


RESEARCH

Open Access



Soluble amyloid-beta isoforms predict downstream Alzheimer's disease pathology

Guilherme Povala^{1,2}, Bruna Bellaver¹, Marco Antônio De Bastiani¹, Wagner S. Brum¹, Pamela C. L. Ferreira^{1,3}, Andrei Bieger¹, Tharick A. Pascoal³, Andrea L. Benedet^{4,5}, Diogo O. Souza^{1,6}, Ricardo M. Araujo², Bruno Zatt², Pedro Rosa-Neto^{4,5,7*}, Eduardo R. Zimmer^{1,8,9*}  and for the Alzheimer's Disease Neuroimaging Initiative

Abstract

Background: Changes in soluble amyloid-beta (A β) levels in cerebrospinal fluid (CSF) are detectable at early preclinical stages of Alzheimer's disease (AD). However, whether A β levels can predict downstream AD pathological features in cognitively unimpaired (CU) individuals remains unclear. With this in mind, we aimed at investigating whether a combination of soluble A β isoforms can predict tau pathology (T+) and neurodegeneration (N+) positivity.

Methods: We used CSF measurements of three soluble A β peptides (A β_{1-38} , A β_{1-40} and A β_{1-42}) in CU individuals (n = 318) as input features in machine learning (ML) models aiming at predicting T+ and N+. Input data was used for building 2046 tuned predictive ML models with a nested cross-validation technique. Additionally, proteomics data was employed to investigate the functional enrichment of biological processes altered in T+ and N+ individuals.

Results: Our findings indicate that A β isoforms can predict T+ and N+ with an area under the curve (AUC) of 0.929 and 0.936, respectively. Additionally, proteomics analysis identified 17 differentially expressed proteins (DEPs) in individuals wrongly classified by our ML model. More specifically, enrichment analysis of gene ontology biological processes revealed an upregulation in myelination and glucose metabolism-related processes in CU individuals wrongly predicted as T+. A significant enrichment of DEPs in pathways including biosynthesis of amino acids, glycolysis/gluconeogenesis, carbon metabolism, cell adhesion molecules and prion disease was also observed.

Conclusions: Our results demonstrate that, by applying a refined ML analysis, a combination of A β isoforms can predict T+ and N+ with a high AUC. CSF proteomics analysis highlighted a promising group of proteins that can be further explored for improving T+ and N+ prediction.

Keywords: Alzheimer's disease, Amyloid-beta, Tau pathology, Neurodegeneration, Machine learning, Proteomics

Background

Alzheimer's disease (AD) is the most prevalent neurodegenerative disease worldwide [1]. Its main neuropathological features involve the deposition of two proteins,

amyloid- β (A β) and tau, into insoluble aggregates in the brain [2, 3]. Indeed, the most accepted AD theoretical model suggests that A β dysmetabolism triggers a cascade of downstream pathological events, including tau pathology, synaptic dysfunction, and neurodegeneration, which leads to cognitive decline and, ultimately, to dementia [4, 5].

This theoretical model relies on data derived from cross-sectional and longitudinal multicentric studies using multiple biomarkers. Currently, AD biomarkers are divided into two main classes: biofluid-based [blood and

*Correspondence: pedro.rosa@mcgill.ca; erzimmer@gmail.com

¹ Graduate Program in Biological Sciences: Biochemistry, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil

⁴ Translational Neuroimaging Laboratory, The McGill University Research Centre for Studies in Aging, 6825 LaSalle Boulevard, Verdun, QC H4H 1R3, Canada

Full list of author information is available at the end of the article



cerebrospinal fluid (CSF)] and neuroimaging [magnetic resonance imaging (MRI) and positron emission tomography (PET)] [6]. These biomarkers constitute the basis of the National Institute on Aging-Alzheimer's Association (NIA-AA) Research Framework proposed for clinical studies, which adopted the A/T/(N) system for amyloid, tau, and neurodegeneration biomarkers [7]. In each category, biomarkers are dichotomized to indicate a normal or abnormal status [7].

Importantly, this system relies on the amyloid cascade hypothesis, i.e., the linear chain A β positivity (A+) \rightarrow tau positivity (T+) \rightarrow neurodegeneration positivity (N+) \rightarrow cognitive symptoms [4, 5]. However, around 30% of cognitively unimpaired (CU) individuals are A+ but do not present any other AD pathological features [8–10]. Thus, A+, usually indexed by CSF A β_{1-42} or PET, does not infer *per se* if an individual presents or will develop tau pathology or neurodegeneration. Therefore, it is clear that other biological processes are also critical in the progression toward clinical symptoms.

In this study, we asked (i) whether a combination of A β isoforms, measured in the CSF, would be capable of predicting downstream pathological biomarkers and (ii) what biological processes are related to an increase in A β isoforms' prediction power over downstream AD pathology. To answer these inquiries, we aimed at predicting T+ and N+ using a combination of demographics and A β isoforms levels in the CSF (A β_{1-38} , A β_{1-40} , and A β_{1-42}) as input features in machine learning models (ML). We also evaluated whether CSF proteomic analyses could reveal altered biological processes heterogeneity in individuals wrongly classified in ML models.

Methods

ADNI description

Data used in this article are available at the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). ADNI is a longitudinal multicentric study launched in 2004, as a result of a public-private partnership, including the Foundation for the National Institutes of Health and the National Institute on Aging alongside contributors from many other sources. The study is currently in its 4th phase (ADNI1, ADNI GO, ADNI2, and ADNI3) and has recruited over 2300 participants in North America, to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of AD. More information on the study design can be found in adni.loni.usc.edu/about/.

Eligibility criteria

In this study, data from 318 CU subjects were collected from ADNI1 and ADNI2 database. Specific criteria for

inclusion in this study were the availability of CSF levels of A β_{1-38} , A β_{1-40} , and A β_{1-42} proteins measured by 2D-ultra-performance liquid chromatography-tandem mass spectrometry (2D-UPLC-MS/MS). ADNI's inclusion and diagnostic criteria have been described elsewhere [11].

CSF biomarker collection and analysis

CSF A β_{1-38} , A β_{1-40} , and A β_{1-42} peptide levels were measured using the 2D-UPLC-MS/MS method (Waters® XEVO-TQ-S), which had been previously described [12] and has been recently revalidated. This updated technique has been recognized as an accepted analytical reference by the Joint Committee for Traceability in Laboratory Medicine (JCTLM), in whose database it was published under the JCTLM Identification Number C12RMP1. For defining T+ and N+, p-tau (Thr-181) and t-tau levels used in this study were measured by the Elecsys® immunoassay, with T+ defined as CSF p-tau (181-Thr) > 19.2 pg/mL and N+ defined as CSF t-tau > 242 pg/mL [13]. Data for the 2D-UPLC-MS/MS and Elecsys® methods are available, respectively, at the ADNI database under the file names "UPENNMSMSABETA.csv" and "UPENNBIOMK9_04_19_17.csv".

Statistical analysis

All statistical analyses were performed in GraphPad Prism 8. Data are expressed as mean \pm standard deviation (SD). Normality was evaluated using histograms and quantile plots. Because samples did not have Gaussian distributions, comparisons between groups were carried out using MannWhitney test. P-values of less than 0.05 were reported as statistically significant.

Machine learning framework

We developed a ML framework that combines multiple techniques and models to predict T+ and N+ with the use of CSF A β isoform levels, demographic information and APOE $\epsilon 4$ status. The framework was coded in Python (version 3.6.8, <https://www.python.org/>), using the scikit-learn (version 0.20.2, <https://scikit-learn.org/>) and xgboost (version 0.81, <https://xgboost.readthedocs.io/>) libraries. The supervised ML algorithms used in our framework are composed of Logistic Regression, Naive Bayes, k-Nearest Neighbors (kNN), Support Vector Classifier (SVC), Decision Trees, Random Forest, Gradient Boosting, XGBoost, and AdaBoost.

