proteomics implicate mechanisms beyond those of the two primary AD pathologies. Specifically, proteins associated with MCI-to-dementia conversion were significantly enriched in extracellular matrix (ECM)-related processes, suggesting a critical role of ECM in disease progression. Rahman et al.⁴⁰ demonstrated that ECM-related proteins are not only present in amyloid plaques but also actively contribute to the initiation and accumulation of A β aggregates. Additionally, Sun et al.⁴¹ reported that the ECM can impair synaptic transmission and exacerbate the pathological processes underlying AD. Indeed, proteins driving the enrichment of ECM includes those not associated with A β or Tau, for example, the tissue remodeling protein MMP12, highlights the complex inflammatory and neurodegenerative processes involved. MMP12 belongs to the matrix metalloproteinase (MMP) family. MMP-2 and MMP-9 from this family were found to influence amyloid-beta peptide clearance⁴². MMP-12 may be involved in AD associated neuroinflammation and neurodegeneration⁴³. Studies in mice suggested that MMP12 was strongly upregulated during aging, and MMP12 deficiency reduced neuroinflammation during aging⁴⁴ suggesting the neuroinflammatory role of MMP12. Interestingly, TIMP4, a member from the TIMP (tissue inhibitors of metalloproteinases) family that helps to counterbalance MMP activity by inhibiting excessive matrix degradation, significantly associated with A β and Tau. TIMP4 has been found elevated in AD, even at the early stage^{45–47}, vascular dementia⁴⁸ and cerebral amyloid angiopathy⁴⁹, and was implicated in inflammatory processes⁵⁰. Our results suggest disturbed balance of these factors contributing to inflammation and neurodegeneration during dementia conversion.

In the co-expression modules analysis, five modules were associated with at least one AD CSF biomarker. However, only the blue (neuronal) and magenta (metabolic) modules were linked to MCI-to-dementia conversion risk, and with opposing directions of correlation. The pathways associated with the blue module are integral to neuronal structure remodeling and synaptic organization—processes known to be disrupted in dementia—suggesting a pathogenic role in the progression of AD dementia. In contrast, the metabolic magenta module may function as a compensatory mechanism against dementia pathogenesis. The blue module is the sole module whose driver proteins were selected in the best predictive model for future dementia conversion. These driver proteins include SMOC1, which ranked highly (3rd) based on SHAP values, along with GAS1 (24th) and RNASE1 (26th), underscoring their importance in predicting dementia progression.

SMOC1 (SPARC-related modular calcium-binding protein 1) has emerged as a key protein signature of Alzheimer's disease across multiple independent studies utilizing various tissue types and orthogonal proteomics technologies. Numerous proteomics studies have reported elevated SMOC1 in brain tissue and CSF in both sporadic AD and autosomal dominant AD (ADAD)^{51–56}. Notably, this elevation occurred decades before symptom onset in ADAD⁵⁷. SMOC1 has also been shown to colocalize and interact with amyloid and Tau pathologies in the brain⁵². Furthermore, SMOC1 was included in a CSF 48-protein panel in a study that demonstrated significantly improved prediction performance for the change of cognitive decline and dementia severities when combined with A β and Tau⁵⁸. In our study, SMOC1 was selected as a top-ranked feature in the “best

model”, which comprises 48 plasma protein signatures for predicting MCI-to-dementia conversion. Higher plasma abundance of SMOC1 was associated with accelerated dementia progression, further supporting its critical role as a potential biomarker for dementia progression.

LRRN1 (Leucine-rich repeat neuronal 1) has been identified as the most important protein feature in the “best” model. While the specific function and mechanism of LRRN1 remain to be elucidated, as a protein with leucine-rich repeat domains, LRRN1 may play a crucial role as an organizer of synaptic specificity and diversity⁵⁹. We found that LRRN1 was significantly associated with APOE4 genotype and CSF A β and Tau (Fig. 2C, Supplementary Table 5), corroborating the observation that LRRN1's association with late-onset AD is APOE4-dependent⁶⁰. These results suggest that LRRN1 may be a key player in A β and Tau-mediated pathological processes and underscore the importance of A β and Tau in dementia conversion.

Protein signatures associated with disease progression may serve as promising drug targets. For instance, CA6 and CA3 are members of carbonic anhydrase family, which have been suggested as potential targets for AD^{25,26,61}. Notably, seven proteins, including SMOC1, were proposed as potential therapeutic targets by AMP-AD consortium members, with SMOC1 being nominated twice. Furthermore, five proteins (LRRN1, SERPINA11, TIMP4, SMOC1, TREM1) were identified as potentially causal for AD disease onset and were also associated with disease progression, highlighting their potential as therapeutic targets for AD intervention.

By integrating a protein-based feature set with demographic factors, we developed a predictive model capable of distinguishing patients at higher and lower risk for MCI-to-dementia conversion. KM analysis revealed distinct survival probability profiles between the risk strata, particularly during the clinically meaningful time window of year 2–5, underscoring the model's potential utility. Furthermore, the model demonstrated comparable performance in the A+/T+ patient subgroups, supporting its potential for more refined clinical applications. The 48-panel model also showed comparable performance to that of CSF pTau-181. Future exploration of these proteins, alongside established plasma biomarkers like pTau-217, has the potential to enhance predictive accuracy.

