Multiomics machine learning identifies sleep and inflammation molecular pathways in prodromal Alzheimer's Disease

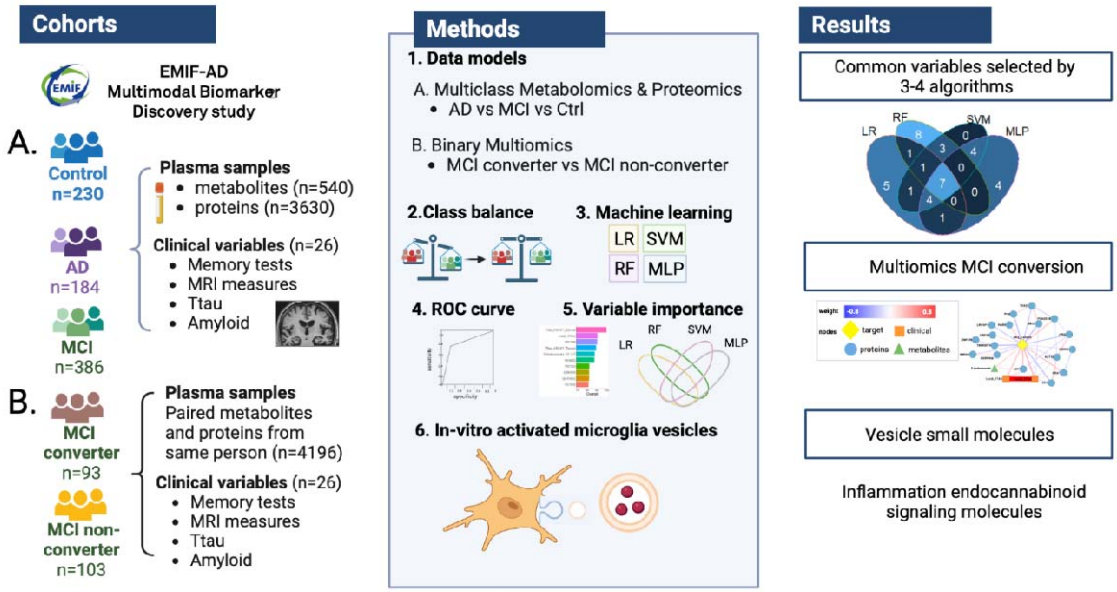
Alicia Gómez-Pascual[1*+], Talel Naccache[2+], Jin Xu[3], Koroush Hooshmand[4], Asger Wretlind[4], Martina Gabrielli[5], Marta Tiffany Lombardo[5,6], Liu Shi[33], Noel J. Buckley[7,8], Betty M. Tijms[9], Stephanie J. B. Vos[10], Mara ten Kate[9], Sebastiaan Engelborghs[11,12], Kristel Sleegers[13,14], Giovanni B. Frisoni[15,16], Anders Wallin[17], Alberto Lleó[18], Julius Pop[19,20], Pablo Martinez-Lage[21], Johannes Streffer[22], Frederik Barkhof[23,24], Henrik Zetterberg[25-27], Pieter Jelle Visser[9,10], Simon Lovestone[7,29], Lars Bertram[30,31], Alejo J. Nevado-Holgado[7], Petroula Proitsi[32], Claudia Verderio[5], Juan A. Botía[1§], Cristina Legido-Quigley[3,4*§]

1. Department of Information and Communications Engineering Faculty of Informatics, University of Murcia, Murcia, Spain.
2. Department of Data Science, City University of London, UK
3. Institute of Pharmaceutical Science, King's College London, London, UK.
4. Steno Diabetes Center Copenhagen, Herlev, Denmark.
5. CNR Institute of Neuroscience, 20854, Vedano al Lambro, Italy
6. School of Medicine and Surgery, University of Milano-Bicocca, 20126, Italy
7. Department of Psychiatry, University of Oxford, UK.
8. Kavli Institute for Nanoscience Discovery
9. Alzheimer Center, VU University Medical Center, Amsterdam, the Netherlands.
10. Department of Psychiatry and Neuropsychology, Alzheimer Centrum Limburg, Maastricht University, Maastricht, the Netherlands.
11. Reference Center for Biological Markers of Dementia (BIODEM), Department of Biomedical Sciences, University of Antwerp, Antwerp, Belgium.
12. Department of Neurology and Bru-BRAIN, UZ Brussel and Center for Neurosciences (C4N), Vrije Universiteit Brussel, Brussels, Belgium.
13. Complex Genetics Group, VIB Center for Molecular Neurology, VIB, Antwerp, Belgium.
14. Institute Born-Bunge, Department of Biomedical Sciences, University of Antwerp, Antwerp, Belgium
15. University of Geneva, Geneva, Switzerland.
16. IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy.
17. Institute of Neuroscience and Physiology, Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden.
18. Neurology Department, Hospital Sant Pau, Barcelona, Spain. Centro de Investigación en Red en enfermedades neurodegenerativas (CIBERNED).
19. University Hospital of Lausanne, Lausanne, Switzerland
20. Geriatric Psychiatry, Department of Mental Health and Psychiatry, Geneva university Hospitals, Geneva, Switzerland.
21. CITA-Alzheimer Foundation, San Sebastian, Spain.
22. AC Immune SA, Lausanne, Switzerland, formerly Janssen R&D, LLC. Beerse, Belgium at the time of study conduct.
23. Department of Radiology and Nuclear Medicine, Amsterdam UMC, Vrije Universiteit, The Netherland
24. Queen Square Institute of Neurology and Centre for Medical Image Computing, University College London, UK.
25. Clinical Neurochemistry Laboratory, Sahlgrenska University Hospital, Mölndal, Sweden.
26. Department of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Mölndal, Sweden.
27. UK Dementia Research Institute at UCL, London, United Kingdom.
28. Department of Neurodegenerative Disease, UCL Institute of Neurology, London, United Kingdom.
29. Janssen Medical (UK), High Wycombe, UK.
30. Lübeck Interdisciplinary Platform for Genome Analytics, University of Lübeck, Lübeck, Germany.
31. Department of Psychology, University of Oslo, Oslo, Norway.
32. Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK.
33. Novo Nordisk Research Centre Oxford (NNRCO), Oxford, UK.

+ Contributed equally
§ Senior authors with equal contribution

*correspondence

Graphical abstract



Abstract

Mild Cognitive Impairment (MCI) is a phase that can precede Alzheimer's Disease (AD). To better understand the molecular mechanisms underlying conversion from MCI to AD, we proposed a multiomics machine learning pipeline (four algorithms) to identify key pathways. Data consisted of metabolites (n=540) and proteins (n=3630) measured in blood plasma coupled with standard clinical tests (n=26). The cohort comprised 230 controls, 386 MCI participants and 184 AD-type dementia participants. Multiclass models showed that oleamide, MMSE and the priority language Z-score were the most relevant variables. Oleamide was increased in the MCI group and further increased in converters (both $P<0.0001$). In-vitro disease-associated microglia were able to synthesize oleamide and excrete it in vesicles. MCI conversion models showed pTau, tTau and JPH3, CFP, synuclein and PI15 proteins as the most relevant. This study uncovered molecular pathways in MCI conversion involved in inflammation (oleamide, CFP), neuronal regulation (JPH3, SNCA) and protein degradation (PI15).