As input features for our framework, we used A β peptide levels (A β_{1-38} , A β_{1-40} , and A β_{1-42}), demographic information (age, sex and years of education), and APOE $\epsilon 4$ status. For feature selection, we evaluated all possible feature combinations, generating 1023 subsets. For each feature subset, we performed the nested cross-validation

Table 1 Hyperparameters evaluated for the machine learning models

Algorithm	Fixed parameters	Iterated parameters
Logistic Regression	solver: lbfgs max_iter: 250 penalty: l2	C: [0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000]
Naive Bayes	–	–
kNN	algorithm: ball_tree leaf_size: 50	n_neighbors: [1,2,3,4,5,6,7,8,9] p: [1,2]
SVC	–	for kernels: [rbf, poly, sigmoid] C: [–4, –3, –2, –1, 0, 1, 2, 3] for kernel: linear gamma: [0.00001, 0.0001, 0.001, 0.01, 0.1] C: [0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000]
Decision Trees	–	max_depth: [1,2,3,4,5,6,7,8,9] criterion: [gini, entropy]
Random Forest	–	max_depth: [3,4,5,8,10] n_estimators: [5, 20, 50, 100, 200, 500, 1000]
Gradient Boosting	–	max_depth: [3,4,5,8,10] learning_rate: [0.01, 0.05, 0.1, 0.2] n_estimators: [5, 20, 50, 100, 200, 500, 1000]
XGBoost	–	max_depth: [6,7,8] learning_rate: [0.01, 0.025, 0.05, 0.075, 0.1] n_estimators: [5, 20, 50, 100, 200, 500, 1000]
AdaBoost	–	learning_rate: [0.25, 0.5, 1.0, 1.25, 1.5] n_estimators: [20, 50, 100, 150, 200]

kNN: k-Nearest Neighbors; SVC: Support Vector Classifier

(CV) technique. Here, we used the nested CV since we needed to train different ML models together with its hyperparameter optimization. The nested CV has an inner CV loop nested in outer CV. The inner loop is composed of a 2-fold CV, and it is responsible for model selection and hyperparameter tuning, which is similar to a validation set. The outer loop, however, is composed of a 5-fold CV and it is used for error estimation, as a test set. The nested cross-validation uses the area under the curve (AUC) metric to select the best hyperparameters and models. Then, an independent test set is used to test the overall performance of the best model and to generate the AUC result. The hyperparameters evaluated for each ML algorithm used in this work are shown in Table 1. After obtaining the AUC results for tuned ML algorithms with the nested cross-validation, only the model that presented the best performance is chosen for each feature subset. Among all these models, we selected the best one and then extracted the AUC for the independent test set.

CSF proteomics analysis

Processed CSF proteomics data were collected from the ADNI database. Samples were measured using the LC/MS-MRM method [12]. Proteins and peptides were selected based upon their previous detection in CSF, relevance to AD, and previous results from the Rules Based

Medicine (RBM) multiplex immunoassay analysis of ADNI CSF. The final MRM panel consisted of 567 peptides representing 221 proteins. From these 567 peptides, 320 were detectable in > 10% of ADNI samples and are available in the file “CSFMRM.csv”.

From the previously included CU individuals, only 76 presented CSF proteomics data in the ADNI database and were included in further analyses. CSF proteomics analysis was performed comparing T– (n = 55) and T+ (n = 21) individuals and N– (n = 57) and N+ (n = 19). All proteomic analyses were implemented in an R statistical environment. Differentially expressed analysis was computed for T–/T+ and N–/N+ groups independently, using the LIMMA (version 3.46.0) package [14], and considering FDR-adjusted p-value < 0.05 as differentially expressed proteins (DEP) criteria. Finally, functional enrichment analyses of gene ontology (GO) biological processes and KEGG pathways were computed and visualized using the clusterProfiler (version 3.18.1) and Gplot (version 1.0.2) packages [15, 16].

Results

Sample characteristics

We included 318 CU individuals from ADNI, whose CSF had been analyzed with 2D-UPLC-MS/MS. Characteristics of the ADNI cohort and the different A, T, and N status of samples are provided in Table 2. Population characteristics were compared between positive and

Table 2 Sample characteristics

Characteristic	CU	A-	A+	T-	T+	N-	N+
Number of individuals	318	60	50	52	58	67	43
Sex (% female)	50%	51.67%	48%	50%	50%	50.75%	48.84%
Age (y)	75.66 ± 5.22	75.37 ± 5.67	76.01 ± 4.67	73.87 ± 4.54	77.26 ± 5.32 ^{b***}	74.46 ± 4.66	77.52 ± 5.55 ^{c**}
Education (y)	15.73 ± 2.83	15.42 ± 2.68	16.1 ± 2.99	15.77 ± 2.77	15.69 ± 2.91	15.57 ± 3.1	15.98 ± 2.37
MMSE	29.08 ± 1.03	28.98 ± 1.1	29.2 ± 0.95	29.13 ± 0.93	29.03 ± 1.12	29.01 ± 1.05	29.19 ± 1.01
ADAS-Cog	6.42 ± 2.92	6.09 ± 2.91	6.81 ± 2.92	6.18 ± 2.85	6.64 ± 2.99	6.22 ± 2.69	6.73 ± 3.27
APOE ε4 carriers (%)	24.55%	11.67%	40% ^{a***}	15.38%	32.76% ^{b*}	19.40%	32.56%

CU: Cognitively Unimpaired; A+: Amyloid-beta positive; A-: Amyloid-beta negative; T+: Tau positive; T-: Tau negative; N+: Neurodegeneration positive; N-: Neurodegeneration negative; y: year; MMSE: Mini-Mental State Examination; ADAS-Cog: Alzheimer’s Disease Assessment Scale-Cognitive subscale. Statistical differences for numerical characteristics were tested using t test. Statistical differences for sex and APOE status were tested using Fisher’s exact test. (*p < 0.05, **p ≤ 0.01, ***p ≤ 0.001)

^a significantly different from A-, ^b significantly different from T-, ^c significantly different from N-

negative group status for each of the above-mentioned biomarker categories. A+ and T+ showed significantly more APOE ε4 carriers than Aβ negative (A-) and tau negative (T-) groups. As already observed in previous studies, APOE ε4 carriers are associated with decreased Aβ₁₋₄₂ and elevated p-tau in the CSF [14, 15]. T+ and N+ presented elevated age, when compared with T- and neurodegeneration negative (N-) groups, respectively. No significant differences were observed in sex, years of education, Mini-Mental State Examination (MMSE), and Alzheimer’s Disease Assessment Scale-Cognitive Subscale (ADAS-Cog) among groups.

Changes in Aβ soluble isoforms in T+ and N+ CU individuals

Figure 1 compares Aβ isoform levels and their respective ratios between T+ and T- (Fig. 1a), and N+ and N- (Fig. 1b). When comparing T status, T+ group presented higher levels of Aβ₁₋₃₈ (Fig. 1c, T- = 1764 ± 496.1 pg/mL, T+ = 2411 ± 566.95 pg/mL, p < 0.0001) and Aβ₁₋₄₀ (Fig. 1d, T- = 7617 ± 2052 pg/mL, T+ = 10,424 ± 2529 pg/mL, p < 0.0001). Additionally, a decrease in Aβ₁₋₄₂/Aβ₁₋₄₀ (Fig. 1f, T- = 0.1749 ± 0.05, T+ = 0.1381 ± 0.06, p < 0.0001) and Aβ₁₋₄₂/Aβ₁₋₃₈ ratios (Fig. 1g, T- = 0.7610 ± 0.22, T+ = 0.6014 ± 0.25, p < 0.0001) was observed in T+ individuals. However, we did not observe any significant difference in Aβ₁₋₄₂ levels (Fig. 1e, T- = 1353 ± 559.4 pg/mL, T+ = 1492 ± 784 pg/mL, p = 0.41) and Aβ₁₋₄₀/Aβ₁₋₃₈ ratio (Fig. 1h, T- = 4.354 ± 0.42, T+ = 4.329 ± 0.35, p = 0.60) between T+ and T- groups.

For N+ individuals, Aβ₁₋₃₈ (Fig. 1i, N- = 1760 ± 469.6 pg/mL, N+ = 2503 ± 567.2 pg/mL, p < 0.0001), Aβ₁₋₄₀ (Fig. 1j, N- = 7593 ± 1945 pg/mL, N+ = 10,838 ± 2503 pg/mL, p < 0.0001), and Aβ₁₋₄₂ (Fig. 1k, N- = 1328 ± 565.1 pg/mL, N+ = 1575 ± 778.8 pg/mL, p = 0.03) measures were significantly elevated when compared to N-, along with a decrease in Aβ₁₋₄₂/Aβ₁₋₄₀ ratio

(Fig. 1l, N- = 0.1720 ± 0.05, N+ = 0.1411 ± 0.05, p < 0.0001) and Aβ₁₋₄₂/Aβ₁₋₃₈ ratio (Fig. 1m, N- = 0.7483 ± 0.23, N+ = 0.6146 ± 0.25, p < 0.0001). By contrast, Aβ₁₋₄₀/Aβ₁₋₃₈ ratio (Fig. 1n, N- = 4.350 ± 0.41, N+ = 4.336 ± 0.35, p = 0.78) does not differ between N+ and N- groups.