When validating the model in independent datasets, we observed a decline in performance, particularly in EMIF AD Dataset 1. The compromised performance can be attributed to several factors. First, the age distribution in the EMIF AD Dataset 1 differs from that of the F.A.C.E cohort, which may impact the model's applicability and performance, since age is the strongest predictor in the model. Another important factor is pathology burden. EMIF-AD Dataset 1 had much higher AD pathology burden, ~95% individuals were amyloid-positive with higher pTau and tTau in CSF at baseline. The higher pathology burden and similar conversion rate in shorter follow-up time indicate participants in EMIF-AD dataset1 were likely in a more advanced preclinical stage than the participants in F.A.C.E cohort. These differences highlight the impact of cohort composition on model performance and generalization. Definition and diagnosis of dementia conversion across different study sites can introduce phenotype variabilities, further compromising the generalizability of the findings. Additionally, the affinity-based proteomics measures relative abundance, and is sensitive to dataset-specific processing, which can affect reproducibility and complicate the interpretation of results. Several proteins ranked high by SHAP (e.g., SERPINA1, TNNI3, CA3, CANT1) failed proteomics QC in EMIF-AD Dataset 1, likely reducing accuracy.

Several limitations of this study should be acknowledged. First, the sample size in the F.A.C.E cohort is relatively small, which poses challenges for high-dimensional proteomics data. To mitigate this, we employed a bootstrap-based feature selection approach to identify reproducible and stable predictors; however, these protein signatures require validation in larger and diverse cohorts. The two EMIF-AD MBD datasets used for validation are also small and differ from the discovery cohort in key characteristics, including demographic factors, APOE4 carrier frequency, conversion rates, and amyloid status, etc. Such heterogeneity across cohorts complicates validation and limits generalizability. Beyond biological and demographic

factors, systematic biases introduced by sample processing protocols and reagent sensitivity and specificity of proteomic platforms and assays remain a concern, making cross-study comparisons challenging. Addressing these disparities will require the use of high-quality, technology-agnostic reference samples to enable robust cross-study comparisons⁶². By sharing these findings, we aim to contribute to this ongoing discussion and emphasize the need for collaborative efforts within the research community. Furthermore, affinity-based proteomics approaches measure relative protein abundance and are susceptible to epitope effects, meaning that identified signatures will require validation using orthogonal techniques such as mass spectrometry or ELISA. In addition, the plasma proteome tends to overrepresent secreted proteins, which may introduce bias in our findings. Finally, our study cohorts are predominantly European, underscoring the need for future studies that include participants from diverse racial and ethnic backgrounds to ensure broader applicability.

In summary, our study identified 74 plasma proteins associated with MCI-to-dementia conversion and developed a machine learning model using 48 proteins, including SMOC1 and LRRN1, for near-term dementia conversion and risk stratification in patients with MCI. The 48-protein panel demonstrated predictive performance comparable to CSF pTau-181. Proteomics signatures that capture the multifactorial nature of dementia and AD may facilitate the development of precision strategies in dementia clinical research and care. With advances in proteomics techniques and the inclusion of more disease-focused cohorts and samples, our approach could lead to the discovery and validation of new plasma biomarkers, more refined predictive models, and ultimately accelerate therapeutics discovery and development as well as clinical care for patients with MCI and dementia.

Methods

F.ACE cohort

The F.ACE (Fundació ACE) is a nonprofit Alzheimer's center based in Barcelona, Spain dedicated to advancing our understanding of Alzheimer's disease and other dementias. The F.ACE cohort included participants with subjective cognitive impairment, a subset of whom were diagnosed with mild cognitive impairment or dementia. Patients with MCI were followed up to monitor dementia conversion. Cross-sectional data at baseline are available on 1370 subjects, including genotyping array, APOE genotype, lipidomics, ATN classification⁶³, and plasma and CSF proteomics data, along with extensive longitudinal clinical data for 755 patients diagnosed with MCI at baseline⁶⁴. The follow-up data used in this study were collected up to April 2024.

The diagnosis of MCI at baseline was based on impaired cognitive domains, assessed by the neuropsychological battery of ACE (NBACE), with impairment cut-offs specific to age, formal education levels, and a CDR of 0.5^{65–68}. Alzheimer's disease diagnosis followed the 2011 National Institute on Aging and Alzheimer's Association (NIA-AA) guidelines, and dementia was defined according to the DSM-V criteria⁶⁹. A+ and T+ statuses were defined using Gaussian mixture modeling of Aβ 42 and pTau-181 in CSF to establish cutoff values⁷⁰. Dementia conversion was defined as the first clinical follow-up visit with a CDR greater than or equal to 1, a GDS (Geriatric Depression Scale) score greater than or equal to 4, and a primary diagnosis of dementia. The time to MCI-to-dementia conversion status was calculated as the duration between the last follow-up visit (if no conversion occurred) or the follow-up visit indicating conversion and the date of the first clinic visit (baseline). Further details on the clinical characteristics of this cohort can be found elsewhere^{63,64,71,72}.

