Large-scale proteome analysis of CSF implicates altered glucose metabolism in Alzheimer's disease

Daniel J. Panyard[*1], Justin McKetney[*2,3], Yuetiva K. Deming[1,4,5], Autumn R. Morrow[1], Gilda E. Ennis[4], Erin M. Jonaitis[4,6], Carol A. Van Hulle[4,5], Carlos Cruchaga[7,8,9], Chengran Yang[7,8,9], Yun Ju Sung[7,8,9], Henrik Zetterberg[10,11,12,13,14], Kaj Blennow[10,11], Gwendlyn Kollmorgen[15], Ivonne Suridjan[16], Anna Bayfield[15], Barbara B. Bendlin[4,5,6,17], Cynthia M. Carlsson[4,5,6,17], Sterling C. Johnson[4,5,6,17], Sanjay Asthana[4,5,17], Joshua J. Coon[†2,3,18,19], Corinne D. Engelman[†1]

[1]Department of Population Health Sciences, University of Wisconsin-Madison, 610 Walnut Street, 707 WARF Building, Madison, WI 53726, United States of America

[2]National Center for Quantitative Biology of Complex Systems, University of Wisconsin-Madison, Madison, WI 53706, United States of America

[3]Department of Biomolecular Chemistry, University of Wisconsin-Madison, Madison, WI 53506, United States of America

[4]Wisconsin Alzheimer's Disease Research Center, University of Wisconsin-Madison, 600 Highland Avenue, J5/1 Mezzanine, Madison, WI 53792, United States of America

[5]Department of Medicine, University of Wisconsin-Madison, 1685 Highland Avenue, 5158 Medical Foundation Centennial Building, Madison, WI 53705, United States of America

[6]Wisconsin Alzheimer's Institute, University of Wisconsin-Madison, 610 Walnut Street, 9th Floor, Madison, WI 53726, United States of America

[7]Department of Psychiatry, Washington University School of Medicine, St Louis, MO 63110, United States of America

[8]NeuroGenomics and Informatics Center, Washington University School of Medicine, St Louis, MO 63110, United States of America

[9]Hope Center for Neurological Disorders, Washington University School of Medicine, St Louis, MO 63110, United States of America

[10]Department of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology, the Sahlgrenska Academy at the University of Gothenburg, Mölndal, Sweden

[11]Clinical Neurochemistry Laboratory, Sahlgrenska University Hospital, Mölndal, Sweden

[12]Department of Neurodegenerative Disease, UCL Institute of Neurology, London, UK

[13]UK Dementia Research Institute at UCL, London, UK

[14]Hong Kong Center for Neurodegenerative Diseases, Hong Kong, China

[15]Roche Diagnostics GmbH, Penzberg, Germany

[16]Roche Diagnostics International Ltd, Rotkreuz, Switzerland

[17]William S. Middleton Memorial Veterans Hospital, 2500 Overlook Terrace, Madison, WI 53705, United States of America

[18]Morgridge Institute for Research, Madison, WI 53706, United States of America

[19]Department of Chemistry, University of Wisconsin-Madison, Madison, WI 53506, United States of America

[*]These authors contributed equally

[†]Joint corresponding authors

Abstract

A major hallmark of Alzheimer's disease (AD) is the aggregation of misfolded proteins (β-amyloid (A) and hyperphosphorylated tau (T)) in the brain. As these proteins can be monitored by cerebrospinal fluid (CSF) measures, the AD proteome in CSF has been of particular interest. Here, we conducted a proteome-wide assessment of the CSF in an AD cohort among participants with and without AD pathology (n = 137 total participants: 56 A-T-, 39 A+T-, and 42 A+T+; 915 proteins analyzed), identifying a diverse set of proteins in the CSF enriched for extracellular and immune system processes. We then interrogated the proteome using the amyloid, tau, and neurodegeneration (ATN) framework of AD and a panel of 9 CSF biomarkers for neurodegeneration and neuroinflammation. After multiple testing correction, we identified a total of 61 proteins significantly associated with AT group ($P < 5.46 \times 10^{-5}$; strongest was SMOC1, $P = 1.87 \times 10^{-12}$) and 636 significant protein-biomarker associations ($P < 6.07 \times 10^{-6}$; strongest was a positive association between neurogranin and EPHA4, $P = 2.42 \times 10^{-25}$) across all measures except for interleukin-6, which had no significantly associated proteins. Community network and pathway enrichment analyses highlighted three biomarker-associated protein networks: one related to amyloid and tau measures, one to CSF neurogranin, and one to the remaining CSF biomarkers. Glucose metabolic pathways were enriched primarily among the amyloid- and tau-associated proteins, including malate dehydrogenase and aldolase A, both of which were replicated as strongly associated with AD ($P = 1.07 \times 10^{-19}$ and $P = 7.43 \times 10^{-14}$, respectively) in an independent CSF proteomics cohort (n = 717 participants). Comparative performance of the CSF proteome in predicting AT categorization was high (mean AUC range 0.891–0.924 with number of protein predictors ranging from 37-97) relative to other omic predictors from the genome, CSF metabolome, and demographics from the same cohort of individuals. Collectively, these results emphasize the importance of the CSF proteome relative to other omics and implicate glucose metabolic dysregulation as amyloid and tau pathology emerges in AD.

Introduction

Despite much improvement in our understanding of it, Alzheimer's disease (AD) continues to impose an enormous medical, social, and economic toll on society. An estimated 50 million people have dementia worldwide, with that number likely to increase to over 150 million by 2050[1]. AD is the 6th leading cause of death in the U.S. and costs an estimated $290 billion annually for healthcare[2]. Part of the reason for this global impact of AD is the lack of effective therapies for the disease, which is driven in part by an incomplete understanding of its causal mechanisms[3]. The core pathological features of AD are well-described and center on the accumulation of two proteins, amyloid and tau, into amyloid plaques and neurofibrillary tangles[4], for which there are validated cerebrospinal fluid (CSF) biomarkers[5].

In order to better inform research on AD, there has been a shift in the conceptualization of the disease from a focus on clinical signs and symptoms[6] to AD biology measured *in vivo*. Using CSF assays related to amyloid deposition and hyperphosphorylation of tau protein (in addition to neuroimaging), it has become possible to leverage these biomarkers for identifying preclinical AD, mild cognitive impairment (MCI), and AD dementia[7–10]. Most recently, in 2018, an explicit research framework for categorizing AD was proposed by the National Institute on Aging and Alzheimer's Association (NIA-AA). This framework categorized individuals as amyloid positive (A+), tau positive (T+), and/or neurodegeneration positive (N+)[11]. This so-called ATN framework—using ATN-based categorizations rather than more traditional clinical diagnoses as outcomes—provided nosological clarity in studying AD and other forms of dementia.

The use of these biomarker-defined categories is most relevant in multiomic approaches to studying AD pathophysiology, where molecular pathways are interrogated and clear case definitions are essential. Omics research offers immense promise for understanding complex disease by leveraging analyses of millions of molecular features spanning from the genome to the proteome, metabolome, phenome, and beyond[12]. In the field of AD research, each of these individual omic approaches has already been applied extensively. Genomics research has highlighted a number of important loci, from the role of mutations in *APP*, *PSEN1*, and *PSEN2* in early-onset familial AD[13] to late-onset AD genetic risk factors like *APOE*, *CR1*, and *ABCA7*[14–16]. CSF metabolomics studies have identified alterations in cholesterol, sphingolipid, norepinephrine, and other pathways[17,18]. In the CSF proteome, already known to include the amyloid and tau biomarkers for AD, studies have identified altered proteins related to the immune system and inflammation, carbohydrate metabolism, phospholipids, and the regulation of synapses[19–24].