To test whether single Aβ isoforms or its ratios can predict downstream AD pathological processes in CU individuals, we used logistic regression models. The AUC results for predicting T+ and N+ individuals are shown in Table 3. Among all results, Aβ₁₋₃₈ and Aβ₁₋₄₀ seem to be the most reliable features to predict T+, with an AUC of 0.811 for both Aβ isoforms. For predicting N+, Aβ₁₋₃₈ and Aβ₁₋₄₀ showed similar results, with AUCs of 0.847 and 0.855, respectively. On the other hand, Aβ₁₋₄₂ presented an AUC of 0.580 for predicting N+ and 0.529 for T+.

Machine learning framework

Aiming at better predictive models, we proposed a ML framework, which is presented in Fig. 2. Aβ isoforms in the CSF (Aβ₁₋₃₈, Aβ₁₋₄₀, and Aβ₁₋₄₂; measured by 2DUPLCMS/MS), APOE ε4 carrier status, and demographic information (age, sex, and years of education) were used as input features. Besides, for feature generation, Aβ isoforms were used either alone or combined in ratios (Fig. 2a). In the feature subset generation step (Fig. 2b), all possible combinations of features were created (1023 different subsets). Then, for each subset, two models were selected using the nested CV technique (Fig. 2c): one for T+ prediction and another to predict N+ (Fig. 2d).

In our ML framework, to choose the best model for each subset to classify T+ and N+, we evaluated the use of the following ML algorithms: Logistic Regression, Naïve Bayes, kNN, SVC, Decision Trees, Random Forest,

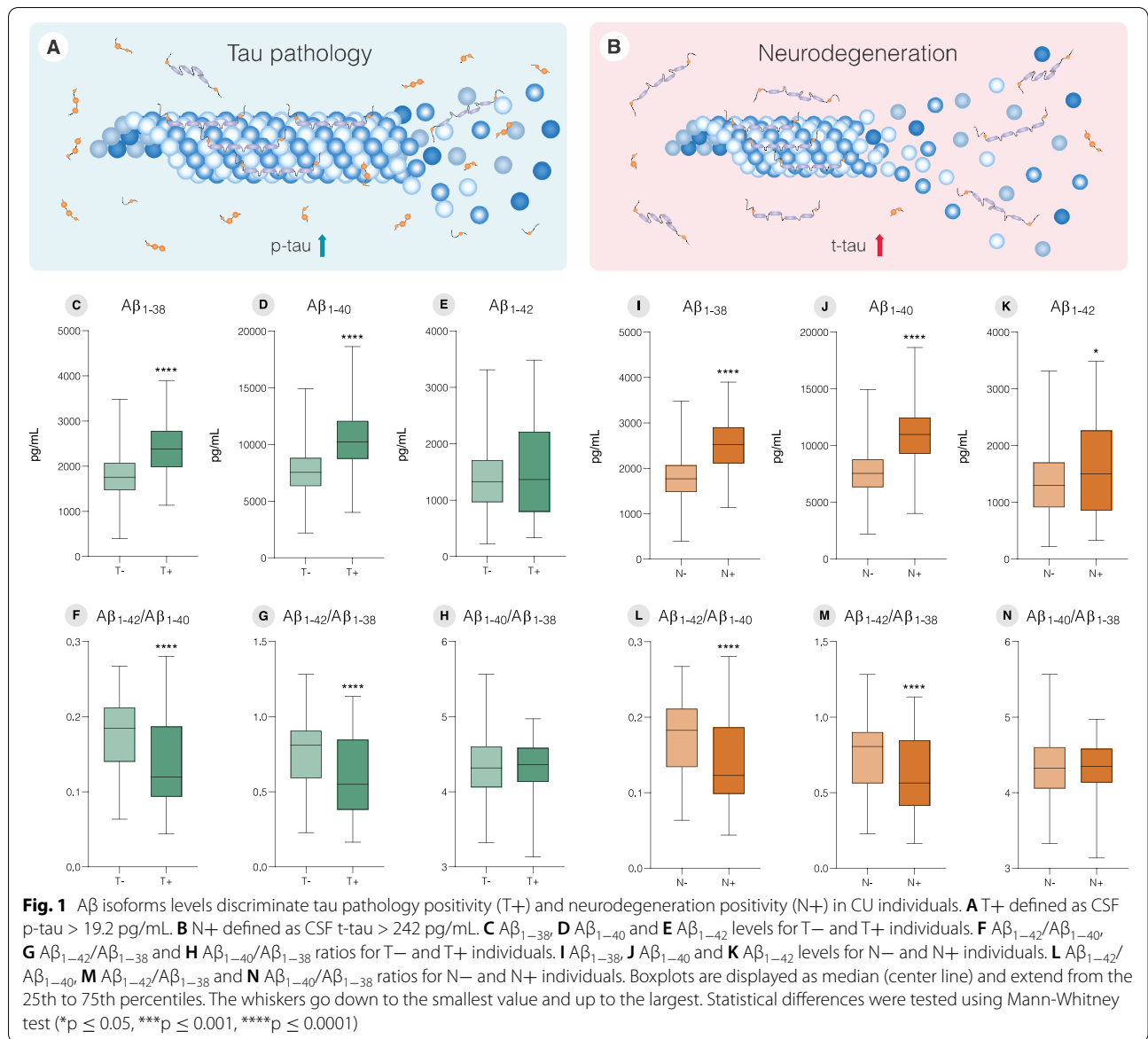


Table 3 AUC results for predicting T+ and N+ in CU individuals using single Aβ isoforms or its ratios

Prediction	Aβ ₁₋₃₈	Aβ ₁₋₄₀	Aβ ₁₋₄₂	Aβ ₁₋₄₂ /Aβ ₁₋₄₀	Aβ ₁₋₄₂ /Aβ ₁₋₃₈	Aβ ₁₋₄₀ /Aβ ₁₋₃₈
T+	0.811	0.811	0.529	0.693	0.682	0.484
N+	0.847	0.855	0.580	0.663	0.652	0.479

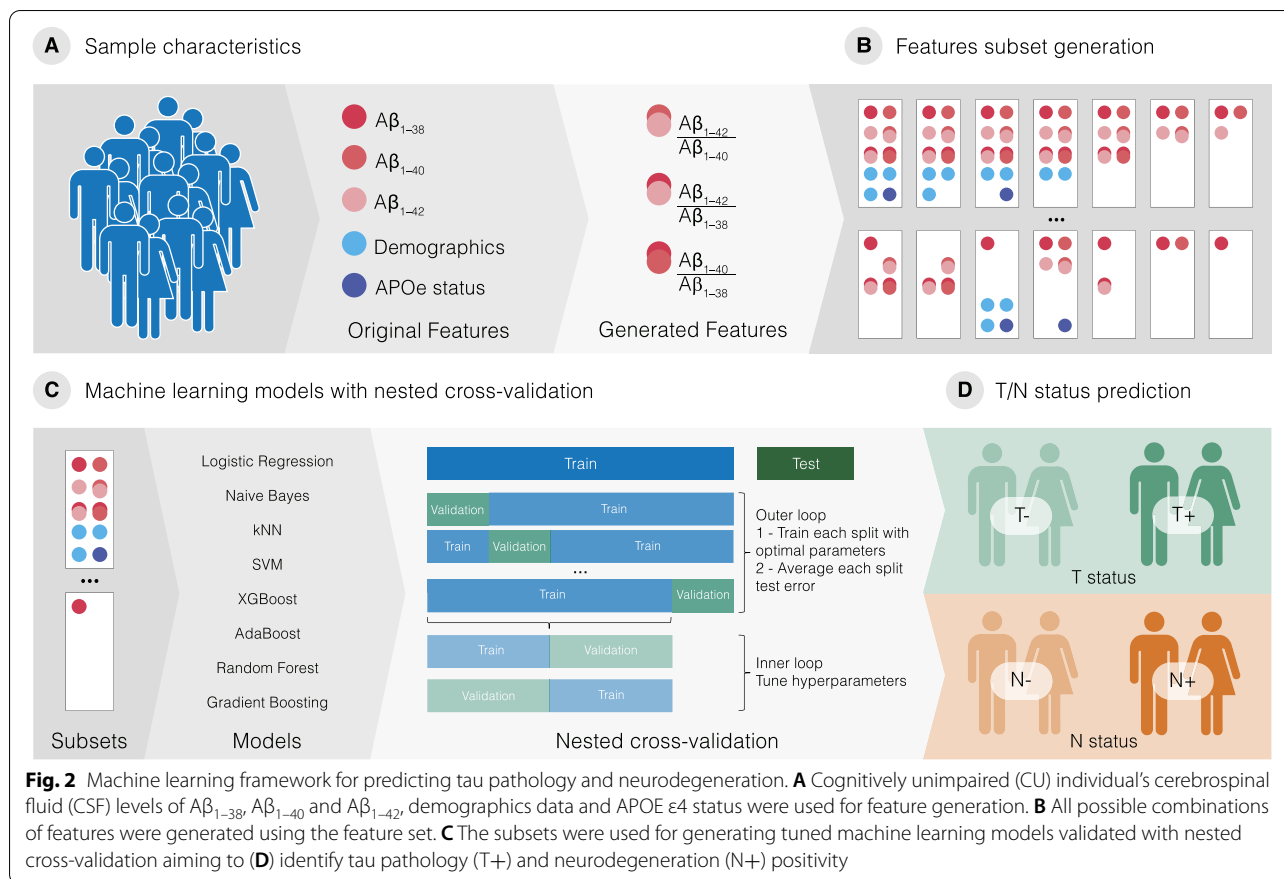
T+: Tau positive; N+: Neurodegeneration positive; Aβ: Amyloid-beta

Gradient Boosting, XGBoost, and AdaBoost within the nested CV technique. For each subset, the best model was defined based on the model’s AUC obtained from the validation set. The top 1 model among the 1023 models (one for each subset) was evaluated using an independent

test set and was defined as the best model to predict T+ or N+.