EMIF-AD MBD cohort

The European Medical Information Framework for Alzheimer's disease Multimodal Biomarker Discovery (EMIF-AD MBD) is a pan-European large-scale research project that aims to accelerate the discovery of diagnostic and prognostic biomarkers for AD as well as to unravel the pathological mechanism using extensive phenotypic and in-depth proteomics data. It is conducted as part of the EMIF-AD project, and the cohort selection is built upon EMIF-AD Catalog (<https://emif-catalogue.eu>) with

certain inclusion criteria, including subjects with normal cognition (NC), MCI and AD-type dementia⁷³.

The diagnosis of MCI in EMIF-AD MBD was made according to the criteria of Petersen⁷⁴. AD type dementia was diagnosed using the National Institute of Neurological and Communicative Disorders and Stroke–Alzheimer's Disease and Related Disorders Association criteria (NINCDS-ADRDA) criteria⁷⁵. Similarly to F.ACE cohort, the time to MCI-to-dementia conversion status was calculated as the duration between the last follow-up visit (if no conversion occurred) or the follow-up visit indicating conversion and the date of the first visit.

The clinical metadata were obtained up until June 2020, and the two proteomics datasets for distinct subjects from different clinic sites were obtained in September 2021. In our study, we included only subjects diagnosed with MCI at baseline who had plasma proteomics data, resulting in an analytic cohort of 170 individuals: 60 from Dataset 1 and 110 from Dataset 2 (Table 1).

Plasma proteomics and data quality control

Plasma proteomics were profiled using SomaScan[®] 7K and SomaScan[®] 4k (SomaLogic Operating Co., Inc., Boulder, Colorado) for the F.ACE and EMIF-AD MBD plasma samples, respectively. The SomaScan platform employs a SOMAmer-based capture array method to quantify the relative abundance of plasma proteins or protein complexes, with fairly good reproducibility^{76,77}. SomaScan platform utilizes SOMAmer (Slow off-rate modified aptamer) reagents that can bind to a specific epitope of a target protein. Multiple somamers annotated to the same gene may bind to different regions of a same protein, different isoforms, different proteoforms from post-translationally modifications and different 3D conformations. To aid interpretation, we report both the SOMAmer name (aptamer ID) and the corresponding targeted gene.

All samples passed SomaLogic's quality control. Additionally, we applied a multi-step quality control (QC) process at both the sample and the somamer levels, following the methodology described by Yang et al.⁷⁸. Briefly, principal component analysis (PCA) was used to cluster the samples, and samples with a standardized PC1 or PC2 value more than 5 standard deviations from the mean were removed. Somamers with a coefficient of variation (CV) > 0.15, a scale factor (SF) > 0.5 and limit of detection (LOD) > 0.85 were also removed. Only somamers measuring human proteins were included in subsequent analyses and log10 transformation was applied. Additionally, we used the VariancePartition package⁷⁹ in R 4.3.3 to explore proteomics expression variance attributed to the biological (age, sex, and BMI) and technical variables, such as plate id, scanner id, plasma sample storage duration in months. No obvious batch effects were observed.

Supplementary Fig. 7 lists the number of samples and somamers included at each QC step. For the F.ACE cohort, 5882 somamers of 1563 samples passed the final quality control. Low-quality proteins were observed in the EMIF-AD MBD Dataset 2 and were flagged. No low-quality proteins were observed in Dataset 1. These low-quality proteins were not excluded, as they may represent relevant signatures identified in the discovery cohort. Due to the difference in the number of proteins in SomaScan 7k and SomaScan 4k panels (Supplementary Fig. 1F), we only used the shared 3006 proteins that passed the quality control in F.ACE cohort for model development and replication.

Protein co-expression network analysis

For protein co-expression modules identification, we included all subjects with available plasma proteomics data, including those with subjective cognitive impairment or diagnosed with AD. The co-expression network was constructed using the WGCNA package (version 1.7.2)³¹. We evaluated 16 parameter combinations, including correlation methods, R-squared, module size and module similarity thresholds, etc., using the F.ACE cohort data comprising 5882 somamers and 1325 samples. The following traits were examined for module-trait correlations: AD CSF biomarkers (pTau-181, tTau, Aβ 42), APOE4 carrier status (0 = non-carrier, 1 = one copy of APOE4, 2 = APOE4 homozygote), MCI-to-dementia conversion

(conversion=1, no conversion at the last follow-up=0), time to conversion in years, age in years and sex (male = 1, female = 0). After evaluating module size and module-trait associations, we selected the final parameter combination: a signed network with minimum module size of 25, no member reassignment after initial module construction, merge height of 0.1, deep split of 4, and soft threshold of 15.

Over-representation analysis (ORA) of Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), WikiPathway and Reactome were performed for the five modules showing protein-trait significance (p -value < 0.05). The background gene set included the 5276 genes corresponding to the 5882 somamers. The ORA analysis was performed using WebGestalt 2019⁸⁰.