Here, we combined the A and T of the ATN framework of AD with a novel CSF proteomics data set comprising 915 proteins generated for 137 participants, building on our recently published pilot study results in an independent sample[25]. We comprehensively profiled the AD CSF proteome, its relationship to AT category, and its association with a diverse set of 9 AD CSF biomarkers covering measures of amyloid, tau, neurodegeneration, and neuroinflammation. These results were then extensively interrogated for pathway-level and network-based patterns, with top findings replicated in an independent cohort. Finally, we combined the proteomics data set with previously generated genome-wide genotypes, 390 CSF metabolites, and demographic information to examine the relative utility of different omics data sets in predicting the AT-based categories. Elucidating the pathophysiology leading

to the development of AD pathology and symptoms of AD dementia is expected to inform the identification of novel, effective drug targets.

Methods

Study participants

The data in this study came from two longitudinal AD cohorts of middle- and older-aged adults: the Wisconsin Registry for Alzheimer's Prevention (WRAP)[26] and the Wisconsin Alzheimer's Disease Research Center (ADRC)[27] (Table 1). Briefly, WRAP includes participants enriched for a parental history of AD dementia who were largely between the ages of 40 and 65 at the time of enrollment, fluent in English, able to perform neuropsychological testing, without a diagnosis or evidence of dementia at baseline, and without any health conditions that might prevent participation in the study. The ADRC study includes participants from one of several subgroups: mild late-onset AD, MCI, cognitively unimpaired middle-aged adults enriched for a parental history of presumed AD dementia, and age-matched healthy older controls (age > 65). Briefly, the ADRC participants were over the age of 45, with decisional capacity, and without a history of certain medical conditions (like congestive heart failure or major neurologic disorders other than dementia) or any contraindication to biomarker procedures. Participants in both the WRAP and ADRC cohorts were given diagnoses of AD, MCI, cognitively unimpaired, and others that were reviewed by a consensus review committee that included dementia-specialist physicians, neuropsychologists, and nurse practitioners[26]. The National Institute of Neurological and Communicative Disorders and Stroke and Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA)[6] and NIA-AA[7] criteria were used in defining the clinical diagnoses without reference to the participants' CSF biomarker status. This study used the STROBE cohort reporting guidelines[28] and was performed as part of the GeneRations Of WRAP (GROW) study, which was approved by the University of Wisconsin Health Sciences Institutional Review Board. Participants in the ADRC and WRAP studies provided written informed consent.

CSF biomarkers

The CSF samples used for the biomarker analyses were acquired from lumbar punctures (LPs) using a uniform preanalytical protocol between 2010 and 2018 as previously described[29]. Samples were collected in the morning using a Sprotte 24- or 25-gauge atraumatic spinal needle, and 22 mL of fluid was collected via gentle extraction into polypropylene syringes and combined into a single 30 mL polypropylene tube. After gentle mixing, samples were centrifuged to remove red blood cells and other debris. Then, 0.5 mL CSF was aliquoted into 1.5 mL polypropylene tubes and stored at -80 degrees Celsius within 30 minutes of collection.

All CSF samples were assayed between March 2019 and January 2020 at the Clinical Neurochemistry Laboratory at the University of Gothenburg. CSF biomarkers were assayed using the NeuroToolKit (NTK) (Roche Diagnostics International Ltd, Rotkreuz, Switzerland), a panel of automated Elecsys® and robust

prototype immunoassays designed to generate reliable biomarker data that can be compared across cohorts. Measurements with the following immunoassays were performed on a cobas e 601 analyzer (Roche Diagnostics International Ltd, Rotkreuz, Switzerland): Elecsys β-amyloid (1–42) CSF (Aβ42), Elecsys Phospho-Tau (181P) CSF (ptau), and Elecsys Total-Tau CSF, β-amyloid (1–40) CSF (Aβ40), and interleukin-6 (IL-6). The remaining NTK panel was assayed on a cobas e 411 analyzer (Roche Diagnostics International Ltd, Rotkreuz, Switzerland), including markers of synaptic damage and neuronal degeneration (neurogranin, neurofilament light protein [NFL], and alpha-synuclein) and markers of glial activation (chitinase-3-like protein 1 [YKL-40] and soluble triggering receptor expressed on myeloid cells 2 [sTREM2]).

A total of nine established CSF biomarkers for AD were analyzed in this study: the Aβ42/Aβ40 ratio, ptau, the ptau/Aβ42 ratio, NFL, alpha-synuclein, neurogranin, YKL-40, sTREM2, and IL-6. Since the CSF biomarker measurements were to be used as outcomes, each biomarker was assessed for skewness using the skewness function of the R package moments (version 0.14)[30]. Any biomarker with a skewness ≥ 2 was transformed with a $\log_{10}$-transformation to better meet the normality assumption of regression. The outcomes that were $\log_{10}$-transformed were NFL and IL-6.

Samples used in this study were then assigned to pathological categories from the NIA-AA ATN research framework[11] using binary cut-offs for CSF amyloid and tau positivity. The development of these research cut-offs is described in detail elsewhere[29]. Briefly, cut-offs were estimated via ROC analysis on a subsample of n = 185 participants (cognitively impaired and unimpaired) who underwent [11C] PiB-PET imaging within two years of an LP. Using the Matlab perfcurve function[31] with an equally weighted cost function[32], the optimal Aβ42/Aβ40 threshold was 0.046 and the optimal ptau/Aβ42 threshold was 0.038. Thresholds for ptau181 were determined by establishing a reference group of 223 CSF amyloid (Aβ42/Aβ40) negative, cognitively unimpaired younger participants (ages 40-60 years). Biomarker positivity thresholds for these analytes were set at +2 standard deviations (SD) above the mean of this reference group (ptau threshold = 24.8 pg/mL). In this study, A+ and T+ were defined based on the CSF Aβ42/Aβ40 and ptau thresholds, respectively. The final pathological categories for this study included amyloid negative and tau negative (A-T-); amyloid positive and tau negative (A+T-); and amyloid positive tau positive (A+T+). The fourth possible category of amyloid negative and tau positive (A-T+) was not included in this study as these samples are considered to represent non-AD pathological change[11].

CSF metabolomics

All samples used in this study had CSF metabolomics data available from the WRAP and ADRC cohorts. The details of the CSF sample collection, handling, and metabolomics profiling have been previously described[33,34]. Briefly, fasting CSF samples were drawn from study participants in the morning through LP and then mixed, centrifuged, aliquoted, and stored at -80 degrees Celsius. Samples were kept frozen until they were shipped overnight to Metabolon, Inc. (Durham, NC), which similarly kept samples frozen at -80 degrees Celsius until analysis. Metabolon used Ultrahigh Performance Liquid Chromatography-Tandem Mass Spectrometry (UPLC-MS/MS) to conduct an untargeted metabolomics analysis of the CSF samples. The metabolites were then annotated with metabolite identifiers, chemical properties, and

pathway information. Metabolite measurements were divided by the median measurement for that metabolite across all samples. Missing values for xenobiotic metabolites were imputed to 0.0001, while missing values for non-xenobiotic metabolites were imputed to half of the minimum value among all other measured samples for that metabolite.