Tau pathology positivity prediction

From our proposed ML framework, 1023 tuned ML models were generated for predicting T+ (Additional



file 1). Figure 3a shows the AUC results for predicting T+ horizontally ordered by AUC – SD. In Fig. 3b, the best 10 models are ranked. Among the 10 models, all of them presented similar results, ranging from 0.877 to 0.887.

The top 1 model was a logistic regression model using $A\beta_{1-42}$, $A\beta_{1-42}/A\beta_{1-40}$, $A\beta_{1-42}/A\beta_{1-38}$, $A\beta_{1-40}/A\beta_{1-38}$, and years of education as input features. The AUC result obtained for the validation set was 0.881 ± 0.024 . For the independent test set, we achieved an AUC of 0.929 (Fig. 3c).

Neurodegeneration positivity prediction

For N+ prediction, we generated another 1023 models using the same method (Additional file 2). The AUC results for the N+ predictions are shown in Fig. 3d horizontally ordered by AUC – SD. The best 10 models were ranked and plotted on the graph represented in Fig. 3e. The best 10 models presented similar results, ranging from 0.909 to 0.915.

A kNN generated the best results, which had $A\beta_{1-42}$, $A\beta_{1-40}$, $A\beta_{1-42}/A\beta_{1-40}$, $A\beta_{1-42}/A\beta_{1-38}$, and $A\beta_{1-40}/A\beta_{1-38}$ as input features. The AUC result for the validation set for this model was 0.915 ± 0.018 . The independent test set achieved an AUC of 0.936 (Fig. 3f).

CSF proteomics of T+ and N+ CU individuals

To address T+ and N+ CU individuals' functional changes in biological processes, we performed CSF-based proteomics analyses. A total of 112 DEPs were observed in the CSF of CU T+ compared to T- subjects (Additional file 3). The enrichment analysis of GO biological processes in T+ individuals evidenced processes related to myelination, synapse and neurogenesis regulation, immune response, carbohydrate metabolism, memory and learning, and glial cell differentiation (Fig. 4a). Figure 4b depicts top 20 GO terms enriched in T+ subjects compared to T-. To identify the most affected pathways related to changes in proteomics profile of T+, we performed an enrichment analysis using canonical pathways described in the KEGG pathway database [17]. This revealed a significant enrichment of 112 DEPs in 4 signaling pathways: "cell adhesion molecules", "biosynthesis of amino acids", "carbon metabolism", and "prion disease" (Fig. 4c-g). Regarding proteomics analysis of N+, we identified 123 DEPs when compared to N- individuals (Additional file 4). Of note, T+ and N+ subjects share 101 DEPs. Functional enrichment analyses revealed an overlap of enriched GO terms in N+ individuals and T+

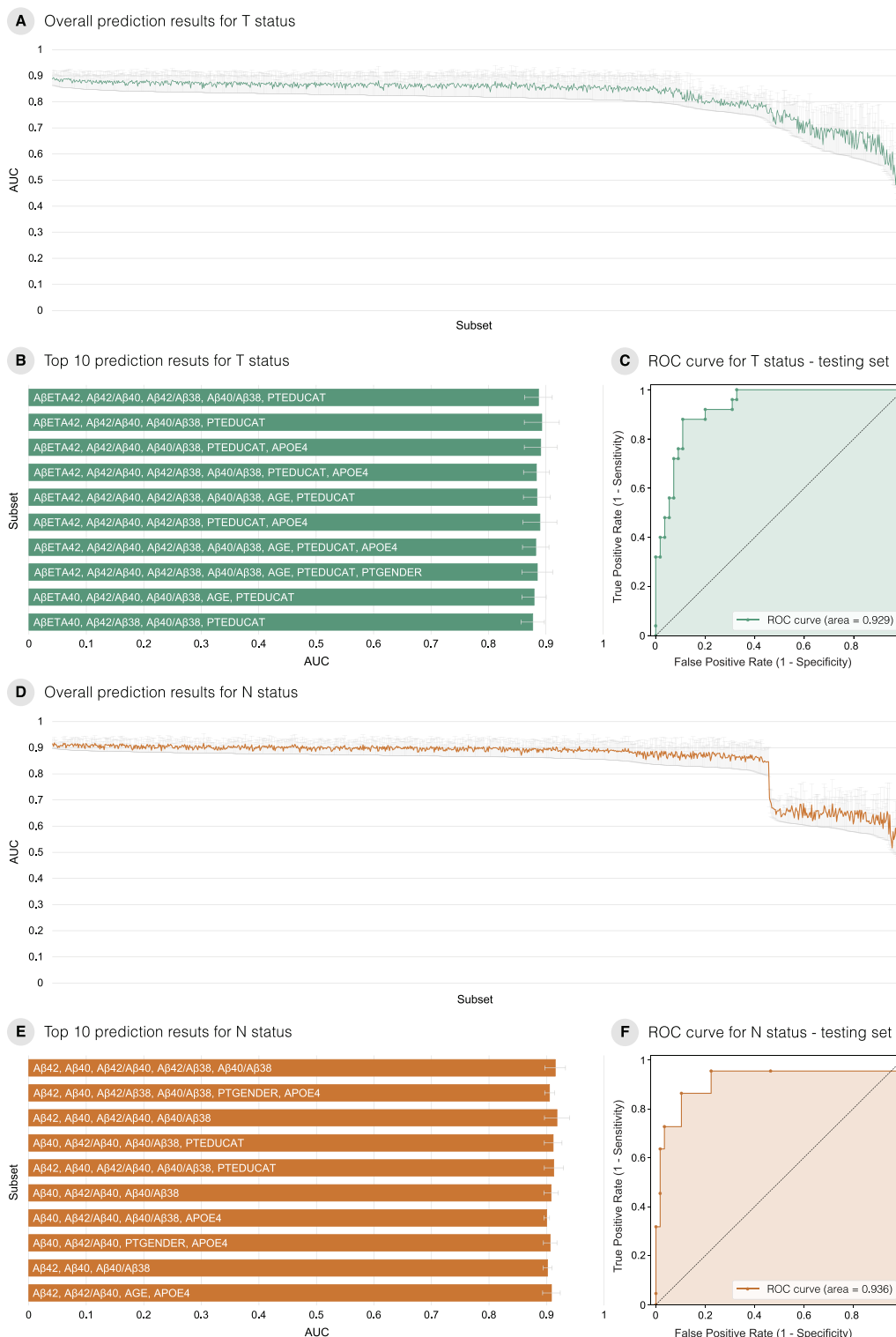
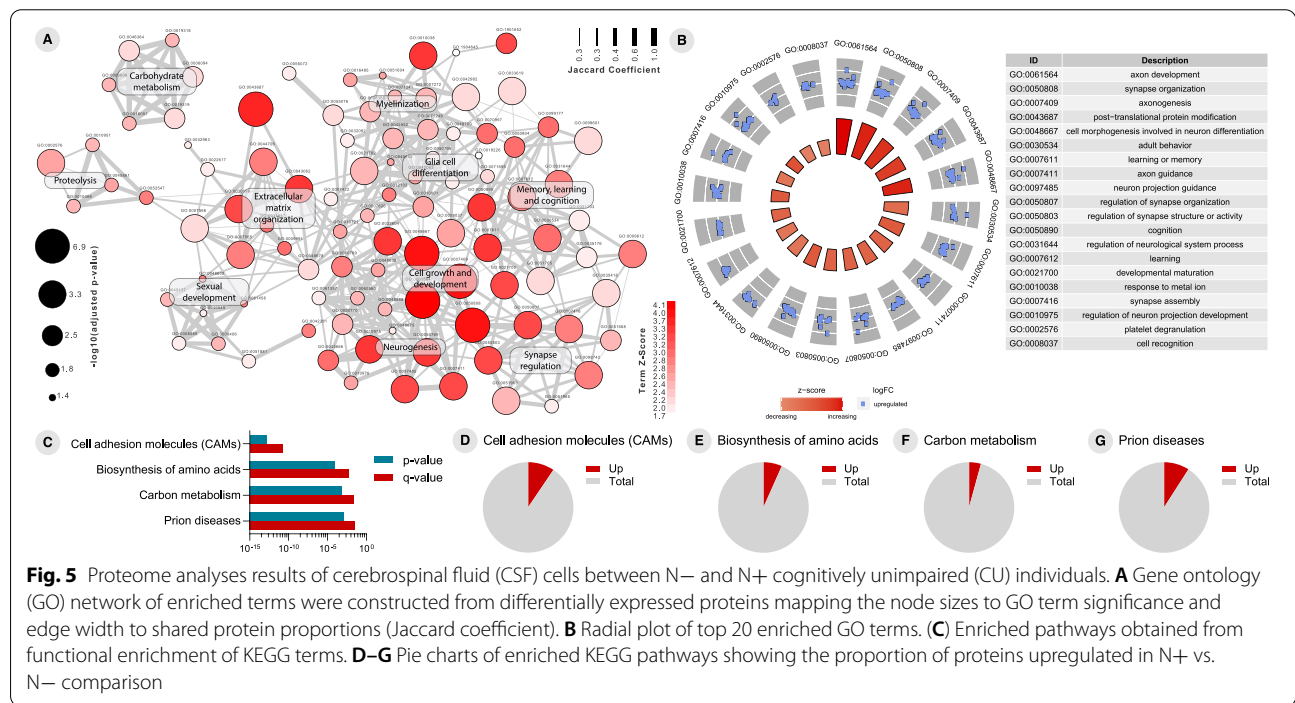
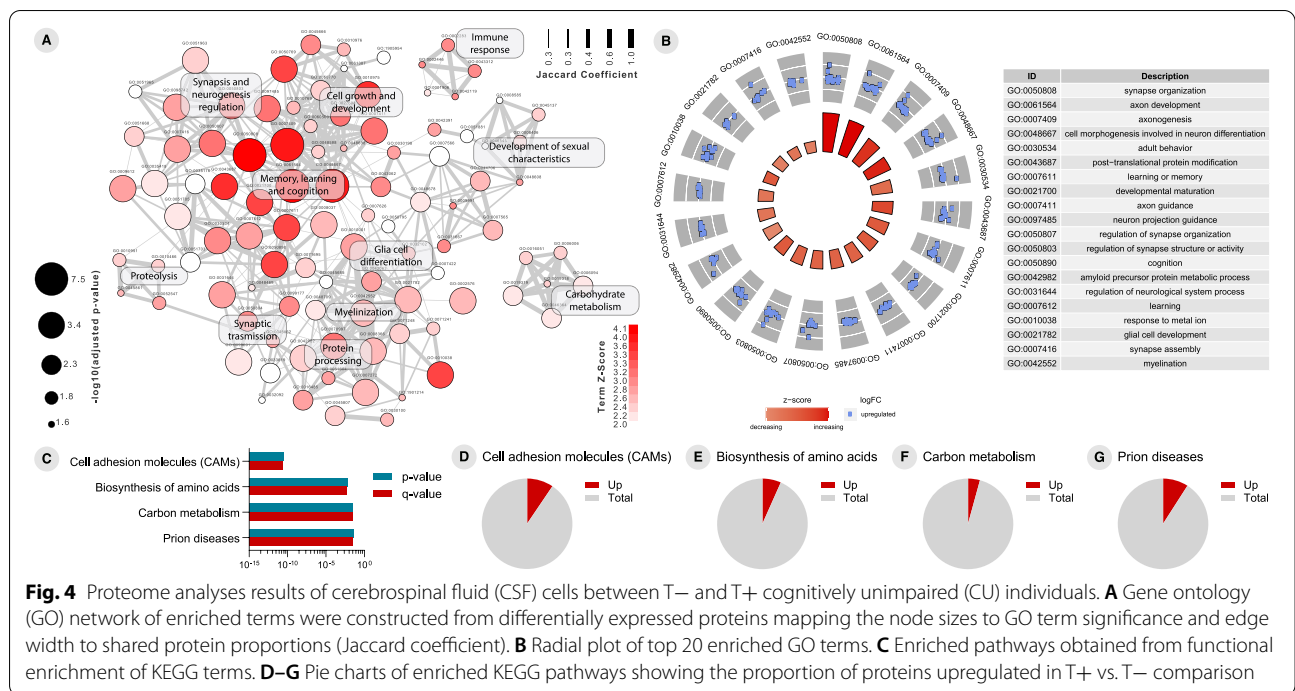