Driver proteins for the significant trait-associated modules were identified based on module membership (MM), which measures the overall correlation of a given protein with the other members in the same module. Proteins with MM > 0.70 and an MM p -value < 0.05 were designated as driver proteins. Additionally, gene-significance (GS) coefficients, which indicate how strongly a gene's expression correlates with a specific trait, were used to identify driver proteins significantly associated with clinical traits of interest (GS p -value < 0.05).

Statistical analysis to identify signatures associated with MCI-to-dementia conversion

For the identification of plasma protein signatures associated with future MCI-to-dementia conversion, we focused on patients diagnosed with MCI at baseline who had longitudinal follow-ups and plasma proteomics data available ($n = 755$).

After QC and transformation of proteomics data, we divided the discovery dataset into training (80%) and testing (20%) sets. To ensure robust downstream model development and prevent data leakage, we identified protein signatures associated with MCI-to-dementia conversion using only the training set. The event of interest was defined as conversion from MCI to dementia within the follow-up period, while censored observations included participants who either remained stable in their MCI diagnosis or reverted to normal cognitive function. Bootstrapped Univariate Cox Proportional Hazard (Cox-PH) regression models were employed to assess the marginal effects of individual plasma proteins on MCI-to-dementia conversion. Bootstrapping ($n = 1000$ iterations) was performed to identify protein signatures that could optimally distinguish MCI patients who progressed to dementia from those who remained stable during follow-up. Each iteration involved random sampling with replacement from the case (MCI-to-dementia) and control (MCI at last follow-up, right-censored) groups. For each protein, a Cox-PH model was fitted, and false discovery rate (FDR) corrections were applied to identify significant associations (FDR < 0.05). Proteins with an FDR < 0.05 in over 500 iterations were deemed significantly associated with MCI-to-dementia conversion. Hazard ratios (HR), 95% confidence intervals (CI), p -values and FDR were reported. Of the significant somamers identified in the discovery training set, their associations were further validated using two independent datasets from the EMIF-AD MBD cohorts via univariate Cox PH model on the full dataset. Replication was defined as a p -value < 0.05 in the validation dataset. The Cox-PH regression model was performed using the R survival package (version 3.7-0)⁸¹.

Cross-sectional correlation of the significant somamers with AD CSF biomarkers (A β 42, pTau-181 and total Tau) and APOE4 carrier status were evaluated using spearman correlation in R. ORA analysis of GO, KEGG and Reactome were conducted using ClusterProfile R package (version 4.14.6)⁸². In addition, we explored the protein-protein interaction network using STRING database with k-mean clustering ($k = 6$)⁸³.

Predictive model development

To evaluate the applicability of identified proteins in predicting the near-future MCI-to-dementia conversion, we developed and evaluated the performance of three models via a 5-fold cross-validation (CV) approach:

1. Demographics model (model 0): base model including only age and sex.
2. Proteomics model (model 1): model including plasma proteins only.
3. Demographics + Proteomics model (model 2): a full model incorporating demographics (age, sex) and proteomics.

To identify the best predictive model, we explored three machine learning (ML) algorithms—Random Survival Forest (RSF), survival Gradient Boost (SGB), and survival Support Vector Machine (sSVM)—coupled with three feature selection methods: Cox PH Lasso, RSF minimal depth, and Recursive Feature Elimination (RFE). Performance was evaluated using 5-fold CV with two survival-specific metrics: Harrell's Concordance Index (C-index) and time-dependent area under the ROC curve (AUC). Model training and evaluation were performed using 5-fold cross validation on the 80% training set. To test if there are statistically significant differences among the three optimal models, we performed 1000X Bootstrapped resampling on the testing set and applied the Wilcoxon signed-rank test on the evaluation metrics.

Given different proteomics panels were used for the F.ACE and EMIF-AD MBD cohorts, to ensure consistency and reliability, we used the significant somamers that are shared between the two panels as starting features in proteomics model (model 1) and full model (model 2). In the full model (model 2), age and sex were also tested as additional features.

Model development and optimization were conducted using Python 3.9 and R 4.3.4. The R environment was interfaced with Python via the rpy2 package⁸⁴. All covariates were scaled and standardized by removing the mean and scaling to unit variance using the StandardScaler function from the scikit-learn library in Python⁸⁵. Cox PH Lasso was performed using the glmnet package in R^{86–88}, with family set to "cox" and the type.measure set to "C". For feature selection, RSF minimal depth was implemented using the randomForestSRC package (version 3.3.3) with default parameters in R⁸⁹. Custom Python codes were developed for RFE utilizing permutation importance. The R environment was interfaced with Python via the rpy2 package⁸⁴. All covariates were scaled and standardized by removing the mean and scaling to unit variance using the StandardScaler function from the scikit-learn library in Python⁸⁵. Cox PH Lasso was performed using the glmnet package in R^{86–88}, with family set to "cox" and the type.measure set to "C". For feature selection, RSF minimal depth was implemented using the randomForestSRC package (version 3.3.3) with default parameters in R⁸⁹. Custom Python codes were developed for RFE utilizing permutation importance. All three machine learning algorithms were implemented in Python using the scikit-survival package⁹⁰. Stratified 5-fold CV was performed using the StratifiedKFold function in Python. Performance metrics were obtained using concordance_index_censored and cumulative_dynamic_auc at time points ranging from 1 to 6 years (intervals of 0.25 years) using the sksurv.metrics package. C-index and time-dependent AUC values were reported for both training and testing sets. The best model was selected based on higher C-index and higher median time-dependent AUCs across CV results. Best model with parameters was saved for validation. As a sensitivity analysis, model was also tested in A+ or T+ subgroups.