The initial data set contained 412 metabolites from 1,172 CSF samples across 687 unique individuals. A total of 13 metabolites that were missing for ≥ 50% of samples were removed. One sample was removed for missing ≥ 40% of metabolite values. A total of 9 metabolites with low variance (interquartile range = 0) were then removed. A $\log_{10}$ transformation was applied to all metabolite values. A total of 220 samples from a clinical trial were excluded from analysis. The processed data set contained 390 metabolites quantified on 951 CSF samples from 609 unique individuals, including all of the individuals here on whom proteomics data were generated.

Genome-wide genotyping

Genome-wide genotypes were also available for all samples in this study. The genotyping in both the WRAP and ADRC had been previously conducted[35]. For the WRAP cohort, DNA from whole blood samples were genotyped with the Illumina Multi-Ethnic Genotyping Array at the University of Wisconsin Biotechnology Center[34]. Pre-imputation quality control (QC) steps included removing samples and variants with a high missingness (> 5%) or inconsistent genetic and self-reported sex. Samples from individuals of European descent were imputed using the Michigan Imputation Server[36] and the Haplotype Reference Consortium (HRC) reference panel[37]. Variants with poor quality ($R^2 < 0.8$) or out of Hardy-Weinberg equilibrium (HWE) were removed after imputation, leaving a total of 1,198 samples with 10,499,994 single nucleotide polymorphisms (SNPs). In the ADRC, whole blood samples were genotyped by the Alzheimer's Disease Genetics Consortium (ADGC) at the National Alzheimer's Coordinating Center (NACC) using the Illumina HumanOmniExpress-12v1_A, Infinium HumanOmniExpressExome-8 v1-2a, or Infinium Global Screening Array v1-0 (GSAMD-24v1-0_20011747_A1) BeadChip assay. Initial quality control was conducted on each chip's data separately, removing variants or samples with high missingness (> 2%), out of HWE ($P < 1x10^{-6}$), or with inconsistent genetic and self-reported sex. The remaining samples were then imputed with the Michigan Imputation Server, phased using Eagle2[38], and imputed to the HRC reference panel. As before, variants of low quality ($R^2 < 0.8$) or out of HWE were removed. The data sets from the different chips were then merged together, leaving a data set with 377 samples of European descent and 7,049,703 SNPs. The WRAP and ADRC data sets were then harmonized to each other and to the 1000 Genomes Utah residents with Northern and Western European ancestry (CEU)[39] data set, using the GRCh37 genome build. Ambiguous SNPs were removed, and then the remaining SNPs were aligned to the same strand and allele orientations as the ADRC data set.

The 137 samples from this study were then extracted from this combined genetic data set and further processed using PLINK[40] (v1.90b6.3). To ensure sufficient data were available for use in the prediction models, only SNPs with no missing data and with a minor allele count of 20 or greater among the 137 samples were retained. Linkage disequilibrium (LD) pruning was then applied using a window size of

1000 kb, an R^2 threshold of 0.1, and the 1000 Genomes CEU samples as the reference data set. The pruning resulted in a data set of 38,652 SNPs.

APOE *genotyping*

Each sample was additionally assigned an *APOE* genotype based on the participant's combination of the ε2, ε3, and ε4 alleles for *APOE* from a separate set of genotyping. DNA was extracted from whole blood samples, which was then genotyped for the *APOE* alleles using competitive allele-specific PCR-based KASP genotyping for rs429358 and rs7412[33].

Proteomics sample selection

Based on the results of our pilot study for CSF proteomics[25], we had estimated *a priori* that a sample of approximately 150 would be sufficient for 80% power to detect most of the observed protein-AD diagnosis associations from the pilot using the R package pwr (version 1.3-0)[41] (Supplementary Figure 1). The process of selecting samples for CSF proteomics generation began by considering all CSF samples from fasted, successful LPs (n = 1,440) from 823 unique participants across WRAP and ADRC. From there, each CSF sample was matched to its closest set of CSF biomarker data, CSF metabolomics data, and consensus conference diagnosis. Samples were excluded if there was insufficient material for proteomics analysis, if they were part of a clinical trial, or if they had been used already in our pilot study. To simplify the downstream analyses, only one sample (the most recent) per participant was considered when there were multiple samples. An approximately equal number of samples per AT-defined subgroup (A-T-, A+T-, A+T+) was selected, prioritizing samples with available genomic data and metabolomic data. A total of 140 samples were selected to have proteomics data generated.

Protein extraction and digestion

CSF protein concentration was determined by protein BCA assay (Thermo Scientific). CSF aliquots were moved to 96-plates and dried down using a SpeedVac Concentrator (Thermo Scientific) before being resuspended in a lysis buffer consisting of 10 mM TCEP, 40 mM CAA, 100 mM Tris pH 8, and 8M urea. The sample solution was then diluted to 25% strength using 100 mM Tris pH 8 before the addition of protease. Trypsin was added to the protein solution at an approximate ratio of 50:1 w/w and digested overnight at ambient temperature. The digestion reaction was quenched by acidification using TFA. Digested peptides were desalted using Strata-X Polymeric Reverse Phase plates (Phenomenex) before being dried down in the SpeedVac Concentrator overnight. Dried down samples were resuspended in 0.2% FA and peptide concentration was determined using a peptide BCA (Thermo Scientific). Peptide samples were injected directly from the 96-well plates.

Offline fractionation

Pooled samples for each of the three disease groups were created by combining 10 µL of CSF from each sample in that disease group. These three pooled samples were then prepared using the extraction and digestion protocol described above. The three desalted, digested peptide solutions were then fractionated using high-pH reverse-phase liquid chromatography. Separation was performed using an Agilent Infinity 2000 HPLC with a 150 mm C18 reverse-phase column (Waters, XBridge Peptide BEH, particle size 3.5 µm). Mobile phase buffer A was a freshly prepared mixture 10 mM ammonium formate pH 9.5, and mobile phase buffer B was a freshly prepared mixture of 80% MeOH, 10 mM ammonium formate pH 9.5. The gradient method was 20 minutes in length with fractions collected from minute 5 to minute 20, with a flow a rate of 800 nL/min across the entire method. The method initiated with a concentration of 5% B before increasing to 35% by minute 2. Percent B increased to 100% by 13 minutes. From 5 to 20 minutes, 32 fractions were collected in round-bottom 96 well plates in a time-based manner. Fractions were concatenated into a total of 16 by combining every other column in the collection plate. Fractionated samples were injected directly from the collection plate.

Online chromatography

We used a method previously developed in our pilot study[25]: a single-shot nano-liquid chromatography-tandem mass spectrometry (nLC-MS/MS) method for quantitative and fast analysis of CSF protein extracts. Reverse phase columns used online with the mass spectrometer were packed using an in-house column packing apparatus described previously[42]. In brief, 1.5 µm BEH particles were packed into a fused silica capillary purchased from New Objective (PicoTip, Stock # PF360-75-10-N-5) at 30,000 psi. During online LC separations, capillary was heated to 50° C and interfaced with mass spectrometer via an embedded emitter. For online chromatography, a Dionex UltiMate 300 nanoflow UHPLC was used with mobile phase A consisting of 0.2% FA and mobile phase B consisted of 70% ACN, 0.2% FA. A flow rate of 310 nL/min was used throughout with the method increasing from 0% to 7% B over the first four minutes. Percent B then increased to 49% B by 59 minutes before a wash step of 100% B from 62 to 67 minutes. Method finished with an equilibration step from minute 68 to 78 of 0% B.