Keywords: Alzheimer's Disease; multiomics, machine learning; mild cognitive impairment; microglia.

Abbreviations: AD = Alzheimer's Disease; MCI = mild cognitive impairment; cMCI = MCI converted to Alzheimer's; sMCI = MCI stable; MMSE = mini mental state examination; NL = normal elderly; ML = machine learning; LR = logistic regression; MLP = multi-layer perceptron network; RF = random forest; SVM = support vector machine.

Introduction

Mild cognitive impairment (MCI) is defined as the symptomatic predementia stage characterized by objective impairment in cognition that does not interfere notably with activities of daily life[1]. It is estimated that over 15% of community dwellers have MCI. The prevalence of MCI increases with age and decreases with education, and it is a heterogeneous and unstable condition[2]. MCI individuals who eventually progress to Alzheimer's Disease (AD) diagnosis are classified as MCI converters (cMCI), while those who remain stable or improve are classified as MCI stable (sMCI). It has been reported that approximately 29% of individuals with prevalent or incident MCI will progress to dementia, while 38% will revert back to a normal cognition diagnosis[3].

The conversion of MCI to dementia has been studied using various data sources, including neuropsychological tests, demographic information, neuroimaging (both structural and functional), genetics, and cerebrospinal fluid (CSF) biomarkers, either alone or in combination[4–10]. With the advancement of high-throughput sequencing technologies, there has been increased interest in "omics" data, which measure different molecules such as the genome, transcriptome, proteome, metabolome, and lipidome[11]. The most common method to quantify proteins, metabolites, and lipids is mass spectrometry, which measures the mass-to-charge ratio of ionized molecules[12]. These measures are taken from biofluids such as plasma and cerebrospinal fluid, as structural and functional changes in the brain can be reflected in these fluids[13].

Many studies investigating MCI progression to AD use traditional methods such as basic statistical tests, correlation and similar metrics[10] as well as models such as Cox regression[5,10] or Bayesian networks[9]. However, the high dimensionality of omics data have directed the interest towards machine learning (ML), a computational approach that delivers further insight into complex data such as omics[14]. Previous research has applied ML algorithms with single omics and clinical data, including lasso regression[4,9], support vector machines (SVM)[4] as well as different deep learning algorithms such as multiple kernel learning (MKL)[4,6] and multimodal recurrent neural network[8]. Among these, our work initiated the use of machine learning-based pipelines for the study of omics related to AD[15–17].

In order to identify early disease pathways involved in the conversion to AD, we analyzed both proteomics and metabolomics data from participants in the European Medical Information Framework for Alzheimer's Disease (EMIF-AD) Multimodal Biomarker Discovery Study[18] using four machine learning algorithms. Our study aims to address two main research questions: 1) which proteins and metabolites are most relevant in differentiating between controls, MCI, and AD? and 2) which proteins and metabolites are most relevant in differentiating between sMCI and cMCI? To specifically examine MCI progression to AD, we created a smaller cohort with paired data and used binary classification models to identify the most relevant proteins and metabolites that differentiate between sMCI and cMCI. Our results indicate the presence of pathways related to inflammation, sleep, neural proteins, and protein degradation. Further exploring the

inflammation connection, we studied microglia in vitro and discovered that they secreted an endocannabinoid called oleamide.

Results

Multiclass models

Multiclass models were created to classify controls, MCI and AD donors using four different machine learning algorithms in two different datasets: proteins with clinical covariates and metabolites with the same covariates. For both approaches, a similar performance was observed between the four different algorithms in the training step, with a mean accuracy of 0.86 for the proteins approach and 0.826 for the metabolites approach. However, when evaluating the models on the test set, SVM stands out from the rest in both approaches. Within the protein models, the mean accuracy of all algorithms on the test set is 0.717, SVM dominates with a 0.882 of accuracy. Analogously, when dealing with the metabolites, the overall accuracy on the test set is 0.717, where SVM precedes with a 0.854 of accuracy (Table 1).

Table 1. Multiclass models performance for proteins and metabolites separately for each algorithm. Results are expressed as accuracy (standard error).

Omics data	Algorithm	Training Accuracy	Test Accuracy
Clinical data and proteins	LR	0.860 (0.034)	0.735 (0.046)
	SVM	0.875 (0.025)	0.882 (0.016)
	RF	0.854 (0.049)	0.729 (0.072)
	MLP	0.857 (0.049)	0.701 (0.048)
Clinical data and metabolomics	LR	0.826 (0.074)	0.717 (0.083)
	SVM	0.843 (0.062)	0.854 (0.014)
	RF	0.872 (0.052)	0.763 (0.075)
	MLP	0.844 (0.063)	0.701 (0.070)

For the protein models, the top 20 most predictive clinical and molecular features were extracted for each algorithm. The overlap of the most relevant proteins selected among the algorithms is represented with a Venn Diagram in Figure 1A. Three clinical variables were selected as relevant by all the algorithms, these are: priority attention z score, priority language z-score and priority memory delayed z score. Amyloid status and MMSE were selected by 3/4 algorithms. Five proteins were selected in two algorithms, these were: Apolipoprotein D (Apo D), Trypsin-1 (PRSS1), Procollagen C-endopeptidase enhancer 1 (PCOLCE), B melanoma antigen 3 (BAGE3) and Semaphorin-6C (SEMA6C).

The same procedure was repeated with metabolite models (Figure 1B). Two clinical variables, MMSE and priority language z-score, were selected by all the algorithms. In addition, one metabolite - oleamide – was selected by all four algorithms. Oleamide concentration in follow up univariate analyses was increased in MCI participants compared to NL (t-test $P<2.2 \cdot 10^{-16}$, t=11.59, df=473.52) and compared to AD (t-test, $P<3.72 \cdot 10^{-8}$, t=5.71, df=217.99). Moreover, oleamide was increased in cMCI compared to sMCI (t-test, $P<8.4 \cdot 10^{-5}$, t=4.01, df=199.01) (see Figure 1D). In additional analysis, oleamide was associated with age in all participants (ANCOVA, $P<2.7 \cdot 10^{-10}$, F=41.28, df=1) and also in MCI converter participants (ANCOVA, $P<7.4 \cdot 10^{-6}$, F=21.07, df=1). Methionine sulfoxide was selected in 3/4 algorithms. AB Zscore, glycerophosphorylcholine (GPC), glycosyl-N-tricosanoyl-sphingadienine-d18:2/23:0, iminodiacetate (IDA), N1-methylinosine, serotonin, 3-hydroxyhippurate and ximenoylcarnitine (C26:1) were selected by two algorithms. The features selected by at least three of the four algorithms in both multiclass models with proteins and multiclass models with metabolites are reported in the Supplementary Table 1.

MCI conversion models

MCI conversion models reported a mean ROC of 0.64 across the different algorithms, with a mean sensitivity of 0.614 and mean specificity of 0.582 (see Supplementary Table 2). All algorithms performed similarly in terms of metrics. The overlap of the top 20 most relevant variables of the four models was evaluated with a Venn Diagram (see Figure 1C). Seven features were selected as relevant in all models. Among them, two clinical variables, local pTau and tTau, tTau Zscore were selected.