To interpret the effect of the selected feature sets in the best model, shapley additive explanations (SHAP) were visualized using a beeswarm plot generated with the shap package in Python³⁵. KM curves for selected features (eg. SMOC1 and LRRN1) were generated in a color-coded fashion by protein expression strata (determined by the median abundance value) using the R survival package (version 3.7-0)⁸¹ and survminer (version 0.5.0)⁹¹.

To investigate the model's ability to identify patients at high risk of dementia conversion, we applied the best model to the testing set to predict survival probabilities at a given time point using the predict_survival_function function in Python. This generated the estimated survival probabilities across time for each testing-set individual. Probabilities at specified time points (e.g., 2 years post-baseline) were used to categorize patients into high- or low-risk groups using a 0.5 threshold. KM plots were generated to compare inferred risk groups at each specified time

point, and log-rank p -values were calculated using Python lifelines library. Model risk discrimination was evaluated at annual intervals from 1 to 5 years post-baseline.

Predictive model validation in independent datasets

To assess the generalizability of the model predictions, we applied the best model to two datasets of EMIF-AD MBD cohorts. Proteomics and demographic features were transformed (normalized or standardized) using parameters derived from the F.ACE cohort, assuming the datasets represent the same population. The best model was reloaded and predicted survival functions were generated using `predict_survival_function` in Python. C-index and time-dependent AUCs were generated the same way as described above, at a sequence of time points ranging from 1 to 4 years with an increment of 1 year for both datasets. Risk strata were inferred similarly to the F.ACE cohort. Patients with predicted survival probability greater than 0.5 at a given time point were categorized as high risk for MCI-to-dementia conversion, while those with probabilities ≤ 0.5 were categorized as low risk. KM plot colored by inferred risk strata were generated and log-rank p -values were calculated using python lifelines package⁹².

All analyses were performed using R 4.4.4 and python 3.9. Visualization were primarily created using seaborn python package⁹³ and ggplot2 R package⁹⁴.

Benchmarking the 48-protein panel with CSF pTau-181

To benchmark our 48-protein based predictive model against a CSF pTau-181 biomarker-based, we focused on participants in the F.ACE cohort with available CSF pTau-181 measurements. Of the 755 individuals, 439 had CSF pTau-181 data. For a valid comparison, we restricted the analysis to this subset. We retrained our plasma-based model (48 proteins + age + sex) and constructed a reference model using CSF pTau-181, age, and sex. Both models were evaluated under identical conditions. The subset dataset was randomly split into 80% training and 20% testing. Within the training set, we applied 5-fold cross-validation for hyperparameter tuning and model selection. Hyperparameter tuning was performed within each fold using grid search to optimize the primary performance metric (C-index). After selecting the optimal configuration, both models were evaluated on the held-out testing set. External validation on EMIF-AD MBD dataset were not conducted because CSF pTau 181 is not measured in this cohort.

Data availability

Results were shared in the main table or supplementary table. Individual-level data are not available.

Received: 11 August 2025; Accepted: 8 December 2025;