Fig. 3 Results for predicting tau pathology (T) and neurodegeneration (N) status. **A** Area under the ROC curve (AUC) results (vertical axis) for all 1023 subsets to predict T status ordered by AUC – standard deviation (SD). **B** AUC results (horizontal axis) for the top 10 models (vertical axis) to predict T status. **C** ROC curve for the best model to predict T status using the independent test set. **D** AUC results (vertical axis) for all 1023 subsets to predict N status ordered by AUC – SD. **E** AUC results (horizontal axis) for the top 10 models (vertical axis) to predict N status. **F** ROC curve for the best model to predict N status using the independent test set



individuals (Fig. 5a). Synapse organization, learning and memory processes, and APP metabolic processes are among the top 20 GO terms enriched in N+ (Fig. 5b). Interestingly, the same 4 KEGG pathways enriched for T+ were found enriched for N+ individuals (Fig. 5c–g).

CSF proteomics analysis for ML wrong predictions

Because Aβ isoforms predicted T+ and N+ outcomes with an AUC of up to 0.936, we next aimed, with a second proteomics analysis, at identifying differences in biological processes occurring in CU individuals that were

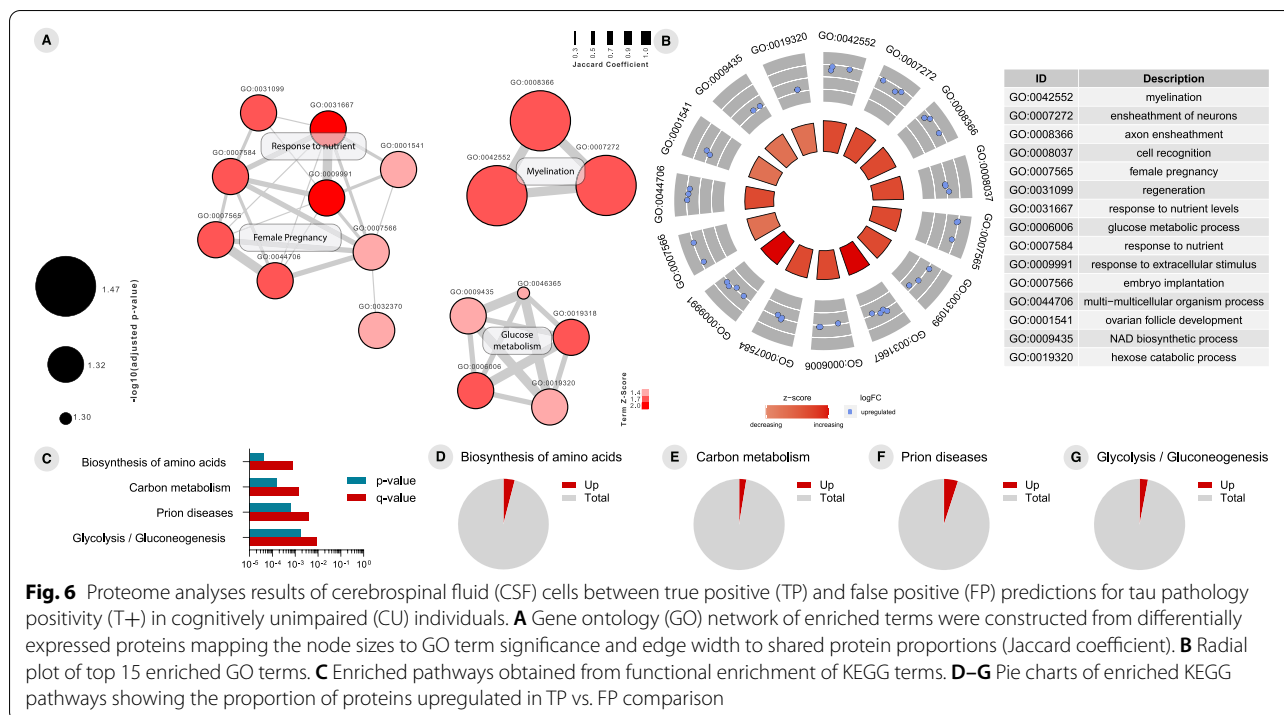


Fig. 6 Proteome analyses results of cerebrospinal fluid (CSF) cells between true positive (TP) and false positive (FP) predictions for tau pathology positivity (T+) in cognitively unimpaired (CU) individuals. **A** Gene ontology (GO) network of enriched terms were constructed from differentially expressed proteins mapping the node sizes to GO term significance and edge width to shared protein proportions (Jaccard coefficient). **B** Radial plot of top 15 enriched GO terms. **C** Enriched pathways obtained from functional enrichment of KEGG terms. **D–G** Pie charts of enriched KEGG pathways showing the proportion of proteins upregulated in TP vs. FP comparison