In addition, four proteins were selected by all the algorithms, these were: peptidase inhibitor 15 (PI15), Properdin (CFP), Alpha-synuclein (SNCA) and Junctophilin-3 (JPH3). None of them was selected as relevant by any algorithm in the multiclass models. Follow up univariate analyses showed that CFP was increased in cMCI compared to sMCI (t.test $P<6.1 \cdot 10^{-4}$, t=3.49, df=173.22) while JPH3 (t.test $P<9.8 \cdot 10^{-4}$, t=-3.35, df=190.46), PI15 (t.test $P<3.3 \cdot 10^{-4}$, t=-3.65, df=191.73) and SNCA (t.test $P<3.7 \cdot 10^{-4}$, t=-3.63, df=182.67) were decreased for cMCI compared to sMCI (see Figure 1E). In additional analysis, SNCA was the only protein also associated with age (ANCOVA, $P<0.0304$, F=4.756, df=1). Proteins selected by three of the four algorithms were Pancreatic hormone (PPY), Phospholipase A2 (PLA2G1B), Carnitine O-acetyltransferase (CRAT), and Tumor necrosis factor receptor superfamily member 19 (TNFRSF19). None were selected as relevant by multiclass models. Finally, for two algorithms, 5-dodecenoic acid (12:1(n-7)) metabolite was selected as relevant. The features selected by three of the four algorithms for the MCI conversion models are reported in the Supplementary Table 1. Features in this table from multiclass models and MCI conversion models were represented in a correlation network where MCI conversion is the target. Proteins or metabolites were arranged around it and linked to each other (see Figure 1F).

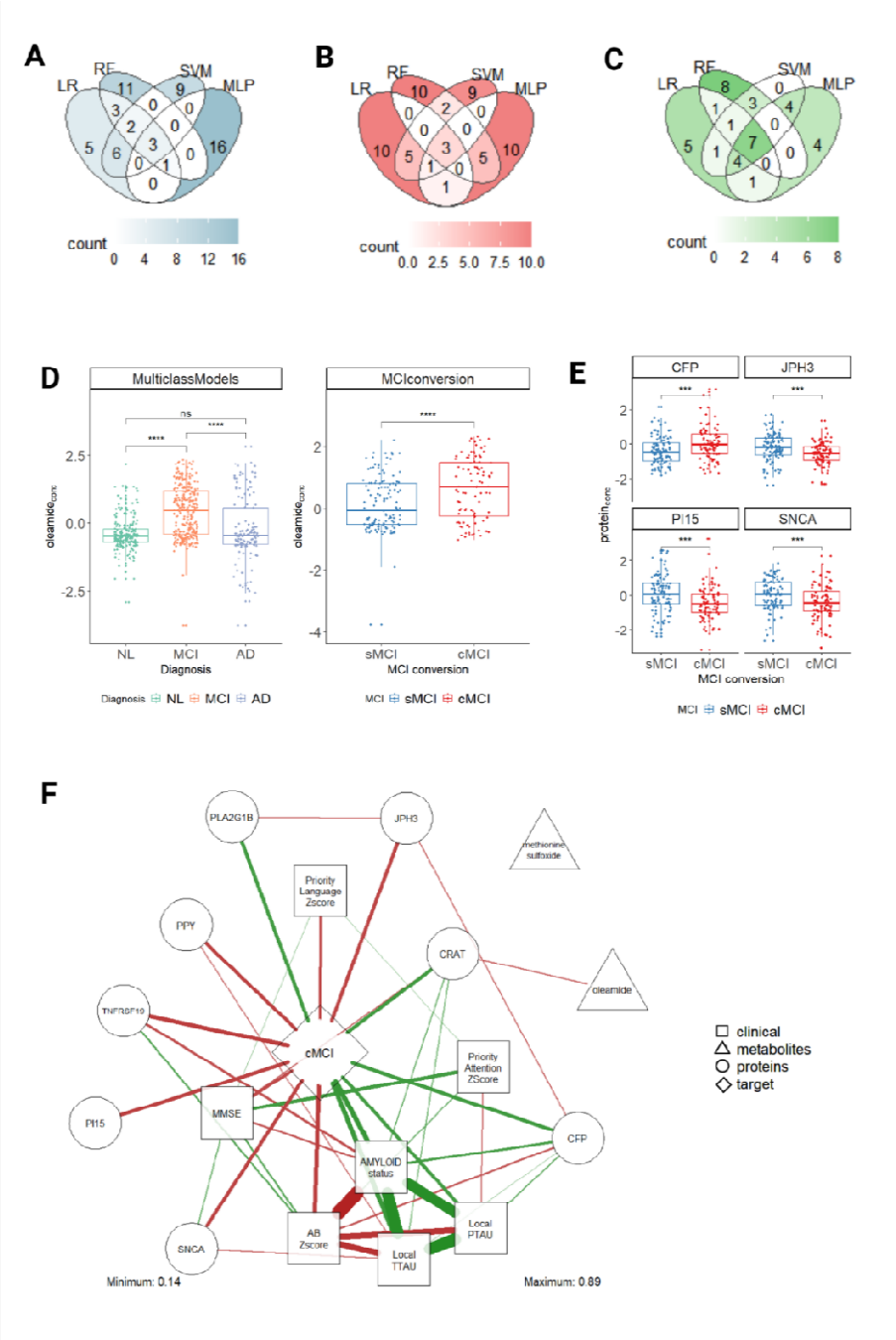


Figure 1. Most relevant clinical features, proteins and metabolites extracted with multiclass models and MCI conversion models.

Venn diagram shows the overlap of the top 20 most relevant predictors of each algorithm for **(A)** multiclass models of proteins, **(B)** multiclass models of metabolites and **(C)** MCI conversion models of proteins and metabolites; **(D)** Oleamide level distribution for NL (n=230), MCI (n=386) and AD (n=184) donors; **(E)** Oleamide concentration distribution for sMCI (n=291) and cMCI (n=100) donors; **(F)** Correlation network including the variables selected as relevant by at least three of the four algorithms proposed in all the approaches: multiclass models for proteins, multiclass models for metabolites and MCI conversion

models. Correlations were estimated with the sMCI (n=103) and cMCI (n=91) donors with paired data (proteomics and metabolomics). Only significant correlations are shown. The minimum and maximum strength of the correlation is shown in the figure. Positive correlations are represented with green color, negative correlations with red. Features are grouped into target, clinical, proteins or metabolites. ns, non-significant; *, P<0.05; **, P<0.01; ***, P<0.001; ****, P<0.0001.

Oleamide in microglia

The LC-MS quantitation for oleamide was carried out on microglia cultures and their secreted extracellular vesicles (EVs) (Supplementary Table 3, see supplementary methods). The concentration of oleamide was normalized by the total protein in cells and EVs. The amount of oleamide was higher in microglia compared to the supernatant, and further concentrated in EVs. Moreover, it can be observed that activated microglia contained higher amounts of oleamide compared to unstimulated cells (see Figure 2.A).