Tandem mass spectrometry

Peptides eluting from the column were ionized by electrospray ionization and analyzed using a Thermo Orbitrap Eclipse hybrid mass spectrometer. Survey scans were collected in the Orbitrap at a resolution of 240,000 with a normalized AGC target of 250% (1e6) with Advanced Precursor Determination engaged across the range of 300–1400 m/z. Precursors were isolated for tandem MS scans using a window of 0.5 m/z, with a dynamic exclusion duration of 22 seconds and a mass tolerance of 15 ppm. Precursors were dissociated using HCD with a normalized collision energy of 25%. Tandem scans were taken over the range 130–1350 m/z using the “rapid” setting with a normalized AGC target of 300% (3e4) and a maximum injection time of 18 ms.

The resulting raw data files were searched in MaxQuant[43,44] using fast LFQ and a full human proteome with isoforms downloaded from UniProt (downloaded June 14, 2017). Oxidation of methionine and

acetylation of the N terminus were allowed as variable modifications, and carbamidomethylation of cysteine was set as a fixed modification. Proteins were searched using an FDR of 1% with a minimum peptide length of 7 and a 0.5 Da MS/MS match tolerance. Matching between runs was utilized, applied with a retention time window of 0.7 minutes. Protein abundance data were extracted in the form of LFQ Intensity from the "proteinGroups.txt" output file. Throughout this manuscript, each protein group is referred to by the first listed majority protein from its annotation from MaxQuant. The protein data were annotated with Entrez IDs (via R package org.Hs.eg.db[45], version 3.11.4), UniProt[46] IDs, and gene information (GENCODE[47], version 37, and the HUGO Gene Nomenclature Committee, HGNC, database[48]). When the gene annotations conflicted or were absent from one of these databases for a given UniProt ID, the gene identifiers were taken in the order of resources listed.

Proteomics quality control

After removing several samples with injection or other technical issues, the proteomics data set included 2,040 proteins across 137 samples. These data underwent a strict quality control pipeline: proteins that were missing for 33% or more of samples (either overall or within an AT group) were removed; samples missing 33% or more of proteins were removed; and proteins with an interquartile range of 0 were removed (Supplementary Figure 2, Supplementary Figure 3). A total of 137 samples with 915 proteins remained (Supplementary Table 1). The label-free quantification (LFQ) values for each protein were then $\log_2$-transformed. The remaining missing values were then randomly imputed based on a normal distribution derived from the lower end of the observed values for that protein (the observed distribution mean was shifted by -1.8 and the SD shrunk by a factor of 0.3) (Supplementary Figure 4). This imputation was performed separately within each AT group. Finally, each protein was Z-score transformed.

Proteomics descriptive analysis

The set of proteins quantified in the CSF was extensively profiled (Supplementary Table 2). The pairwise correlation of all proteins was calculated (nominally significant results with correlation $P < 0.05$ shown in Supplementary Table 3) and then visualized with a heatmap with hierarchical clustering to show the underlying patterns of covariation (R package ComplexHeatmap[49], version 2.4.3) (Figure 1a). The structure was further examined with a principal components analysis (PCA), scree plot (Supplementary Figure 5), and plot of the first two PCs by AT group (R package factoextra[50], version 1.0.7) (Supplementary Figure 6) to assess the presence of independent signals among the CSF proteome and their relationship to the AT groups. A pathway analysis was then performed to examine the differences between the set of proteins quantified in the CSF and the overall human proteome. The enrichment of Gene Ontology (GO) terms[51], Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways[51,52], and Disease Ontology (DO) gene sets[53] among the CSF proteome against the entire human proteome was calculated using the R packages clusterProfiler[54] (version 3.16.1), DOSE[55] (version 3.14.0), and ReactomePA[56] (version 1.32.0) (Figure 1b, Figure 1c, Supplementary Table 4). The presence of clusters

was then assessed with Gaussian mixture modeling using the R package mclust[57] (version 5.4.6). The number of clusters (3) was chosen based on the elbow of the plot of the Bayesian Information Criterion (BIC) (Supplementary Figure 7). Enrichment of gene set ontologies across the clusters was repeated with the GO, KEGG, and DO sets and plotted (Figure 1d, Supplementary Table 5).

Protein-AT category associations

The primary objective was the identification of differentially expressed proteins across the three AT groups. This analysis was performed using an analysis of covariance (ANCOVA) model comparing each protein across the three groups, controlling for age at LP and sex (Supplementary Table 6). A Bonferroni correction for the number of proteins tested ($P = 0.05 / 915 = 5.46 \times 10^{-5}$) was used for reporting significant results. The distributions of the top-associated proteins across the AT spectrum were plotted (Figure 2a). To assess whether signal enrichment was likely due to an artifact, the ANCOVA analyses were repeated with randomly permuted AT group labels. A quantile-quantile (Q-Q) plot was generated to assess the presence of signal enrichment across the proteome for AT-related differences and to compare the permuted and non-permuted analyses (Figure 2b). Since the *APOE* gene is known to have a significant effect on AD risk, we examined whether *APOE* genotype was driving the observed AT-protein associations. The ANCOVA analyses were repeated but with the count of *APOE* ε4 alleles included as an additional covariate. The same Bonferroni correction was used as before. The resulting AT-associated proteins were compared to the results from the original ANCOVA analyses (Supplementary Figure 8). The set of associated proteins from the non-permuted analysis was then assessed for enriched GO, KEGG, and DO gene sets against the human proteome as before (Figure 2c, Supplementary Table 7).

To examine the direction of effect of each protein, a logistic regression was performed with A+T+ (vs. A-T-) as the outcome and a protein as the main predictor, controlling for age at LP and sex and using the same Bonferroni threshold for significance as the ANCOVA analyses. The sample size for the logistic regression was smaller (n = 98) due to the exclusion of the A+T- samples. The overlap between the set of significantly associated proteins and the set of significantly associated proteins from the ANCOVA analysis was displayed in a Venn diagram (R package ggVennDiagram[58], version 1.0.7) (Supplementary Figure 9), and the odds ratios were presented in a volcano plot (Supplementary Figure 10).

Protein-CSF biomarker associations

Each protein was then tested for association with each of the 9 CSF biomarkers (the Aβ42/Aβ40 ratio, ptau, the ptau/Aβ42 ratio, NFL, alpha-synuclein, neurogranin, YKL-40, sTREM2, and IL-6). Linear regression models were used to regress each CSF biomarker on each protein, controlling for age at CSF sample and sex and using a Bonferroni correction for the total number of tests ($9 \times 915 = 8{,}234$ tests; $P = 0.05 / 8{,}234 = 6.07 \times 10^{-6}$) (Table 2, Supplementary Table 8). The results were summarized with a Q-Q plot showing the signal enrichment for each biomarker along with a sensitivity analysis where the regression models were repeated with the biomarker values randomly permuted to test the robustness of each biomarker's signal enrichment (Supplementary Figure 11). The cross-biomarker relationships

among the significantly associated proteins were then visualized as a bipartite graph using the R package tidygraph (version 1.2.0)[59] and the Fruchterman-Reingold algorithm. A community structure network analysis was performed using the greedy hierarchical agglomeration algorithm[60] implemented in igraph (version 1.2.5)[61] to identify clusters of proteins among the protein-biomarker associations. An upset plot was then created showing the set of significantly associated proteins unique to each subset of biomarkers (Supplementary Figure 12) using the UpSetR package (version 1.4.0)[62]. Pathway enrichment analyses were performed as before comparing each biomarker's set of significantly associated proteins (Supplementary Table 9).