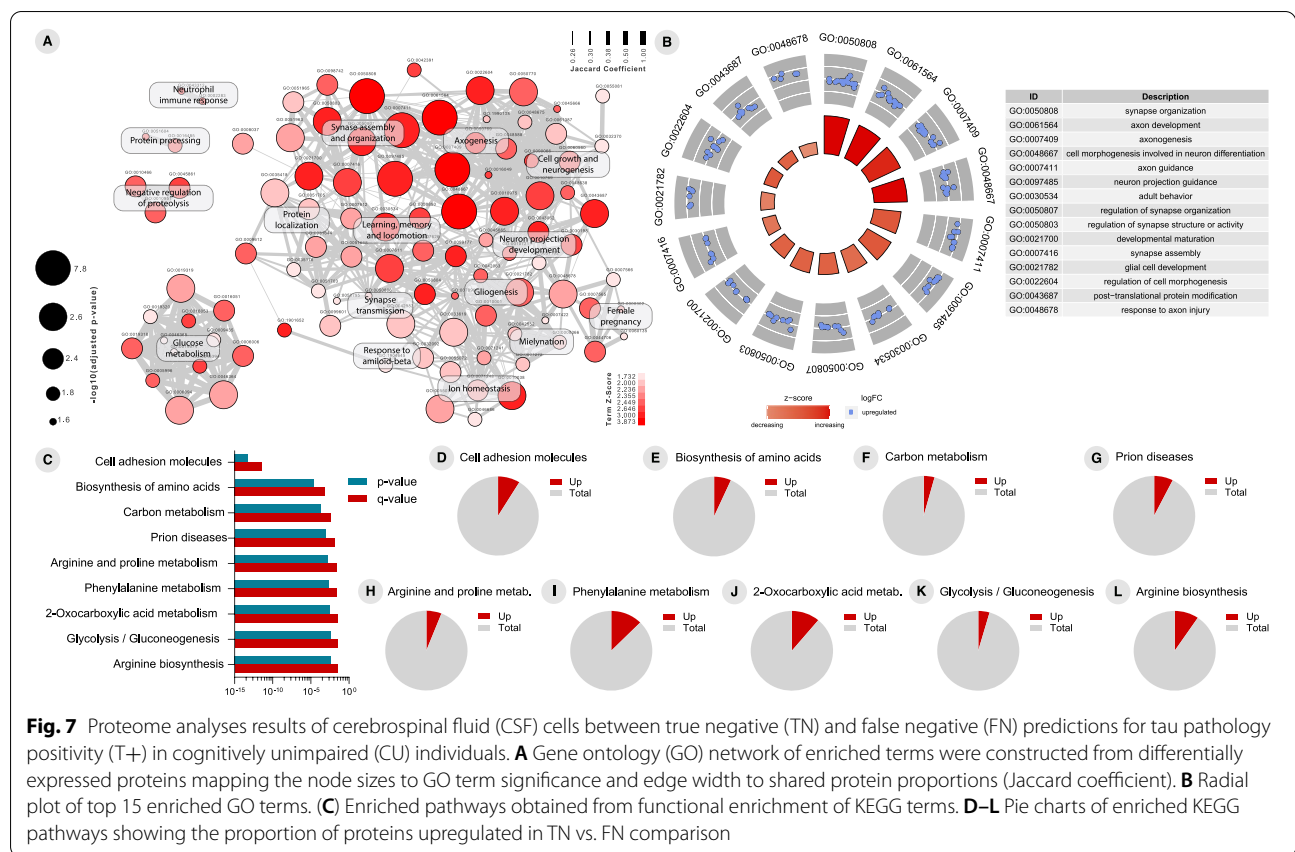
wrongly classified by our ML model in the test set. First, we stratified the ML predictions for T+ in false-positive (n = 17), false-negative (n = 23), true-positive (n = 51), and true-negative (n = 147). Proteomic analyses for N+ prediction model was not carried out, since few wrong predictions were generated, leading to a small sample size.

We identified 17 upregulated DEPs between true-positive and false-positive (Fig. 6a) and 67 upregulated DEPs between true-negative and false-negatives for T+ individuals (Fig. 7a). Interestingly, enrichment analysis of GO biological processes revealed that processes related to myelinization, and glucose metabolism are enriched when comparing false-positive and true-positive predictions for T+ (Fig. 6a, b). When considering the false-negative and true-negative predictions for T+, DEPs related to glucose metabolism, synapse transmission, gliogenesis, and axogenesis appeared among the enriched GO terms (Fig. 7a, b). Finally, to recognize the most affected pathways related to changes in proteomics profile of individuals that were wrongly predicted, we performed an enrichment analysis using canonical pathways described in the KEGG pathway database. This revealed a significant enrichment of DEPs in pathways including “biosynthesis of amino acids”, “glycolysis/gluconeogenesis”, “carbon metabolism”, “cell adhesion molecules”, and “prion disease” (Figs. 6c–g and 7c–l).

Discussion

In the present study, we demonstrated that ML models using combined Aβ soluble isoforms can predict downstream AD pathological processes, T+ and N+, in CU individuals with better results than Aβ isoforms independently. In the generated models, a higher AUC was achieved for predicting N+ when comparing with the T+. Our proteomics analysis identified several biological processes and signaling pathways altered at pre-symptomatic phase of AD. These findings are especially relevant when considering that AD pathological processes initiate around 20–30 years before the occurrence of the first clinical symptoms [18–22]. Finally, we identified DEPs among individuals wrongly classified as T+ by ML that can be further explored to improve prediction performance of the models.

The notion that Aβ triggers tau hyperphosphorylation and neurodegeneration has been corroborated by multiple experimental studies [23–26]. In fact, Höglund and colleagues demonstrated that CU individuals with amyloidosis presented increased levels of p-tau181 and t-tau in the CSF [27]. However, the diagnostic value of Aβ1–42 has been explored in the literature delivering, though, only modest accuracy for AD prediction [28, 29]. Accordingly, here we demonstrated a poor AUC of 0.580 for N+ and 0.529 for T+ prediction modeled using the Aβ1–42 isoform by itself, the most used CSF biomarker in the diagnosis of AD. *Per se*, the poorly explored isoform Aβ1–38 (AUC of 0.847) along with Aβ1–40 (AUC of 0.811)



were the most accurate predictors for both T+ and N+, respectively. In clinical studies, the Aβ₁₋₄₂/Aβ₁₋₃₈ ratio has been capable of significantly discriminating AD from other forms of dementia [30–32] and shown to be negatively correlated with CSF p-tau levels in AD patients [31]. Additionally, a slight increase in Aβ₁₋₃₈ levels was found in a disease-specific manner in the CSF of AD subjects [32, 33]. Nevertheless, a meta-analysis pointed no significant difference in Aβ₁₋₃₈ levels between AD individuals and control group after comparing eight studies [34]. Cullen and colleagues more recently demonstrated that higher CSF Aβ₁₋₃₈ levels are negatively associated with cognitive decline and risk of developing AD [35]. In this context, it is evident that the potential of this isoform to add information in the preclinical stage of the disease remains under-explored.

In this work, we showed that a logistic regression model could predict T+ using multiple input features, with an AUC of 0.929. It has been demonstrated that Aβ dysmetabolism is capable of triggering the conversion from a normal to a toxic state of tau-dependent synaptic dysfunction [23]. As well, a synergistic interaction between Aβ and tau pathology is likely to occur in AD, rather than the sum of their independent effects [36–38]. Bilgel and colleagues showed that a higher

baseline amyloid load in CU individuals was associated with steeper cognitive decline [39]. In parallel, we hereby demonstrated that amyloid isoforms levels can predict N+ in CU individuals with an AUC of 0.936 using a kNN model. The combination of Aβ isoforms, especially those including smaller Aβ isoforms, seems to help to deliver the best results to predict N+. Indeed, limited *in vivo* evidence shows significant correlations between Aβ₁₋₄₂ levels in the CSF and neurodegeneration in CU individuals [27]. On the other hand, the importance of Aβ₁₋₄₂ isoform as a toxic amyloid specie has been extensively demonstrated [23–26]. In the context of isoform production, literature indicates that Aβ₁₋₃₈ is partially formed by cleavage of the Aβ₁₋₄₂ isoform [40]. Also, it seems that no further cleavage of Aβ₁₋₃₈ occurs, resulting in a “more stable” isoform of Aβ, easier to detect [40]. One could argue that a more prominent amyloid dysmetabolism, with higher rates of cleavage of Aβ₁₋₄₂ into Aβ₁₋₃₈, might be a crucial process that seems to drive tau pathology and neurodegeneration. However, the already described [41] faster turnover of Aβ₁₋₄₂ might be accounting for its poor predictive value in our model. Accordingly, our model shows an important role for less explored Aβ isoforms as indicators of emerging tau pathology and neurodegeneration. In addition to CSF, AD blood biomarkers have

been gaining attention in recent years [42]. Due to their scalability, blood biomarkers will generate large datasets highly suited for ML prediction models.