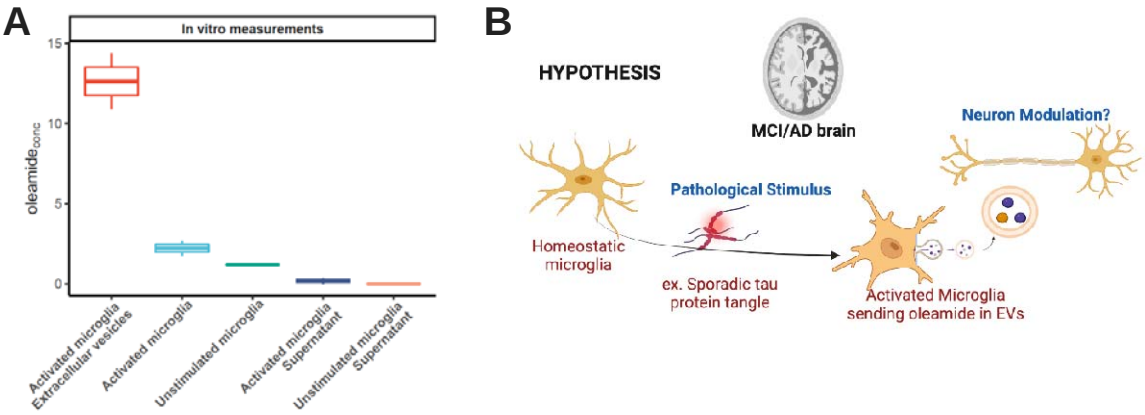


Figure 2. Oleamide in microglia. (A) Oleamide concentration distribution per microglia state and EVs. Concentration is normalized by protein amount in cells and EVs (n=2). **(B)** Microglia phenotypes and oleamide hypothesis. The present study provides the first evidence of oleamide being present in increased amount in plasma from MCI participants and enriched in microglia extracellular vesicles released in the pericellular space. A hypothesis is presented in which microglia activated by fibrils in MCI/AD brain release oleamide in EVs to possibly modulate or inhibit neuronal transmission. It is known that EV production increases in MCI and AD brain.[19,20] This mechanism, possibly via CB1 receptors would reduce the activity of neurons in the brain, similarly to sleep-inducing.

Discussion

The use of artificial intelligence in uncovering disease-specific pathways has gained great interest in recent years. Machine learning models, by their nature, have the ability to analyze vast amounts of data and detect complex molecular interactions, such as non-linear relationships. In this study, we used the EMIF-AD dataset and applied machine learning algorithms to identify key molecules involved in the conversion of MCI to AD. Four different algorithms (logistic regression, support vector machines, random forest, and multi-layer perceptron) were utilized to analyze two different groups: i) controls vs MCI vs AD, and ii) cMCI vs sMCI. Our discussion focuses on the molecules that were consistently identified as relevant by all four algorithms and thus assumed to be robustly linked to the disease status.

Model performance

The four algorithms showed a similar performance in the training of multiclass models for both proteins (0.86 mean accuracy) and metabolites (0.826 mean accuracy). All the algorithms showed a lower accuracy of prediction for the MCI class, possibly due to the heterogeneity covered by the MCI diagnosis (Supplementary Figure 1). SVM outperformed all other algorithms in the test data, reaching values of accuracy of 0.882 for proteins and 0.854 for metabolites. Despite SVM outperforming all other algorithms, we focused on molecules repeatedly selected by all algorithms for pathway discovery pipeline.

Performance was in line with previous work. The smaller plasma metabolite dataset by Kim et al.[21] was used by Stamate et al.[16] to demonstrate that plasma metabolites have the potential to match well-established AD CSF biomarkers. The study compared Deep Learning (DL), Extreme Gradient Boosting (XGBoost) and Random Forest (RF) models and had 357 participants (115 AD and 242 control). The DL model produced an AUC value of 0.85, XGBoost model 0.88 whereas the RF model resulted in a 0.85 AUC value (all values based on a 95% confidence interval). Additionally, the same proteomics data alone was investigated by Shi et al.[22], where ML algorithms were used to classify the donors into amyloid positive and amyloid negative participants. In this study, Lasso Regression and SVM models parsed a predictive panel composed of 44 proteins, age and the risk gene APOE4 achieving an AUC of 0.68 in the replication group.

Clinical features

Clinical covariates were included in all the models as they help with the pathway interpretability. For example, in the multiclass models, the MMSE test and the priority language score were ranked high for both proteins and metabolites models while these variables did not play an important role in MCI conversion. Instead, Tau-related measures (pTau and tTau) were shown by the binary model to be relevant. This is interesting because Tau or amyloid measures are deemed to be a more accurate and objective measure for diagnosis. The other important features among the clinical metadata were known risk factors of AD, such as gender, age, years of education and the volume of the hippocampus which is known to be related to the memory function in the brain. However, all these were not chosen consistently by the models (see Supplementary Table 1).

Multiclass models most relevant proteins and metabolites

Multiclass models (NL vs. MCI (all) vs. AD) for proteins showed low repeatability but the metabolite models showed that one lipid, oleamide, was selected as a predictor by all algorithms while methionine sulphate was chosen by three of them.

Oleamide is a molecule thought to be synthesized in the brain to aid with sleep[23] and is a potent endogenous endocannabinoid[24]. The increased levels in the MCI group suggested that sleep deprivation could be involved in memory and cognitive impairment. Poor sleep elevates oleamide and memory problems can arise as a symptom of poor sleep. A link with AD conversion is less clear but in one of our previous studies on a smaller dataset, we found that oleamide was associated with elevated amyloid levels in the brain[21]. In sensitivity analyses, oleamide was significantly higher in cMCI (see Figure 1D) but also associated with age. Sleep has been shown by Xie et al.[25] to be a clearance mechanism in the brain and perhaps oleamide could be implicated in sleep but also amyloid or tau clearance.

Furthermore, methionine is a sulfur-containing essential amino acid. Among many different functions, it intervenes in the biosynthesis of glutathione to counteract oxidative stress and inflammation[26].

Microglia and extracellular vesicle (EV) analyses

The discovery of oleamide and its possible role in sleep regulation and amyloid/tau clearance prompted the investigation of microglial involvement in this process. Microglia are known to release a variety of signalling molecules that impact synaptic transmission in response to injury or inflammation and play a crucial role in maintaining balance in neuronal networks[27]. Additionally, microglia release EVs into their surroundings, which carry molecules and distribute them throughout the nervous system[28,29]. EV release from microglia increases in MCI and AD patients[19,20] and microglia are also involved in sleep regulation[30–32]. Endocannabinoids have been linked to learning, memory and long-term plasticity.[33,34] Furthermore, endocannabinoids were shown to be involved in the formation of synapses and the growth of new neurons[35]. Our previous work showed that anandamide was transported in microglia EVs, and that anandamide could inhibit presynaptic transmission in target GABAergic neurons.[36] The present study provides the first evidence of oleamide being present in microglia and enriched in the EVs they release in the pericellular space. Figure 2B presents a hypothesis where microglia activated by fibrils in MCI/AD brains release oleamide in EVs to possibly modulate or inhibit neuronal transmission.

MCI conversion models most relevant proteins and metabolites

The results of the machine learning models used in this study to predict MCI conversion revealed four key proteins: peptidase inhibitor 15, properdin, alpha-synuclein, and junctophilin-3. These proteins were selected by all four algorithms used.