Replication of top results in the Knight ADRC

A replication data set from the Knight ADRC was used to validate findings from the main analyses. The Knight ADRC data set included samples from the brain (n = 360 participants), CSF (n = 717), and plasma (n = 490). The recruited individuals from the Knight ADRC cohort were evaluated by Clinical Core personnel of Washington University. For brain samples, a brain autopsy was performed by the Knight ADRC neuropathological core, and AD status was determined by a postmortem neuropathological analysis. Neuropathological phenotypes including Braak tau, Consortium to Establish a Registry for Alzheimer's Disease (CERAD) Aβ, α-synuclein pathology, postmortem interval (PMI), age at onset (AAO), age at death, and brain weight were obtained for all brain samples. For individuals with CSF and plasma data, cases received a clinical diagnosis of AD in accordance with standard criteria, and AD severity was determined using the clinical dementia rating (CDR) scale[63] at the time of lumbar puncture (for CSF samples) or blood draw (for plasma samples). Brain tissues were collected from fresh frozen human parietal lobes. CSF and plasma samples were collected in the morning after an overnight fast and immediately centrifuged and stored at -80°C until assayed[64]. The Institutional Review Boards of Washington University School of Medicine in St. Louis approved the study; research was performed in accordance with the approved protocols.

For deep omics characterization in the brain, CSF, and plasma tissues, the level of 1,305 proteins were quantified using the SOMAscan assay, a multiplexed, aptamer-based platform[65]. The assay covers a dynamic range of 10^8 and measures all three major categories: secreted, membrane, and intracellular proteins. The proteins cover a wide range of molecular functions and include proteins known to be relevant to human disease. As previously described by Gold et al.[65], modified single-stranded DNA aptamers are used to bind specific protein targets that are then quantified by a DNA microarray. Protein concentrations are quantified as relative fluorescent units (RFU). Aliquots of 150 µL of tissue were sent to the Genome Technology Access Center at Washington University in St. Louis for protein measurement.

Quality control was performed at the sample and aptamer levels using control aptamers (positive and negative controls) and calibrator samples. At the sample level, hybridization controls on each plate were used to correct for systematic variability in hybridization. The median signal over all aptamers was used to correct for within-run technical variability. This median signal was assigned to different dilution sets

within each tissue. For brain and CSF samples, a 20% dilution rate was used. For plasma samples, three different dilution sets (40%, 1%, and 0.005%) were used.

As described in detail[66], additional quality control was performed by identifying and removing protein and analyte outliers by applying four criteria: 1) Minimum detection filtering. If the analyte for a given sample was less than the limit of detection (LOD), the sample was deemed an outlier. Collectively, if the number of outliers given an analyte was less than 15% of the total sample size, the analyte was kept. 2) Flagging analytes based on the scale factor difference. 3) Coefficient of variation (CV) of calibrators lower than 0.15, where the CV for each aptamer was calculated as the standard deviation divided by the mean of each calibrator at the raw protein level. 4) Interquartile range (IQR) strategy. Outliers were identified if the subject was located 1.5-fold of the IQR outside of either end of the distribution given the $\log_{10}$-transformation of the protein level. Analytes were kept after passing all the criteria above for the downstream statistical analysis. An orthogonal approach was used to call subject outliers based on IQR. After this second removal of analytes, subject outliers were examined and removed again.

To obtain the proteomic signatures of sporadic AD status, differential abundance analysis was performed by using linear regression of the log-transformed protein levels. In all three tissues, sporadic AD status was considered as a main predictor. In each tissue, we performed surrogate variable analysis while protecting status and age to correct for unmeasured heterogeneity[67]. Age at death or at measurement, gender, and the resulting surrogate variables were included as covariates. To validate identified proteins using the related traits, analyses using AAO and AD neuropathology characteristics (Braak scores and CDR at death) were considered for brain data, AAO and CSF ptau/Aβ ratio were considered for CSF data, and AAO was considered for plasma data.

The minimum principal component (PC) number that cumulatively explained 95% of the variance for each tissue after QC was obtained. The number of PCs was 169, 230, and 75 in CSF, plasma, and brain data, respectively. These numbers corresponded to 2.96×10^{-4}, 2.17×10^{-4}, and 6.67×10^{-4}, respectively, as the Bonferroni-corrected P value thresholds. The use of these thresholds was more conservative than the false discovery rate (FDR).

The Knight ADRC analyses were used as a replication data set for the top findings from the main analyses performed in the Wisconsin ADRC and WRAP, focusing on the significantly associated proteins from the top implicated biological pathway from the protein-AT group and protein-biomarker analyses. The association of these proteins were compared to the results from the Knight ADRC association tests conducted in the CSF, brain, and plasma to see if their associations and directions of effect were replicated using the significance thresholds defined above (Table 3, Supplementary Table 10).

Secondary analysis of insulin-related proteins

Based on the results of the AT category and biomarker associations and the pathway analysis suggesting a relationship with glucose regulation (described below), the set of proteins excluded during the QC process due to low sample size was examined for proteins related to insulin signaling pathways, including any of the GLUT proteins (SLC2A family), insulin (INS), insulin receptor (INSR), insulin-like

growth factor 1 (IGF1), IGF-1 receptor (IGF1R), insulin receptor substrate 1 (IRS1), IRS 2 (IRS2), phosphoinositide 3-kinase (PI3K), RAC-alpha serine/threonine-protein kinase (AKT1), mechanistic target of rapamycin (mTOR), and glycogen synthase kinase 3 (GSK3A). Proteins that failed the missingness threshold of 33% but were present for 50% or more of samples were investigated further but without the use of imputed data points. The relationships between the proteins and AT category (Supplementary Figure 13) and the CSF biomarkers (Supplementary Figure 14) were plotted, with ANCOVA and linear regression analyses to test for association between the proteins and AT group and the CSF biomarkers performed as previously (Supplementary Table 11).

Multiomic prediction of amyloid and tau

The CSF proteomic data set was then combined with the CSF metabolomic, genome-wide genotyping, and demographic (age at sample and sex) data sets. After the quality control steps described previously for each ome, all 137 samples had values for all of the multiomic features (915 proteins, 390 metabolites, 38,652 SNPs, and 2 demographic features). The multiomic data set was then used to predict different biomarker positivity states[29]: Aβ42/Aβ40-positive, ptau-positive, and ptau/Aβ42-positive. Each ome (CSF proteome, CSF metabolome, genome, and demographics) was used individually along with a fifth multiomics predictor set (comprising all omes) to predict each outcome with an elastic net[68] model (R package glmnet[69], version 4.0-2; alpha parameter = 0.5).

For each biomarker and predictor pair, the procedure was the same. First, one-third of the data was held out as a testing set and the remaining two-thirds used as the training set. Within the training set data, 100-iteration, 3-fold cross-validation was used to select the best lambda value (11 possible values ranging from 10^{-5} to 1) according to AUC using the tidymodels[70] R package (version 0.1.3). The best-performing model was then run on the entire training data set using the chosen lambda and used to predict the outcome on the held-out testing data set. The performances of the different omic models were then compared with ROC curves and 2D histograms showing the raw biomarker levels against the predicted classifications for each biomarker for each subject (Figure 4). The mean model metrics across each of the 4,000 folds were calculated (Supplementary Table 12).