A β isoforms used in combination seems key for predicting T+ and N+, but do not completely explain all the aspects of AD downstream events. Thus, it is believed that simultaneous phenomena, that account for AD heterogeneity, are taking place in the brains of these individuals. In this context, CSF proteomics has been increasingly applied in the attempt to discover novel biomarkers for AD. However, it is mainly focused in comparing CU and AD individuals [43, 44]. Here, we showed A β pathology-dependent changes at protein level occurring in the CSF of CU individuals. Similarly, Whelan and colleagues performed a multiplex proteomics analysis in the CSF of CU A+ and A- patients and found two DEPs significantly altered: Chitinase 3-like protein (YKL-40) and SPARC-related modular calcium binding protein 2 (SMOC2) [45]. The great number of DEPs between CU T+ and T- subjects identified in our study allowed the further determination of biological processes and signaling pathways significantly enriched in these individuals. Additionally, significant differences in DEPs and its associated biological processes and signaling pathways were observed when comparing right and wrong ML predictions for T+. Interestingly, DEPs identified in other studies comparing CU and AD were also found in our analysis of ML wrong predictions for T+ [44]. In specific, YKL-40, SOD1, PKM, and glucose metabolism related proteins are among the DEPs found in both studies. The degree of similarity between studies seems to highlight a robust pattern of change rather than a cohort-specific effect. These results might shed light to key proteins that can be further explored to improve ML performance for predicting T+ and N+.

Conclusions

Our findings indicate that the use of ML models with A β isoforms as input features might help to predict individuals with AD downstream pathology. In addition, CSF proteomics analysis highlighted a promising group of proteins potentially driving tau pathology, which can be further explored for improving future T+ and N+ prediction. Finally, the combination of methodologies used here—ML and proteomics—may help to further understand AD pathology heterogeneity.

Abbreviations

2D-UPLC-MS/MS: 2D-ultra-performance liquid chromatography-tandem mass spectrometry; A+: Amyloid-beta positivity; A-: A β negative; AD: Alzheimer's disease; ADAS-Cog: Alzheimer's Disease Assessment Scale-Cognitive Subscale; ADNI: Alzheimer's Disease Neuroimaging Initiative; AUC: Area under the curve; A β : Amyloid-beta; CSF: Cerebrospinal fluid; CU: Cognitively unimpaired; DEP:

Differentially expressed proteins; GO: Gene ontology; JCTLM: Joint Committee for Traceability in Laboratory Medicine; ML: Machine learning; MMSE: Mini-Mental State Examination; MRI: Magnetic resonance imaging; N+: Neurodegeneration positivity; N-: Neurodegeneration negative; NIA-AA: National Institute on Aging-Alzheimer's Association; p-tau: Phosphorylated tau; PET: Positron emission tomography; RBM: Rules Based Medicine; SD: Standard deviation; SMOC2: SPARC-related modular calcium binding protein 2; T+: Tau pathology positivity; T-: Tau pathology negative; t-tau: Total tau; YKL-40: Chitinase 3-like protein.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13578-021-00712-3>.

Additional file 1. Machine learning results for predicting tau pathology positivity (T+). Table containing features, AUC and standard deviation results for all 1023 models for predicting tau pathology positivity.

Additional file 2. Machine learning results for predicting neurodegeneration positivity (N+). Table containing features, AUC and standard deviation results for all 1023 models for predicting neurodegeneration positivity.

Additional file 3. Differentially expressed proteins (DEPs) in the cerebrospinal fluid (CSF) of cognitively unimpaired (CU) tau pathology positive (T+) compared to negative (T-) subjects. Table containing Protein ID, p-value, adjusted p-value, t-value and logFC for differentially expressed proteins in the cerebrospinal fluid of cognitively unimpaired tau pathology positive compared to negative subjects.

Additional file 4. Differentially expressed proteins (DEPs) in the cerebrospinal fluid (CSF) of cognitively unimpaired (CU) neurodegeneration positive (N+) compared to negative (N-) subjects. Table containing Protein ID, p-value, adjusted p-value, t-value and logFC for differentially expressed proteins in the cerebrospinal fluid of cognitively unimpaired neurodegeneration positive compared to negative subjects.

Acknowledgements

Data used in preparation of this manuscript were obtained from the ADNI database (adni.loni.usc.edu). The list of ADNI investigators can be found online at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Authors' contributions

Conceptualization: GP, BB, WSB, BZ, EZ. Methodology: GP, BB, WSB, MADB. Software: GP. Investigation: GP. Visualization: GP, MADB. Supervision: BB, BZ, RMA, EZ. Writing—original draft: GP, BB, WSB, PCLF, EZ. Writing—review & editing: GP, BB, WSB, MADB, PCLF, TAP, ALB, PRN, DOS, BZ, EZ. All authors read and approved the final manuscript.

Funding

GP receives financial support from CAPES [88882.345577/2019-01]. BB receives financial support from CAPES [88887.336490/2019-00]. PRN receive grants from CIHR [MOP-11-51-31; FRN, 152985], Alzheimer's Association [NIRG-12-92090; NIRP-12-259245] and FRQS [2020-VICO-279314]. ERZ receives grants from CNPq [435642/2018-9; 312410/2018-2], Instituto Serrapilheira [Serra-1912-31365], FAPERGS/MS/CNPq/SESRS-PPSUS [30786.434.24734.23112017], ARD/FAPERGS [54392.632.30451.05032021] and Alzheimer's Association [AARGD-21-850670].

Availability of data and materials

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The dataset supporting the conclusions of this manuscript is available at the ADNI website (<http://adni.loni.usc.edu/>).

Declarations

Ethics approval and consent to participate

ADNI was ethically approved by the institutional review board of all participating sites, subjects provided written informed consent.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Graduate Program in Biological Sciences: Biochemistry, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil. ²Graduate Program in Computing, Universidade Federal de Pelotas (UFPEL), Pelotas, Brazil. ³Department of Neurology and Psychiatry, University of Pittsburgh, Pittsburgh, USA. ⁴Translational Neuroimaging Laboratory, The McGill University Research Centre for Studies in Aging, 6825 LaSalle Boulevard, Verdun, QC H4H 1R3, Canada. ⁵Montreal Neurological Institute, 3801 University Street, H3A 2B4 Montreal, QC, Canada. ⁶Department of Biochemistry, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil. ⁷Douglas Hospital, McGill University, 6875 La Salle Blvd-FBC room 3149, Montreal, QC H4H 1R3, Canada. ⁸Department of Pharmacology, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil. ⁹Graduate Program in Biological Sciences: Pharmacology and Therapeutics, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil.