Decreased levels of alpha-synuclein and junctophilin (JPH3) were found in the blood of cMCI participants, which may suggest that some individuals had mixed Parkinson's Disease pathology and motor symptoms. Notably, the presence of dual motor and memory symptoms has been shown to increase the risk of developing dementia.[42] The JPH3 protein is

particularly noteworthy because it is a neuron specific protein. JPH3 is regulated in a unique neuron-restricted fashion to control the electrical excitability of neurons in different brain regions and is involved in the regulation of intracellular calcium signalling. JPH3 has also been linked to Huntington-like disease-2[37]. Properdin was the only protein that was elevated in cMCI compared to sMCI. It belongs to the complement system, a well-established pathway of inflammation. This supports the hypothesis that inflammation in the brain produced by fibrils exacerbates the progression of AD. The last protein selected by all algorithms was peptidase inhibitor 15 (PI15), which is an inhibitor against trypsin and, although unexplored in brain, may play a role in protein degradation in the central nervous system.

The network analysis was performed to understand any additional relationships between the selected molecules. The results visualised that the converters had relatively higher levels of brain biomarkers such as tau and amyloid which was also reported in the demographics supplementary table 4. Additionally, the average MMSE score was lower for cMCI individuals (mean of 25) compared to sMCI individuals (mean of 26). Two new proteins, PLA2G1B and CRAT, showed weak positive correlation with conversion. PLA2 is a lipase enzyme involved in the hydrolysis of phospholipids that has been previously linked to AD using a lipidomics approach[38]. CRAT on the other hand is a carnitine acetyltransferase that has been associated with the severe neurometabolic disorder Leigh syndrome[39]

Limitations

Machine learning applications to omics data have some limitations, such as collinearity, which can increase bias, and small dataset size. The cMCI and sMCI groups in this study were not matched in biomarkers, as tau and amyloid were elevated in cMCI participants compared to sMCI. Nonetheless, the study has the advantage of using comprehensive clinical data, including CSF biomarkers and imaging data, which helps in pathway discovery while maintaining clinical explainability. The authors were unable to find any existing cohort with both omics data and follow-up outcomes for MCI individuals, which would have further improved the validity of this study. Oleamide measurements have been carried out on EVs and microglia in primary culture, an artificial condition far from the *in vivo* setting. In future studies, isolation of microglia-derived EVs from MCI/AD brain tissue and/or body fluids may help to overcome this limitation.

Conclusion

In conclusion, this study revealed new and potentially significant molecular markers for MCI conversion, including the lipid oleamide which is linked to sleep and memory and can be secreted by microglia via EVs, JPH3 which is a protein found exclusively in the brain with potential memory function, and PI15 which is a peptidase inhibitor of unknown brain function. The pipeline also confirmed the significance of established risk factors such as synuclein and the complement cascade.

Methods

The EMIF-AD Multimodal Biomarker Discovery Study

This study employed data from the European Medical Information Framework for Alzheimer's Disease (EMIF-AD) Multimodal Biomarker Discovery Study[18]. EMIF-AD Multimodal Biomarker Discovery is a cross-cohort study consisting of collated data from 11 European cohorts that aims to discover novel diagnostic and prognostic markers for AD-type dementia by performing analyses in multiple biomarker modalities. In the present study, we used data from 230 normal elderly people (NL), 184 AD participants and 386 participants with MCI. Of 386 MCI participants, 100 were later diagnosed with AD-type dementia (defined as AD converting MCI [cMCI]), 219 remained as MCI (defined as stable MCI [sMCI]) and 67 participants do not have this information available. The average follow-up length was 2.49 years.

From all cohorts, available data on demographics, clinical information, neuropsychological testing, cognition and Aβ status data were gathered. The cognitive tests used varied across centers and only the Mini Mental State Examination (MMSE)[40] was administered everywhere and was available for nearly all subjects. At least one test from the following cognitive domains was performed: memory, language, attention, executive functioning and visuoconstruction. For each cognitive domain, a primary test was selected.

For metabolomics, the relative levels of 540 plasma metabolites were measured in fasting blood samples using three different mass spectrometry methods. Details on the analytical method and data treatment can be found in Min Kim et al.[21]. Effectively, area counts for each metabolite in each sample were extracted from the raw data. The extracted area counts were then normalized to correct for variation resulting from instrument inter-day tuning differences. Metabolite levels below limit of quantification were replaced with 1 while metabolites with more than 20% missing were excluded from the further analysis. Subsequently, the metabolomics data were log transformed to allow the data to be normally distributed and then each metabolite was scaled to have a mean value of 0 and a standard deviation value of 1.

For proteomics analysis, plasma protein levels were assessed in plasma using the SOMAscan assay platform (SomaLogic Inc.). This aptamer-based assay enables the simultaneous measurement of up to 3630 proteins. Samples were grouped and measured separately. To ensure data consistency across assay runs, 40 subjects were tested in both batches. A detailed description of this process has been previously published.[15,41]

Machine learning approaches

Our aim is to discover the most relevant clinical characteristics, proteins and metabolites involved in two different tasks: (1) classifying samples into controls, MCI and AD donors and (2) distinguishing between converters and non-convertors from MCI to AD. We propose

a general pipeline based on a competitive setting in which four different ML algorithms do their best to create these two models and then variables that are the most relevant predictors in classifying samples correctly are extracted from the constructed models. We demonstrate its generality by applying the same pipeline in Python and R setups for topics (1) and (2) respectively.

Pipeline to create multiclass models of NL, MCI and AD donors

To identify the most relevant proteins and metabolites implicated in the identification of NL, MCI and AD donors, multiclass models were created. Proteins and metabolites were treated separately in their corresponding pipelines, but the same clinical covariates were included in both. Multiclass models based on proteins were created using 3630 proteins and 26 clinical covariates from a total of 230 controls, 184 AD donors and 386 MCI donors. Multiclass models for metabolites were created using 540 metabolites with identical clinical covariates from a total of 207 controls, 136 AD donors and 276 MCI donors. Clinical covariates are included in all the models. Statistics of the clinical characteristics of this dataset has been previously published by Shi et al.[22]

The same pipeline was applied to create both multiclass models for proteins with covariates and the multiclass models for metabolites with covariates. Classes were first balanced using the SMOTE approach[42] with the 'imbalanced-learn' python package. A 5-fold cross-validation repeated three times was applied with the 'Repeated Stratified K-Fold' function when fitting the models to address overfitting due to the shape of the data. Four different machine learning algorithms were applied with the 'Scikit-learn' python package[43]. These are well known representatives of four different classifier building approaches, i.e., logistic regression (LogisticRegression), random forest (RandomForestClassifier), support vector machines (SVM) and artificial neural network (MLPClassifier). Each algorithm's hyperparameters were optimized using this same package. The performance of each algorithm was evaluated using ROC curves (Scikit-learn and matplotlib packages)[44].

The most relevant variables for each model were extracted using feature importance methods from the Scikit-learn package. For logistic regression models, feature relevance was represented by the magnitude of the coefficient of the model. For random forest models, feature importance was obtained with "feature_importances_" function, calculated as the mean and standard deviation of accumulation of the impurity decrease within each tree. As SVM and MLP models do not incorporate specific approaches for calculating feature importance, we used a general approach based on predictor permutation through the 'sklearn.inspection.permutation_importance' function, where the values in each feature column were shuffled, the effect on the model prediction accuracy was observed and these steps were repeated for all features. Finally, the features selected as relevant (top 20) by at least three of the four algorithms were selected for further analysis.