Results

Sample summary

CSF samples from 137 WRAP and ADRC participants were selected as described in the Methods, roughly evenly distributed across the three AT groups of interest (Table 1, Supplementary Figure 1). Most (102, 74.5%) of the participants were cognitively unimpaired at the time of the sample, with 16 (11.7%) and 19 (13.9%) participants having an MCI or AD dementia diagnosis, respectively. The age and sex distributions across the AT categories varied, with worse AT pathology having a higher average participant age and a greater proportion of males. The amyloid and tau measures reflected the AT

categorizations as expected. The remaining CSF biomarkers showed a general increase with increasing AT pathology with the exception of IL-6, which fluctuated across the groups.

Table 1. Summary of sample demographics and CSF biomarkers

	Overall	A-T-	A+T-	A+T+
Participants (N, %)	137	56 (40.9)	39 (28.5)	42 (30.7)
Age at CSF sample (mean, SD)	66.1 (8.3)	61.8 (6.7)	67.6 (7.2)	70.6 (8.5)
Female (N, %)	82 (59.9)	39 (69.6)	23 (59)	20 (47.6)
Aβ42/Aβ40 (mean, SD)	0.048 (0.019)	0.068 (0.011)	0.039 (0.006)	0.030 (0.007)
ptau/Aβ42 (mean, SD)	0.045 (0.034)	0.018 (0.006)	0.044 (0.014)	0.082 (0.034)
ptau (pg/mL) (mean, SD)	23.5 (12.0)	15.9 (4.2)	18.6 (3.6)	38.1 (10.9)
NFL (pg/mL) (mean, SD)	119.3 (74.2)	80.8 (25.9)	112.9 (57.1)	176.6 (94.4)
neurogranin (pg/mL) (mean, SD)	902.4 (357.6)	752.9 (251.9)	711.7 (199.7)	1278.8 (303.3)
alpha-synuclein (pg/mL) (mean, SD)	178.0 (77.6)	145.7 (58.1)	143.2 (45.9)	253.4 (71.3)
YKL-40 (ng/mL) (mean, SD)	164.9 (62.5)	133.0 (37.3)	153.9 (48.5)	217.7 (67.6)
sTREM2 (ng/mL) (mean, SD)	8.5 (2.5)	7.9 (2.2)	7.6 (2.0)	9.9 (2.8)
IL-6 (pg/mL) (mean, SD)*	5.0 (4.5)	5.1 (4.7)	5.5 (5.5)	4.5 (2.7)

*2 samples were missing IL-6 measurements and were thus excluded from analyses of IL-6

CSF proteomics descriptive analyses

The nLC-MS/MS analysis, MaxQuant identification, and LFQ quantification generated a total of 2,040 protein groups across the participants. After the proteomics quality control steps (Supplementary Table 1, Supplementary Figure 2, Supplementary Figure 3, Supplementary Figure 4), 915 proteins remained (Supplementary Table 2). Included in these proteins were YKL-40 (correlation with immunoassay measurement = 0.352, $P = 2.40 \times 10^{-5}$), sTREM2 (correlation with immunoassay measurement of sTREM2 = 0.490, $P = 1.26 \times 10^{-9}$), apolipoprotein E (APOE), amyloid precursor protein (APP; correlation with Aβ42 = 0.136, P = 0.114), amyloid-like protein-1 (APLP1), and APLP2. The tau protein was not reliably quantified by nLC-MS/MS in our samples. Little difference in protein missingness was seen by AT group (Supplementary Figure 2, Supplementary Figure 3). The CSF proteome showed a rich correlation structure with both larger clusters and smaller pockets of highly correlated proteins (Figure 1a, Supplementary Table 3). Further interrogation with PCA underscored this complexity, with the first 4 PCs collectively explaining only half (49.89%) of the total variance (Supplementary Figure 5), with the top 2 PCs not explained by either AT or sex (Supplementary Figure 6).

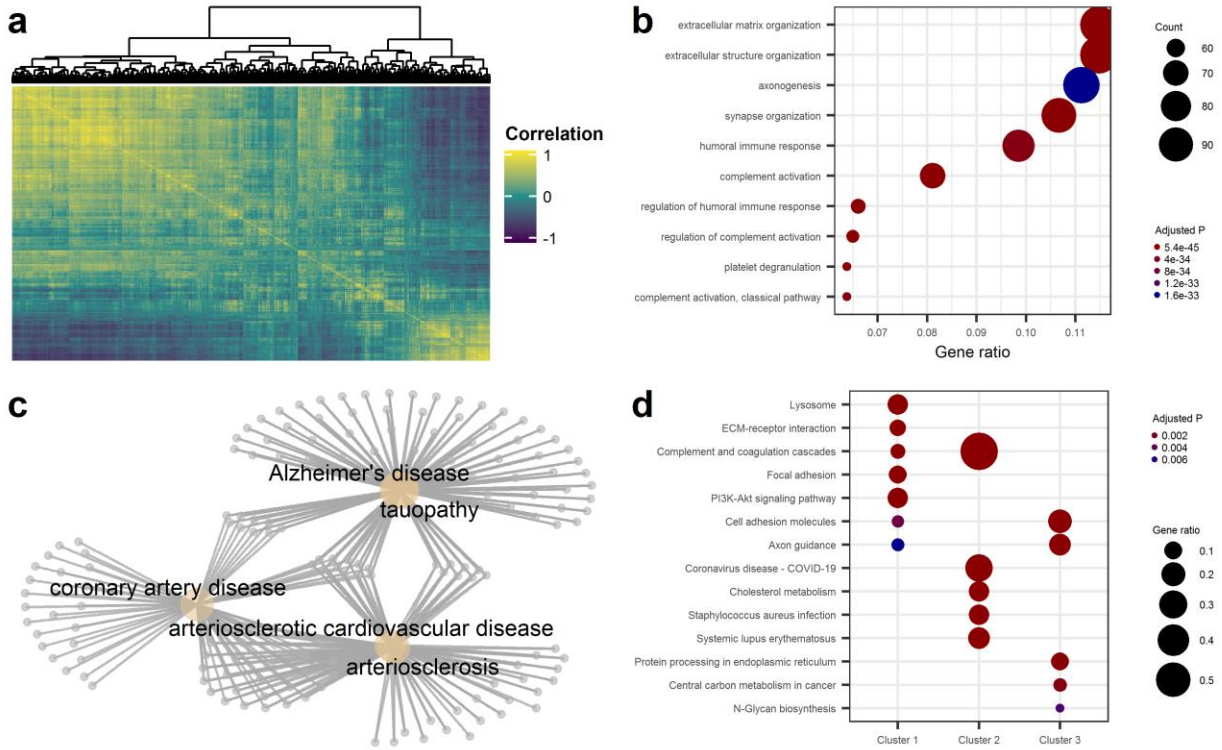

Figure 1: CSF proteomics descriptive analyses

a) A correlation heatmap of the CSF proteins (n = 915) across all participants (n = 137) is shown. The dendrogram above the heatmap shows the results of hierarchical clustering of the proteins. An intricate set of correlation patterns can be seen, with both large clusters of proteins (e.g., top-left and bottom-right of the plot) and small local clusters seen throughout. **b)** An enrichment plot of the GO biological process terms among quantified proteins in the CSF compared to the entire human genome is shown. Significantly enriched processes included extracellular processes, processes involving axons and synapses, and immune system processes. **c)** A network representation of enriched DO terms among the CSF proteome is shown. The tan nodes represent significantly enriched disease terms, and the gray nodes represent proteins whose genes were associated with those terms. Three clusters of terms emerged: Alzheimer's disease and tauopathy; coronary artery disease; and arteriosclerotic cardiovascular disease and arteriosclerosis. **d)** A cluster comparison plot of KEGG pathways between the three main clusters of CSF proteins is shown. The clusters were generated by Gaussian mixture modelling. Cluster one shows an enrichment of extracellular matrix and cell cycle pathways; cluster two shows enrichment of immune system-related pathways; and cluster three shows enrichment of a handful of other pathways.