Published online: 09 January 2026

References

1. WHO. Global Status Report on the Public Health Response to Dementia (WHO, 2021).
2. Alzheimer's Association Report. 2020 Alzheimer's disease facts and figures. *Alzheimer's Dement.* **16**, 391–460 (2020).
3. Duyckaerts, C., Delatour, B. & Potier, M. C. Classification and basic pathology of Alzheimer disease. *Acta Neuropathol.* **118**, 5–36 (2009).
4. Lambert, J. C. et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
5. Lambert, J. C. et al. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat. Genet.* **41**, 1094–1099 (2009).
6. Wightman, D. P. et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat. Genet.* **53**, 1276–1282 (2021).
7. Western, D. et al. Proteogenomic analysis of human cerebrospinal fluid identifies neurologically relevant regulation and implicates causal proteins for Alzheimer's disease. *Nat. Genet.* **56**, 2672–2684 (2024).
8. Tijms, B. M. et al. Cerebrospinal fluid proteomics in patients with Alzheimer's disease reveals five molecular subtypes with distinct genetic risk profiles. *Nat. Aging* **4**, 33–47 (2024).
9. Tijms, B. M. et al. Pathophysiological subtypes of Alzheimer's disease based on cerebrospinal fluid proteomics. *Brain* **143**, 3776–3792 (2020).
10. Jack, C. R. Jr. et al. Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's Association Workgroup. *Alzheimers Dement* **20**, 5143–5169 (2024).
11. Albert, M. S. et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* **7**, 270–279 (2011).
12. McKhann, G. M. et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* **7**, 263–269 (2011).
13. Sperling, R. A. et al. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* **7**, 280–292 (2011).
14. Lovell, M. A. A potential role for alterations of zinc and zinc transport proteins in the progression of Alzheimer's disease. *J. Alzheimers Dis.* **16**, 471–483 (2009).
15. Tabuas-Pereira, M. et al. Prognosis of early-onset vs. late-onset mild cognitive impairment: comparison of conversion rates and its predictors. *Geriatrics* **1**, 11 (2016).
16. Giau, V. V., Bagyinszky, E. & An, S. S. A. Potential fluid biomarkers for the diagnosis of mild cognitive impairment. *Int. J. Mol. Sci.* **20**, 4149 (2019).
17. Chan, K. C. A. L. et al. Analysis of the human serum proteome. *Clin. Proteom.* **1**, 101–225 (2004).
18. Ashton, N. J. et al. Diagnostic accuracy of a plasma phosphorylated tau 217 immunoassay for Alzheimer disease pathology. *JAMA Neurol.* **81**, 255–263 (2024).
19. Lai, R., Li, B. & Bishnoi, R. P-tau217 as a reliable blood-based marker of Alzheimer's disease. *Biomedicines* **12**, 1836 (2024).
20. Ashton, N. J. et al. Diagnostic accuracy of the plasma ALZpath pTau217 immunoassay to identify Alzheimer's disease pathology. medRxiv (2023).
21. Hansson, O. et al. The Alzheimer's Association appropriate use recommendations for blood biomarkers in Alzheimer's disease. *Alzheimers Dement* **18**, 2669–2686 (2022).
22. Guo, Y. et al. Plasma proteomic profiles predict future dementia in healthy adults. *Nat. Aging* **4**, 247–260 (2024).
23. Walker, K. A. et al. Proteomics analysis of plasma from middle-aged adults identifies protein markers of dementia risk in later life. *Sci. Transl. Med.* **15**, eadf5681 (2023).
24. Walker, K. A. et al. Large-scale plasma proteomic analysis identifies proteins and pathways associated with dementia risk. *Nat. Aging* **1**, 473–489 (2021).
25. Provinsi, G. et al. A new kid on the block? Carbonic anhydrases as possible new targets in Alzheimer's disease. *Int. J. Mol. Sci.* **20**, 4724 (2019).
26. Canepa, E. et al. FDA-approved carbonic anhydrase inhibitors reduce amyloid beta pathology and improve cognition, by ameliorating cerebrovascular health and glial fitness. *Alzheimers Dement* **19**, 5048–5073 (2023).
27. Nominated Targets. 2024-12-18]; <https://agora.adknowledgeportal.org/genes/nominated-targets> (2024).
28. Sun, B. B. et al. Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).
29. Jiang, T. et al. Soluble TREM1 concentrations are increased and positively correlated with total tau levels in the plasma of patients with Alzheimer's disease. *Aging Clin. Exp. Res.* **31**, 1801–1805 (2019).