Pipeline to create MCI conversion models

To get a more specific overview of the biochemistry behind MCI conversion to AD, binary models that classify MCI donors into cMCI or sMCI were created using the same ML-based approach. For this purpose, only MCI donors with MCI conversion information and both proteins and metabolites information were selected, taking a total of 103 sMCI and 93 cMCI. The descriptive statistics of clinical characteristics for this dataset are reported in Supplementary Table 4. Proteins (n=3630), metabolites (n=540) and clinical variables (n=26) were treated together in the same model.

The pipeline to create the models are based on the functions from the 'caret' R package[45]. First, classes were balanced using a downsampling approach (downSample function). Since the number of samples is much smaller than the original set, the complete dataset was used for training. A 3-folds cross-validation repeated 10 times ('trainControl' function) was applied to avoid overfitting. In addition, the same four machine learning algorithms were applied, including logistic regression ('glmnet' method), random forest ('rf' method), support vector machines ('svmLinear' method) and artificial neural networks ('mlpWeightDecay' method). Hyperparameter tuning was carried out to define our own grid, included in the 'tuneGrid' parameter. The top 20 most relevant features of each algorithm were extracted using the 'varImp' function. Among them, the features selected as relevant (top 20) in at least three of the four algorithms were selected.

Correlation network

For the multiclass models and MCI conversion models from each approach, the features selected as relevant by at least three of the four algorithms were kept since we considered them to be the most reliable. All these features were put together using a network-based approach, where MCI conversion is the target, and the rest of the features (clinical features, proteins or metabolites) can be linked to the target and between themselves. To this end, a correlation network was created. First, correlations were estimated using the 'cor_auto' function from the 'qgraph' R package [46]. Then, the correlation network was represented using the 'qgraph' function from this same package (graph="cor").

Microglia in-vitro experiments

Microglial culture preparation and EV isolation

Pure murine primary microglia, established as in Gabrielli et al[47], have been stimulated or not with 1:20 Granulocyte-Macrophage Colony-Stimulating Factor (GM-CSF) from murine GM-CSF-transfected X63 cells[48]. GM-CSF is member of the colony-stimulating factor superfamily that induces microglial proliferation, migration and upregulation of surface markers[49]. Supernatants have been cleared from cell debris before storing. Cells have been scraped in physiological solution, pelleted and stored in methanol. Extracellular vesicles

(EVs) have been isolated through differential centrifugation from the cell supernatant upon 30 minutes ATP stimulation at 110,000xg[50] and stored at -80C°.

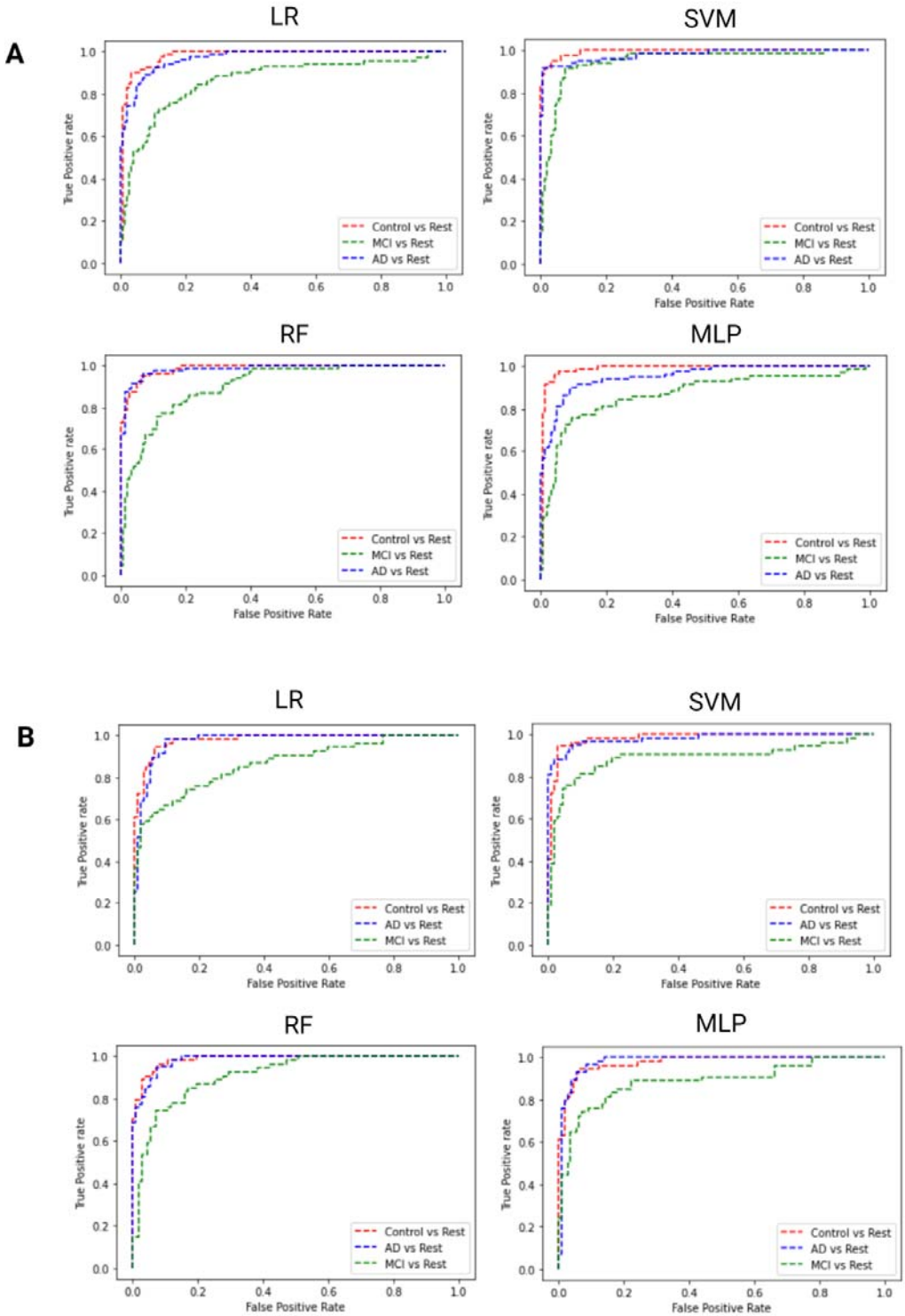
Sample preparation

Oleamide was extracted by the addition of 1 ml of ethanol to the dried microglia and extracellular vesicles. The tubes were then vortexed, for 5s, shaken at 1500 RPM for 5 min at 4 °C, vortexed, for 5s, and subsequently centrifuged at 12,000 g for 5 min at 4°C. Next, the supernatants were evaporated to dryness using a speedvac cold trap concentrator. Each dried sample was reconstituted in a 50 µl methanol: toluene (9:1, v/v) mixture, vortexed for 10s, transferred to glass vials with micro-inserts, capped immediately, and injected into the Ultra High-Performance Liquid Chromatography (UHPLC)-MS/MS system. For details on LC-MS/MS oleamide quantitation with the pure deuterated standard see supplementary methods.