Pathway enrichment analysis comparing the proteins quantified in the CSF to the entire human proteome revealed significant enrichment of terms related to extracellular, neuronal, immune system, and platelet pathways (Figure 1b, Supplementary Table 4). Significantly enriched DO pathways among the cohort included three groups of proteins with some small overlap: Alzheimer's disease and tauopathy; coronary artery disease; and arteriosclerotic cardiovascular disease and arteriosclerosis (Figure 1c).

Given the apparent presence of clusters of proteins based on the correlation structure, the CSF proteome was divided into 3 clusters based on a Gaussian mixture model (Supplementary Figure 7). These three clusters were then compared to each other for the differential enrichment of biological

pathways (Supplementary Table 5). The KEGG terms revealed a pattern where the smallest cluster (2) was enriched for immune system and cholesterol pathways while the other two larger clusters were enriched for extracellular and metabolism-related pathways (Figure 1d).

Protein-AT associations

The ANCOVA tests revealed 61 statistically significant associations between proteins and AT group after multiple testing correction ($P < 5.46 \times 10^{-5}$), with a total of 496 (54.2%) of the proteins nominally associated ($P < 0.05$) (Supplementary Table 6). The differences in distribution of the top ten proteins revealed a number of different patterns in relation to amyloid and tau pathology (Figure 2a). Some proteins increased (S4R371, SMOC1) or decreased (FBLN1) consistently as AT pathology increased. Other proteins did not change from A- to A+ but did change from T- to T+ (AATC, ALDOA, GUAD, DDAH1, CRYM, and F5H5Q2). Overall, there was an enrichment of statistical signal across the proteome that was not seen in the permutation sensitivity analysis (Figure 2b). Controlling for the *APOE* ε4 allele count did not substantially change the results, with 53 of the 61 proteins remaining significantly associated when the *APOE* variable was added to the ANCOVA models (Supplementary Figure 8). Among the set of significantly associated proteins, a number of biological pathways were enriched relative to the full human proteome (Supplementary Table 7), including extracellular matrix, secretory granule, and vesicle lumen GO cellular component terms; peptidase regulation GO molecular function terms; and glucose metabolic KEGG pathways (Figure 2c).

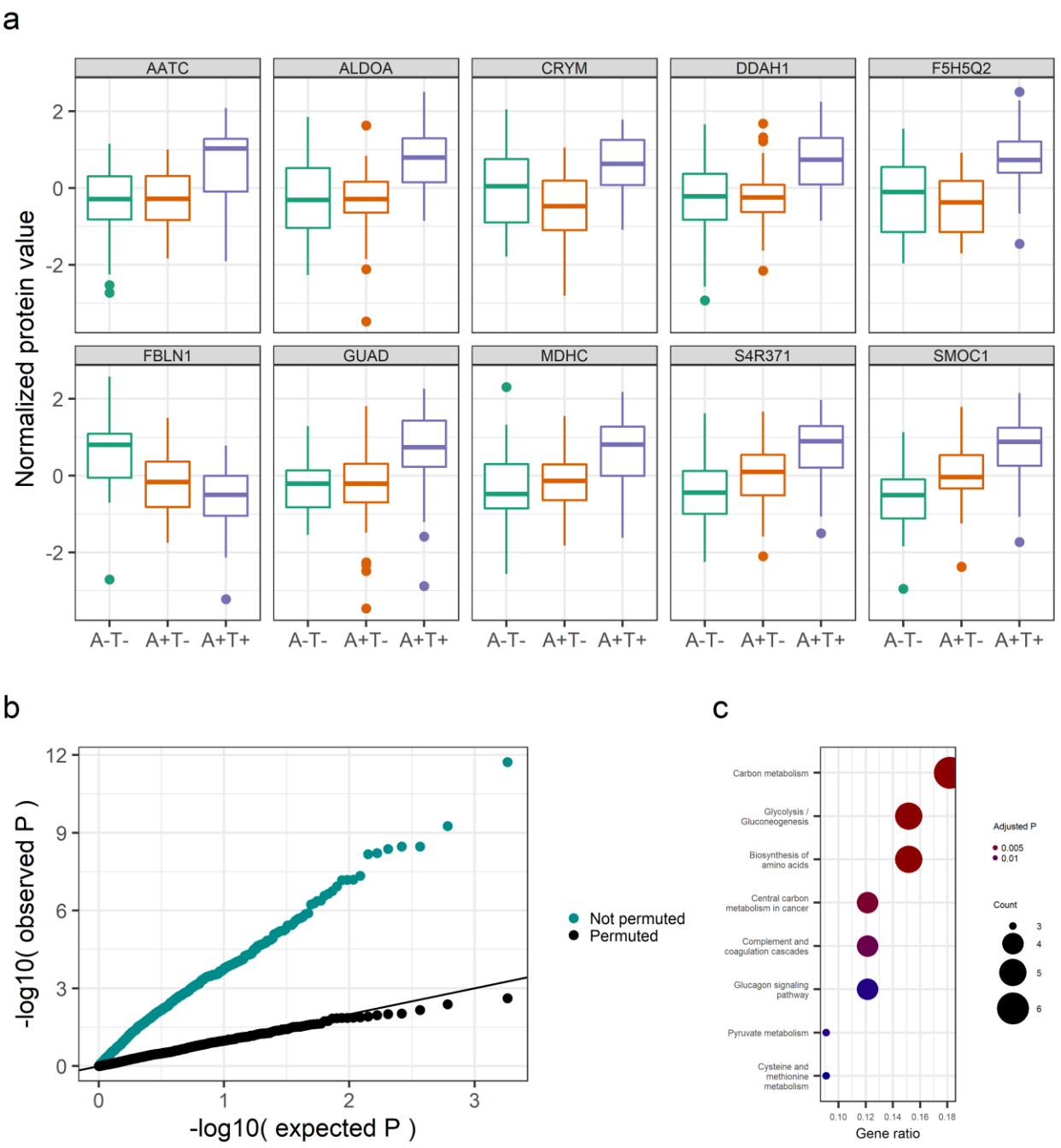

Figure 2: Proteins associated with AT group

a) The distributions of the top 10 most significantly associated proteins with AT group are shown. A number of different patterns were seen across increasing AT pathology, including proteins that increased consistently, decreased consistently, or increased only with tau positivity. **b)** The quantile-quantile (Q-Q) plots of the protein-AT ANCOVA association tests are shown above, with the distribution of P values shown separately for the original ("Not permuted") and permuted data sets. Substantial signal enrichment was seen across the CSF proteome, with that signal absent in the permuted data set. **c)** The enriched KEGG pathways among the AT-associated proteins are shown, revealing a general perturbation of metabolic pathways.