30. Hok-A-Hin, Y. S. et al. Neuroinflammatory CSF biomarkers MIF, sTREM1, and sTREM2 show dynamic expression profiles in Alzheimer's disease. *J. Neuroinflammation* **20**, 107 (2023).
31. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinforma.* **9**, 559 (2008).
32. Nieves-Rodriguez, S. et al. Leveraging machine learning to predict Parkinson's disease using pre-symptomatic proteomics data. *Brain awaf303* (2025).
33. Shen, Y. et al. CSF proteomics identifies early changes in autosomal dominant Alzheimer's disease. *Cell* **187**, 6309–6326.e15 (2024).
34. Shah, A. M. et al. Large-scale plasma proteomics identifies novel proteins and protein networks associated with heart failure development. *Nat. Commun.* **15**, 528 (2024).
35. Scott, M. & Su-In, L. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4765–4774 (2017).
36. Pichet Binette, A. et al. Combining plasma phospho-tau and accessible measures to evaluate progression to Alzheimer's dementia in mild cognitive impairment patients. *Alzheimers Res. Ther.* **14**, 46 (2022).
37. Lombardi, G. et al. Role of blood P-Tau isoforms (181, 217, 231) in predicting conversion from MCI to dementia due to Alzheimer's disease: a review and meta-analysis. *Int. J. Mol. Sci.* **25**, 12916 (2024).
38. Hijssen, E. H. et al. Plasma phosphorylated tau 217 and phosphorylated tau 181 as biomarkers in Alzheimer's disease and frontotemporal lobar degeneration: a retrospective diagnostic performance study. *Lancet Neurol.* **20**, 739–752 (2021).
39. Arranz, J. et al. Correction: diagnostic performance of plasma pTau. *Alzheimers Res. Ther.* **16**, 168 (2024).
40. Rahman, M. M. & Lendel, C. Extracellular protein components of amyloid plaques and their roles in Alzheimer's disease pathology. *Mol. Neurodegener.* **16**, 59 (2021).
41. Sun, Y. et al. Role of the extracellular matrix in Alzheimer's disease. *Front. Aging Neurosci.* **13**, 707466 (2021).
42. Hernandez-Guillamon, M. et al. Sequential amyloid-beta degradation by the matrix metalloproteinases MMP-2 and MMP-9. *J. Biol. Chem.* **290**, 15078–15091 (2015).
43. Zipfel, P. et al. Matrix metalloproteinases as new targets in Alzheimer's disease: opportunities and challenges. *J. Med. Chem.* **63**, 10705–10725 (2020).
44. Liu, Y. et al. Matrix metalloproteinase-12 contributes to neuroinflammation in the aged brain. *Neurobiol. Aging* **34**, 1231–1239 (2013).
45. Kivisakk, P. et al. Plasma biomarkers for prognosis of cognitive decline in patients with mild cognitive impairment. *Brain Commun.* **4**, fcac155 (2022).
46. Whelan, C. D. et al. Multiplex proteomics identifies novel CSF and plasma biomarkers of early Alzheimer's disease. *Acta Neuropathol. Commun.* **7**, 169 (2019).
47. Qin, W. et al. Elevated plasma angiogenesis factors in Alzheimer's disease. *J. Alzheimers Dis.* **45**, 245–252 (2015).
48. Lindbohm, J. V. et al. Plasma proteins, cognitive decline, and 20-year risk of dementia in the Whitehall II and atherosclerosis risk in communities studies. *Alzheimers Dement.* **18**, 612–624 (2022).
49. Jakel, L. et al. Altered brain expression and cerebrospinal fluid levels of TIMP4 in cerebral amyloid angiopathy. *Acta Neuropathol. Commun.* **12**, 103 (2024).
50. Koskivirta, I. et al. Tissue inhibitor of metalloproteinases 4 (TIMP4) is involved in inflammatory processes of human cardiovascular pathology. *Histochem. Cell Biol.* **126**, 335–342 (2006).
51. Roberts, J. A. et al. Unbiased proteomics and multivariable regularized regression techniques identify SMOC1, NOG, APCS, and NTN1 in an Alzheimer's disease brain proteomic signature. *NPJ Aging* **9**, 18 (2023).
52. Balcomb, K. et al. SMOC1 colocalizes with Alzheimer's disease neuropathology and delays Abeta aggregation. *Acta Neuropathol.* **148**, 72 (2024).
53. Li-Kroeger, D. et al. The SMOC1 proteomics network M42 controls multiple signaling modalities, brain homeostasis and toxicity in fly Alzheimer's disease models. *Alzheimer's. Dement.* **20**, e089554 (2024). p.
54. Shen, Y. et al. Systematic proteomics in Autosomal dominant Alzheimer's disease reveals decades-early changes of CSF proteins in neuronal death, and immune pathways. <https://doi.org/10.1101/2024.01.12.24301242> (2024).
55. Balcomb, K. et al. SMOC1 colocalizes with Alzheimer's disease neuropathology and delays Aβ aggregation. *Acta Neuropathol.* **148**, 72 (2024).
56. Morato, X. et al. Associations of plasma SMOC1 and soluble IL6RA levels with the progression from mild cognitive impairment to dementia. *Brain Behav. Immun. Health* **42**, 100899 (2024).
57. Johnson, E. C. B. et al. Cerebrospinal fluid proteomics define the natural history of autosomal dominant Alzheimer's disease. *Nat. Med.* **29**, 1979–1988 (2023).
58. Haque, R. et al. A protein panel in cerebrospinal fluid for diagnostic and predictive assessment of Alzheimer's disease. *Sci. Transl. Med.* **15**, eadg4122 (2023).
59. Schroeder, A. & de Wit, J. Leucine-rich repeat-containing synaptic adhesion molecules as organizers of synaptic specificity and diversity. *Exp. Mol. Med.* **50**, 1–9 (2018).
60. Frick, E. A. et al. Serum proteomics reveal APOE-epsilon4-dependent and APOE-epsilon4-independent protein signatures in Alzheimer's disease. *Nat. Aging* **4**, 1446–1464 (2024).
61. Mishra, K. A. & Sethi, K. K. Unveiling tomorrow: Carbonic anhydrase activators and inhibitors pioneering new frontiers in Alzheimer's disease. *Arch. Pharm.* **358**, e2400748 (2025).
62. Cui, X. et al. Comparative characterization of human accelerated regions in neurons. *Nature* **640**, 991–999 (2025).
63. Orellana, A. et al. Establishing in-house cutoffs of CSF Alzheimer's disease biomarkers for the AT(N) stratification of the Alzheimer center Barcelona cohort. *Int. J. Mol. Sci.* **23**, 6891 (2022).
64. Boada, M. et al. Design of a comprehensive Alzheimer's disease clinic and research center in Spain to meet critical patient and family needs. *Alzheimers Dement.* **10**, 409–415 (2014).
65. Alegret, M. et al. Normative data of a brief neuropsychological battery for Spanish individuals older than 49. *J. Clin. Exp. Neuropsychol.* **34**, 209–219 (2012).
66. Petersen, R. C. et al. Mild cognitive impairment: a concept in evolution. *J. Intern. Med.* **275**, 214–228 (2014).
67. Lopez, O. L. et al. Risk factors for mild cognitive impairment in the Cardiovascular Health Study Cognition Study: part 2. *Arch. Neurol.* **60**, 1394–1399 (2003).
68. Petersen, R. C. et al. Mild cognitive impairment: clinical characterization and outcome. *Arch. Neurol.* **56**, 303–308 (1999).
69. de Rojas, I. et al. Common variants in Alzheimer's disease and risk stratification by polygenic risk scores. *Nat. Commun.* **12**, 3417 (2021).
70. Brosseron, F. et al. Soluble TAM receptors sAXL and sTyro3 predict structural and functional protection in Alzheimer's disease. *Neuron* **110**, 1009–1022.e4 (2022).
71. Moreno-Grau, S. et al. Genome-wide association analysis of dementia and its clinical endophenotypes reveal novel loci associated with Alzheimer's disease and three causality networks: the GR@ACE project. *Alzheimers Dement.* **15**, 1333–1347 (2019).
72. Alegret, M. et al. Cut-off scores of a brief neuropsychological battery (NBACE) for Spanish individual adults older than 44 years old. *PLoS One* **8**, e76436 (2013).
73. Bos, I. et al. The EMIF-AD Multimodal Biomarker Discovery study: design, methods and cohort characteristics. *Alzheimers Res. Ther.* **10**, 64 (2018).
74. Petersen, R. C. Mild cognitive impairment as a diagnostic entity. *J. Intern. Med.* **256**, 183–194 (2004).