Supplementary material

Supplementary methods. Analysis of oleamide by UHPLC-MS/MS

Samples were analyzed in dynamic multiple reaction mode (dMRM) on an Agilent 1290 Infinity UHPLC system connected to an Agilent 6460 triple quadrupole (QqQ) mass spectrometer (MS) from Agilent Technologies Inc. (Santa Clara, CA, USA). Waters HSS T3 2.1 × 100 mm, 1.8 µm column protected by a C18 HSS T3 VanGuard Pre-column (100Å, 1.8 µm, 2.1 mm × 5 mm) both from (Waters, Taastrup, DK.). The column temperature was maintained at 45°C throughout the run with a flow rate of 0.4 mL/min. The injection volume was 5 µL, and a binary solvent mixture was used. Solvent A contained water and solvent B acetonitrile/isopropanol (67/33) (v/v). 0.1% formic acid (v/v) was added to both solvents. The following gradient was used for positive mode analysis: 0-1 min, 60% A, 1-2 min, ramping to 20% A, 2-8 min, 0% A, 8-9 min, 0% A, 9-9.2 min, ramping back to 60% A, 9.2-12 min, 60% A. Instrument-dependent parameters for mass spectrometry were as follows: The nitrogen drying gas flow and the temperature was 12 L/min and 325 °C, respectively. The capillary voltage was 3500. The nebulizer pressure was controlled at 45 psi. The nitrogen sheath gas flow and temperature were kept at 11.0 L/min and 325 °C, respectively. Supplementary Table 2 summarizes dMRM transitions, retention times, fragmentator voltages, and collision energies used for oleamide.



Supplementary figure 1. ROC curves for multiclass models using four different algorithms with **(A)** proteins and clinical features, **(B)** metabolites and clinical features.

Supplementary Table 1. Clinical characteristics, proteins and metabolites selected as relevant by at least three of the four algorithms in one of the following approaches: the multiclass models with proteins, the multiclass models with metabolites and the MCI conversion models.

	Multiclass models proteins and clinical variables				**Multiclass models metabolites and clinical variables**				**MCI conversion models**			
Characteristics	LR	SVM	RF	MLP	LR	SVM	RF	MLP	LR	SVM	RF	MLP
MMSE	1	1	1	0	1	1	1	1	0	0	0	0
Priority Language Zscore	1	1	1	1	1	1	1	1	0	0	0	0
Priority Memory Delayed Zscore	1	1	1	1	0	0	1	1	0	0	0	0
Priority Attention Zscore	1	1	1	1	0	0	1	1	0	0	0	0
Priority Memory Immediate Zscore	1	1	1	0	0	0	1	0	0	0	1	0
Ttau ASSAY Zscore	1	1	0	0	0	0	1	1	1	1	1	1
AMYLOID status	1	0	1	1	0	0	1	1	1	0	0	0
Local PTAU	0	0	0	0	0	0	1	0	1	1	1	1
Local TTAU	0	0	0	0	0	0	1	0	1	1	1	1
PI15	0	0	0	0	0	0	0	0	1	1	1	1
JPH3	0	0	0	0	0	0	0	0	1	1	1	1
CFP	0	0	0	0	0	0	0	0	1	1	1	1
CRAT	0	0	0	0	0	0	0	0	1	1	1	0
SNCA	0	0	0	0	0	0	0	0	1	1	1	1
oleamide	0	0	0	0	1	1	1	1	0	0	0	0
methionine sulfoxide	0	0	0	0	1	1	0	1	0	0	0	0

Supplementary Table 2. Performance of the MCI conversion models for proteins and metabolites for each algorithm. Standard errors are indicated in the brackets.

Omics data	Algorithm	ROC Training data	Sensitivity Training data	Specificity Training data
Clinical data, proteins and metabolites	LR	0.640 (0.048)	0.614 (0.091)	0.582 (0.089)
	SVM	0.630 (0.043)	0.620 (0.090)	0.575 (0.101)
	RF	0.662 (0.067)	0.621 (0.112)	0.588 (0.092)
	MLP	0.641 (0.058)	0.633 (0.103)	0.549 (0.091)

Supplementary Table 3. MS/MS parameters for the analysis of primary fatty acid amides by dMRM in positive mode. Q1, quadrupole 1 m/z ratio; Q3, quadrupole 3 m/z ratio; R_t, retention time (min); CE, collision energy.

Analyte	Q1 (m/z)	Q3 (m/z)	R_t (min)	Frag (v)	CE (v)
Oleamide	282.2	69.2	4.6	92	21
		55.2		92	45

Supplementary Table 4. Demographics of participants included in the MCI conversion approach by diagnosis. One-way analysis of variance (ANOVA) and chi-square tests were used to compare continuous and binary variables, respectively. Aβ status was defined by the CSF Aβ42/40 of the central analyses, using a cutoff value of < 0.061 to determine abnormality. Standard errors are indicated in the brackets.

Characteristics	sMCI	cMCI	*P* value
n	103	91	NA
Age mean (SD) y	69.17 (8.61)	70.76 (7.49)	0.176
Male sex N (%)	52 (50.5)	50 (54.9)	0.634
Education mean (SD) y	11.35 (3.24)	11.25 (3.45)	0.832
MMSE mean (SD)	26.15 (2.69)	24.98 (2.82)	0.003
Phosphorylated TAU mean (SD)	66.12 (33.58)	83.91 (37.00)	0.001
Amyloid positive status (%)	71 (68.9)	81 (89.0)	0.001
Priority Attention Z-score mean (SD)	-0.64 (1.40)	-0.96 (1.52)	0.121
Priority Language Z-score mean (SD)	-0.76 (1.27)	-1.30 (1.37)	0.004
Priority Memory Immediate Z-score mean (SD)	-1.26 (1.32)	-1.78 (1.18)	0.004
APOE E4+ N (%)	55 (53.4)	58 (63.7)	0.190

References

1. Gauthier, S. *et al*. Mild cognitive impairment. *The Lancet* **367**, 1262–1270 (2006).

2. Bai, W. *et al*. Worldwide prevalence of mild cognitive impairment among community dwellers aged 50 years and older: a meta-analysis and systematic review of epidemiology studies. *Age Ageing* **51**, afac173 (2022).

3. Roberts, R. O. *et al*. Higher risk of progression to dementia in mild cognitive impairment cases who revert to normal. *Neurology* **82**, 317–325 (2014).

4. Varatharajah, Y., Ramanan, V. K., Iyer, R., Vemuri, P., & Alzheimer's Disease Neuroimaging Initiative. Predicting Short-term MCI-to-AD Progression Using Imaging, CSF, Genetic Factors, Cognitive Resilience, and Demographics. *Sci. Rep.* **9**, 2235 (2019).

5. Heister, D. *et al*. Predicting MCI outcome with clinically available MRI and CSF biomarkers. *Neurology* **77**, 1619–1628 (2011).

6. Hinrichs, C., Singh, V., Xu, G. & Johnson, S. C. Predictive markers for AD in a multi-modality framework: An analysis of MCI progression in the ADNI population. *NeuroImage* **55**, 574–589 (2011).

7. Huang, K. *et al.* A multipredictor model to predict the conversion of mild cognitive impairment to Alzheimer's disease by using a predictive nomogram. *Neuropsychopharmacology* **45**, 358–366 (2020).

8. Lee, G., Nho, K., Kang, B., Sohn, K.-A. & Kim, D. Predicting Alzheimer's disease progression using multi-modal deep learning approach. *Sci. Rep.* **9**, 1952 (2019).