Received: 30 October 2021 Accepted: 11 November 2021

Published online: 11 December 2021

References

- Collaborators GBDD. Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* 2019;18(11):88–106.
- Serrano-Pozo A, Frosch MP, Masliah E, Hyman BT. Neuropathological alterations in Alzheimer disease. *Cold Spring Harb Perspect Med.* 2011;1(1):a006189.
- Perl DP. Neuropathology of Alzheimer's disease. *Mt Sinai J Med.* 2010;77(1):32–42.
- Selkoe DJ. The molecular pathology of Alzheimer's disease. *Neuron.* 1991;6(4):487–98.
- Hardy JA, Higgins GA. Alzheimer's disease: the amyloid cascade hypothesis. *Science.* 1992;256(5054):184–5.
- Blennow K, Zetterberg H. Biomarkers for Alzheimer's disease: current status and prospects for the future. *J Intern Med.* 2018;284(6):643–63.
- Jack CR, Jr, Bennett DA, Blennow K, Carrillo MC, Feldman HH, Frisone GB, et al. A/T/N: an unbiased descriptive classification scheme for Alzheimer disease biomarkers. *Neurology.* 2016;87(5):539–47.
- Aizenstein HJ, Nebes RD, Saxton JA, Price JC, Mathis CA, Tsopelas ND, et al. Frequent amyloid deposition without significant cognitive impairment among the elderly. *Arch Neurol.* 2008;65(11):1509–17.
- Jack CR, Jr, Lowe VJ, Weigand SD, Wiste HJ, Senjem ML, Knopman DS, et al. Serial PIB and MRI in normal, mild cognitive impairment and Alzheimer's disease: implications for sequence of pathological events in Alzheimer's disease. *Brain.* 2009;132(Pt 5):1355–65.
- Pike KE, Savage G, Villemagne VL, Ng S, Moss SA, Maruff P, et al. Beta-amyloid imaging and memory in non-demented individuals: evidence for preclinical Alzheimer's disease. *Brain.* 2007;130(Pt 11):2837–44.
- Petersen RC, Aisen PS, Beckett LA, Donohue MC, Gamst AC, Harvey DJ, et al. Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. *Neurology.* 2010;74(3):201–9.
- Korecka M, Waligorska T, Figurski M, Toledo JB, Arnold SE, Grossman M, et al. Qualification of a surrogate matrix-based absolute quantification method for amyloid-beta(4)(2) in human cerebrospinal fluid using 2D UPLC-tandem mass spectrometry. *J Alzheimers Dis.* 2014;41(2):441–51.
- Schindler SE, Gray JD, Gordon BA, Xiong C, Batrla-Utermann R, Quan M, et al. Cerebrospinal fluid biomarkers measured by Elecsys assays compared to amyloid imaging. *Alzheimers Dement.* 2018;14(11):1460–9.
- Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015;43(7):e47.
- Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS.* 2012;16(5):284–7.
- Walter W, Sanchez-Cabo F, Ricote M. GPlot: an R package for visually combining expression data with functional analysis. *Bioinformatics.* 2015;31(17):2912–4.
- Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000;28(1):27–30.
- Bateman RJ, Xiong C, Benzinger TL, Fagan AM, Goate A, Fox NC, et al. Clinical and biomarker changes in dominantly inherited Alzheimer's disease. *N Engl J Med.* 2012;367(9):795–804.
- Braak H, Braak E. Diagnostic criteria for neuropathologic assessment of Alzheimer's disease. *Neurobiol Aging.* 1997;18(4 Suppl):S85–8.
- Fagan AM, Xiong C, Jasielec MS, Bateman RJ, Goate AM, Benzinger TL, et al. Longitudinal change in CSF biomarkers in autosomal-dominant Alzheimer's disease. *Sci Transl Med.* 2014;6(226):226ra30.
- Morris JC, Price JL. Pathologic correlates of nondemented aging, mild cognitive impairment, and early-stage Alzheimer's disease. *J Mol Neurosci.* 2001;17(2):101–18.
- Villemagne VL, Burnham S, Bourgeat P, Brown B, Ellis KA, Salvado O, et al. Amyloid beta deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: a prospective cohort study. *Lancet Neurol.* 2013;12(4):357–67.
- Bloom GS. Amyloid-beta and tau: the trigger and bullet in Alzheimer disease pathogenesis. *JAMA Neurol.* 2014;71(4):505–8.
- Nisbet RM, Polanco JC, Ittner LM, Gotz J. Tau aggregation and its interplay with amyloid-beta. *Acta Neuropathol.* 2015;129(2):207–20.
- Jacobs HL, Hedden T, Schultz AP, Sepulcre J, Perea RD, Amariglio RE, et al. Structural tract alterations predict downstream tau accumulation in amyloid-positive older individuals. *Nat Neurosci.* 2018;21(3):424–31.
- Jeong S. Molecular and cellular basis of neurodegeneration in Alzheimer's disease. *Mol Cells.* 2017;40(9):613–20.
- Hoglund K, Kern S, Zettergren A, Borjesson-Hansson A, Zetterberg H, Skoog I, et al. Preclinical amyloid pathology biomarker positivity: effects on tau pathology and neurodegeneration. *Transl Psychiatry.* 2017;7(1):e995.
- Hampel H, Toschi N, Baldacci F, Zetterberg H, Blennow K, Kilimann I, et al. Alzheimer's disease biomarker-guided diagnostic workflow using the added value of six combined cerebrospinal fluid candidates: Aβeta1-42, total-tau, phosphorylated-tau, NFL, neurogranin, and YKL-40. *Alzheimers Dement.* 2018;14(4):492–501.
- Khoonsari PE, Shevchenko G, Herman S, Remnestal J, Giedraitis V, Brundin R, et al. Improved differential diagnosis of Alzheimer's disease by integrating ELISA and mass spectrometry-based cerebrospinal fluid biomarkers. *J Alzheimers Dis.* 2019;67(2):639–51.
- Mulugeta E, Londos E, Ballard C, Alves G, Zetterberg H, Blennow K, et al. CSF amyloid beta38 as a novel diagnostic marker for dementia with Lewy bodies. *J Neurol Neurosurg Psychiatry.* 2011;82(2):160–4.
- Welge V, Fiege O, Lewczuk P, Mollenhauer B, Esselmann H, Klafki HW, et al. Combined CSF tau, p-tau181 and amyloid-beta 38/40/42 for diagnosing Alzheimer's disease. *J Neural Transm (Vienna).* 2009;116(2):203–12.
- Wiltfang J, Esselmann H, Bibl M, Smirnov A, Otto M, Paul S, et al. Highly conserved and disease-specific patterns of carboxyterminally truncated Aβeta peptides 1-37/38/39 in addition to 1-40/42 in Alzheimer's disease and in patients with chronic neuroinflammation. *J Neurochem.* 2002;81(3):481–96.
- Bibl M, Mollenhauer B, Lewczuk P, Esselmann H, Wolf S, Trenkwalder C, et al. Validation of amyloid-beta peptides in CSF diagnosis of neurodegenerative dementias. *Mol Psychiatry.* 2007;12(7):671–80.
- Olsson B, Lautner R, Andreasson U, Ohrfelt A, Portelius E, Bjerke M, et al. CSF and blood biomarkers for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis. *Lancet Neurol.* 2016;15(7):673–84.
- Cullen NC, Janelidze S, Palmqvist S, Stomrud E, Mattsson-Carlgren N, Hansson O. CSF Aβ38 levels are associated with Alzheimer-related decline: implications for γ-secretase modulators. *medRxiv.* 2021:2021.01.31.21250702.

36. Pascoal TA, Mathotaarachchi S, Mohades S, Benedet AL, Chung CO, Shin M, et al. Amyloid-beta and hyperphosphorylated tau synergy drives metabolic decline in preclinical Alzheimer's disease. *Mol Psychiatry*. 2017;22(2):306–11.
37. Pascoal TA, Mathotaarachchi S, Shin M, Benedet AL, Mohades S, Wang S, et al. Synergistic interaction between amyloid and tau predicts the progression to dementia. *Alzheimers Dement*. 2017;13(6):644–53.
38. Busche MA, Hyman BT. Synergy between amyloid-beta and tau in Alzheimer's disease. *Nat Neurosci*. 2020;23(10):1183–93.
39. Bilgel M, An Y, Helpfrey J, Elkins W, Gomez G, Wong DF, et al. Effects of amyloid pathology and neurodegeneration on cognitive change in cognitively normal adults. *Brain*. 2018;141(8):2475–85.
40. Okochi M, Tagami S, Yanagida K, Takami M, Kodama TS, Mori K, et al. gamma-secretase modulators and presenilin 1 mutants act differently on presenilin/gamma-secretase function to cleave Abeta42 and Abeta43. *Cell Rep*. 2013;3(1):42–51.
41. Patterson BW, Elbert DL, Mawuenyega KG, Kasten T, Ovod V, Ma S, et al. Age and amyloid effects on human central nervous system amyloid-beta kinetics. *Anna Neurol*. 2015;78(3):439–53.
42. Henrik Z, Samantha CB. *Molecular Brain*. 2019;12:26.
43. Sathe G, Na CH, Renuse S, Madugundu AK, Albert M, Moghekar A, et al. Quantitative proteomic profiling of cerebrospinal fluid to identify candidate biomarkers for Alzheimer's disease. *Proteomics Clin Appl*. 2019;13(4):e1800105.
44. Bader JM, Geyer PE, Muller JB, Strauss MT, Koch M, Leypoldt F, et al. Proteome profiling in cerebrospinal fluid reveals novel biomarkers of Alzheimer's disease. *Mol Syst Biol*. 2020;16(6):e9356.
45. Whelan CD, Mattsson N, Nagle MW, Vijayaraghavan S, Hyde C, Janelidze S, et al. Multiplex proteomics identifies novel CSF and plasma biomarkers of early Alzheimer's disease. *Acta Neuropathol Commun*. 2019;7(1):169.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