75. McKhann, G. et al. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* **34**, 939–944 (1984).
76. Candia, J. et al. Assessment of variability in the plasma 7k SomaScan proteomics assay. *Sci. Rep.* **12**, 17147 (2022).
77. Haslam, D. E. et al. Stability and reproducibility of proteomic profiles in epidemiological studies: comparing the Olink and SOMAScan platforms. *Proteomics* **22**, e2100170 (2022).
78. Yang, C. et al. Genomic atlas of the proteome from brain, CSF and plasma prioritizes proteins implicated in neurological disorders. *Nat. Neurosci.* **24**, 1302–1312 (2021).
79. Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinforma.* **17**, 1–13 (2016).
80. Liao, Y. et al. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* **47**, W199–W205 (2019).
81. Lin, H., & Zelterman, D. Modeling Survival Data: Extending the Cox Model (Taylor & Francis, 2002).
82. Yu, G. et al. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
83. Szklarczyk, D. et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).
84. Gautier, L. rpy2: A simple and efficient access to R from Python. Available from: <https://rpy2.github.io> (2008).
85. Buitinck, L. et al. API design for machine learning software: experiences from the scikit-learn project. <https://doi.org/10.48550/arXiv.1309.0238> (2013).
86. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1 (2010).
87. Simon, N. et al. Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* **39**, 1 (2011).
88. Tay, J. K., Narasimhan, B. & Hastie, T. Elastic net regularization paths for all generalized linear models. *J. Stat. Soft.* **106**, 1–31 (2023).
89. Ishwaran, H. & Kogalur, U. B. Fast unified random forests for survival, regression, and classification (RF-SRC). R package version, Vol. 2 (2019).
90. Pölsterl, S. Scikit-survival: a library for time-to-event analysis built on top of scikit-learn. *J. Mach. Learn. Res.* **21**, 1–6 (2020).
91. Kassambara, A., Kosinski, M. & Biecek, P. Survminer: Drawing Survival Curves using 'ggplot2'. CRAN: Contributed Packages (2016).
92. Davidson-Pilon, C. Lifelines: survival analysis in Python. *J. Open Source Softw.* **4**, 1317 (2019).
93. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
94. Wickam, H. *ggplot2: Elegant Graphics for Data Analysis*. Vol. 16 (Springer-Verlag, 2021).

Acknowledgements

We thank the F.ACE and the EMIF-AD MBD participants for their invaluable contributions to these research resources. We would like to thank Dr. Xiangwei Peng for his contribution in developing RFE customized python codes for feature selection in model building.

Author contributions

Y.Z. conceptualized and designed the study with input from B.S., C.W., S.M. and S.L.. J.X. and Y.Z. developed analysis plan and interpreted results. J.X. developed codes and performed most of the analysis. Y.Z. performed clinical data processing for EMIF-AD MBD dataset. B.S managed the datasets of F.ACE and EMIF-AD cohorts. K.H., Y.Z. and L.H. compiled cis-MR results using CSF and plasma pQTLs. R.L.A, R.P. and A.C. manage and maintain the F.ACE cohort follow ups, clinical data records, sample collection and data generation. C.S.A. led the collaboration between F.ACE institute and J&J, and sponsored the F.ACE cohort proteomics data generation. J.X. drafted the first version of manuscript with inputs from Y.Z. All authors provided critical review on the manuscript. The company legal team reviewed the intellectual content and approved the submitted version.

Competing interests

The authors declare no competing interests. J.X., K.Y.H., L.H., B.S., S.M., Z.S., A.C.S., S.L. and Y.Z. are employees of Johnson & Johnson. C.D.W. was employed by Johnson & Johnson when this study was performed and is now employed by RCSI University of Medicine and Health Sciences.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s44400-025-00053-9>.

Correspondence and requests for materials should be addressed to Yanfei Zhang.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025