9. Liu, H., Zhou, X., Jiang, H., He, H. & Liu, X. A semi-mechanism approach based on MRI and proteomics for prediction of conversion from mild cognitive impairment to Alzheimer's disease. *Sci. Rep.* **6**, 26712 (2016).

10. Smith, E. E. *et al*. Magnetic Resonance Imaging White Matter Hyperintensities and Brain Volume in the Prediction of Mild Cognitive Impairment and Dementia. *Arch. Neurol.* **65**, 94–100 (2008).

11. Manzoni, C. *et al*. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Brief. Bioinform.* **19**, 286–302 (2018).

12. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).

13. Janelidze, S. *et al*. Plasma P-tau181 in Alzheimer's disease: relationship to other biomarkers, differential diagnosis, neuropathology and longitudinal progression to Alzheimer's dementia. *Nat. Med.* **26**, 379–386 (2020).

14. Lin, E. & Lane, H.-Y. Machine learning and systems genomics approaches for multi-omics data. *Biomark. Res.* **5**, 2 (2017).

15. Shi, L. *et al.* Discovery and validation of plasma proteomic biomarkers relating to brain amyloid burden by SOMAscan assay. *Alzheimers Dement.* **15**, 1478–1488 (2019).

16. Stamate, D. *et al.* A metabolite-based machine learning approach to diagnose Alzheimer-type dementia in blood: Results from the European Medical Information Framework for Alzheimer disease biomarker discovery cohort. *Alzheimers Dement. N. Y. N* **5**, 933–938 (2019).

17. Proitsi, P. *et al.* Association of blood lipids with Alzheimer's disease: A comprehensive lipidomics analysis. *Alzheimers Dement. J. Alzheimers Assoc.* **13**, 140–151 (2017).

18. Bos, I. *et al.* The EMIF-AD Multimodal Biomarker Discovery study: design, methods and cohort characteristics. *Alzheimers Res. Ther.* **10**, 64 (2018).

19. Joshi, P. *et al.* Microglia convert aggregated amyloid-β into neurotoxic forms through the shedding of microvesicles. *Cell Death Differ.* **21**, 582–593 (2014).

20. Agosta, F. *et al.* Myeloid microvesicles in cerebrospinal fluid are associated with myelin damage and neuronal loss in mild cognitive impairment and Alzheimer disease. *Ann. Neurol.* **76**, 813–825 (2014).

21. Kim, M. *et al.* Primary fatty amides in plasma associated with brain amyloid burden, hippocampal volume, and memory in the European Medical Information Framework for Alzheimer's Disease biomarker discovery cohort. *Alzheimers Dement. J. Alzheimers Assoc.* **15**, 817–827 (2019).

22. Shi, L. *et al.* Multiomics profiling of human plasma and CSF reveals ATN derived networks and highlights causal links in Alzheimer's disease. 2022.08.05.22278457 Preprint at https://doi.org/10.1101/2022.08.05.22278457 (2022).

23. Cravatt, B. F. *et al.* Chemical characterization of a family of brain lipids that induce sleep. *Science* **268**, 1506–1509 (1995).

24. Fowler, C. J. Oleamide: a member of the endocannabinoid family? *Br. J. Pharmacol.* **141**, 195–196 (2004).

25. Xie, L. *et al.* Sleep Drives Metabolite Clearance from the Adult Brain. *Science* **342**, 373–377 (2013).

26. Martínez, Y. *et al.* The role of methionine on metabolism, oxidative stress, and diseases. *Amino Acids* **49**, 2091–2098 (2017).

27. Kettenmann, H., Kirchhoff, F. & Verkhratsky, A. Microglia: New Roles for the Synaptic Stripper. *Neuron* **77**, 10–18 (2013).

28. Alger, B. E. & Kim, J. Supply and demand for endocannabinoids. *Trends Neurosci.* **34**, 304–315 (2011).

29. Maccarrone, M., Dainese, E. & Oddi, S. Intracellular trafficking of anandamide: new concepts for signaling. *Trends Biochem. Sci.* **35**, 601–608 (2010).

30. Picard, K. *et al.* Microglial homeostasis disruption modulates non-rapid eye movement sleep duration and neuronal activity in adult female mice. *Brain. Behav. Immun.* **107**, 153–164 (2023).

31. Hristovska, I. *et al.* Sleep decreases neuronal activity control of microglial dynamics in mice. *Nat. Commun.* **13**, 6273 (2022).

32. Garofalo, S. *et al.* Role of Glia in the Regulation of Sleep in Health and Disease. *Compr. Physiol.* **10**, 687–712 (2020).

33. Heifets, B. D. & Castillo, P. E. Endocannabinoid Signaling and Long-Term Synaptic Plasticity. *Annu. Rev. Physiol.* **71**, 283–306 (2009).

34. Kano, M., Ohno-Shosaku, T., Hashimotodani, Y., Uchigashima, M. & Watanabe, M. Endocannabinoid-Mediated Control of Synaptic Transmission. *Physiol. Rev.* **89**, 309–380 (2009).

35. Harkany, T., Mackie, K. & Doherty, P. Wiring and firing neuronal networks: endocannabinoids take center stage. *Curr. Opin. Neurobiol.* **18**, 338–345 (2008).

36. Gabrielli, M. *et al.* Active endocannabinoids are secreted on extracellular membrane vesicles. *EMBO Rep.* **16**, 213–220 (2015).

37. Stevanin, G. *et al.* Huntington's disease-like phenotype due to trinucleotide repeat expansions in the TBP and JPH3 genes. *Brain J. Neurol.* **126**, 1599–1603 (2003).

38. Whiley, L. *et al.* Evidence of altered phosphatidylcholine metabolism in Alzheimer's disease. *Neurobiol. Aging* **35**, 271–278 (2014).

39. Laera, L. *et al.* CRAT missense variants cause abnormal carnitine acetyltransferase function in an early-onset case of Leigh syndrome. *Hum. Mutat.* **41**, 110–114 (2020).

40. Arevalo-Rodriguez, I. *et al.* Mini-Mental State Examination (MMSE) for the early detection of dementia in people with mild cognitive impairment (MCI). *Cochrane Database Syst. Rev.* **7**, CD010783 (2021).

41. Gold, L. *et al.* Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery. *PLOS ONE* **5**, e15004 (2010).

42. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).

43. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

44. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

45. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **28**, 1–26 (2008).

46. Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D. & Borsboom, D. qgraph: Network Visualizations of Relationships in Psychometric Data. *J. Stat. Softw.* **48**, 1–18 (2012).

47. Gabrielli, M. *et al.* Microglial large extracellular vesicles propagate early synaptic dysfunction in Alzheimer's disease. *Brain J. Neurol.* **145**, 2849–2868 (2022).

48. Zal, T., Volkmann, A. & Stockinger, B. Mechanisms of tolerance induction in major histocompatibility complex class II-restricted T cells specific for a blood-borne self-antigen. *J. Exp. Med.* **180**, 2089–2099 (1994).

49. Dikmen, H. O. *et al.* GM-CSF induces noninflammatory proliferation of microglia and disturbs electrical neuronal network rhythms in situ. *J. Neuroinflammation* **17**, 235 (2020).

50. Prada, I. *et al.* Glia-to-neuron transfer of miRNAs via extracellular vesicles: a new mechanism underlying inflammation-induced synaptic alterations. *Acta Neuropathol. (Berl.)* **135**, 529–550 (2018).