When the logistic regression model was used to test for the direction of effect between proteins and A+T+ (vs. A-T-), only 9 proteins were significantly associated with being A+T+, and all of these proteins were also significantly associated in the ANCOVA model (Supplementary Figure 9). All but 1 (FBLN1) of the proteins significantly associated with A+T+ were increased in A+T+ relative to A-T- (Supplementary Figure 10).

Protein-CSF biomarker associations

When each of the 9 CSF biomarkers (Aβ42/Aβ40, ptau, ptau/Aβ42, NFL, alpha-synuclein, neurogranin, YKL-40, sTREM2, and IL-6) was regressed on each CSF protein, a total of 636 protein-biomarker associations were statistically significant after Bonferroni correction ($P < 6.07 \times 10^{-6}$; Supplementary Table 8). As with the protein-AT associations, there was widespread association signal across the proteome with the CSF biomarkers that was not seen in the permutation test, except for IL-6, which had no significantly associated proteins (Supplementary Figure 11). The top 3 significantly associated proteins per biomarker are summarized in Table 2. A total of 119 significantly enriched pathways among the biomarker-specific sets of significantly associated proteins was observed, with glucose metabolic pathways noted to be enriched among amyloid-related biomarkers (Supplementary Table 9).

Table 2. Top 3 significantly associated proteins for each biomarker

Biomarker	UniProt ID	Entrez ID	Protein	Effect size	P value
AB42/AB40	Q9H4F8	64093	SMOC1	-0.010	4.55E-14
AB42/AB40	P40925	4190	MDH1	-0.009	1.44E-09
AB42/AB40	S4R371	2170	FABP3	-0.008	2.69E-08
alpha-synuclein	Q16270	3490	IGFBP7	-50.56	1.04E-18
alpha-synuclein	E7EUF1	5168	ENPP2	-49.04	1.96E-15
alpha-synuclein	P12259	2153	F5	-45.88	6.30E-15
log10(NFL)	Q16270	3490	IGFBP7	-0.092	4.28E-13
log10(NFL)	P23142-4	2192	FBLN1	-0.091	5.55E-12
log10(NFL)	P12259	2153	F5	-0.084	5.11E-11
neurogranin	E9PG71	2043	EPHA4	266.13	2.42E-25
neurogranin	O00451	2675	GFRA2	266.78	1.12E-23
neurogranin	Q8TAG5-2	222008	VSTM2A	259.95	8.53E-23
ptau	A0A087WZS0	66004	LYNX1	6.70	1.05E-14
ptau	P40925	4190	MDH1	6.78	3.29E-14
ptau	E9PG71	2043	EPHA4	6.62	3.81E-14
ptau/AB42	Q9H4F8	64093	SMOC1	0.019	1.83E-15
ptau/AB42	S4R371	2170	FABP3	0.018	1.68E-12
ptau/AB42	P18669	5223	PGAM1	0.017	4.72E-11
sTREM2	Q16270	3490	IGFBP7	-1.68	3.67E-19

sTREM2	E7EUF1	5168	ENPP2	-1.68	5.54E-17
sTREM2	P09486	6678	SPARC	-1.47	3.01E-13
YKL-40	Q16270	3490	IGFBP7	-28.22	4.54E-12
YKL-40	E9PG71	2043	EPHA4	24.60	3.45E-09
YKL-40	P12259	2153	F5	-24.43	4.63E-09

The network plot and subsequent community analysis of the protein-biomarker associations revealed three communities (modularity = 0.256) among the network (Figure 3). One community largely comprised the more traditional AD biomarkers of ptau, ptau/Aβ42, and Aβ42/Aβ40; a second community centered around the proteins associated with neurogranin; and the third community included the remaining biomarkers of alpha-synuclein, YKL-40, NFL, and sTREM2 (IL-6 had no significant protein associations and was not included). The largest number of shared associations across the biomarkers occurred among neurogranin, ptau, and alpha-synuclein, which shared 103 protein associations between at least two of those biomarkers (Supplementary Figure 12).

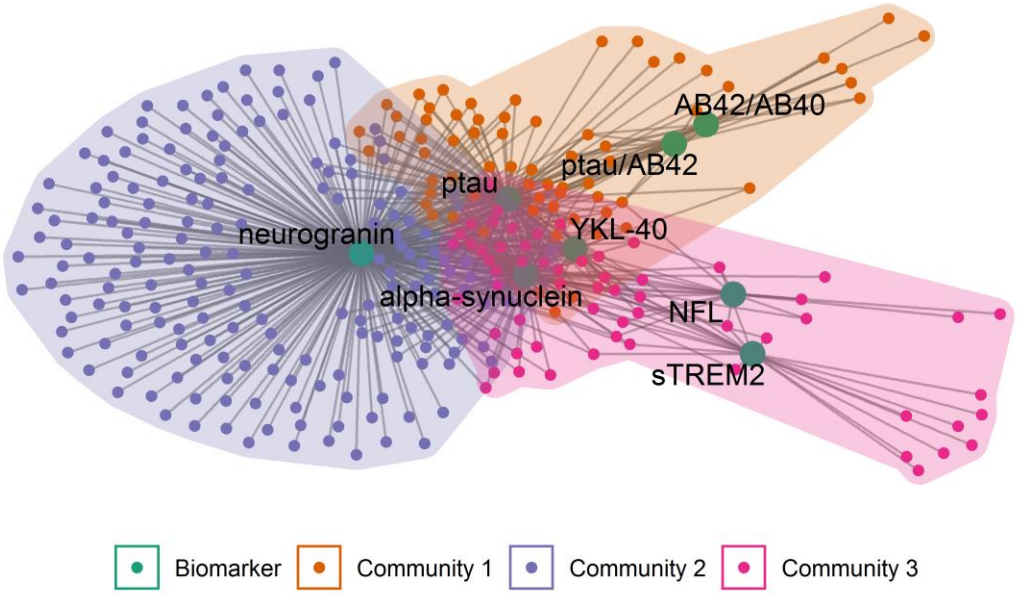

Figure 3: Network analysis of protein-biomarker associations

A bipartite graph representation of the proteins significantly associated with the CSF biomarkers after Bonferroni correction is shown. The nodes representing the biomarkers are larger and in green. Proteins are represented as smaller nodes, with an edge to a biomarker representing a significant association between a protein and a biomarker. The colors of the protein nodes and underlying shaded regions correspond to three distinct communities identified with the fast greedy modularity optimization

algorithm. Three such communities were identified. Community 1 included proteins associated with the more traditional AD biomarkers, such as ptau, ptau/Aβ42, and Aβ42/Aβ40. Community 2 centered around neurogranin, including many proteins uniquely associated with neurogranin. Community 3 included proteins associated with the remaining biomarkers of alpha-synuclein, YKL-40, NFL, and sTREM2. No proteins were significantly associated with IL-6. Many proteins were shared in the center of the network, particularly among neurogranin, ptau, and alpha-synuclein.

Replication of top pathway results in the Knight ADRC

The glucose metabolism pathway (REACTOME ID R-HSA-70326) was significantly enriched among all three of the amyloid and tau measures, so the results of the 9 proteins from this pathway that were significantly associated with one of the amyloid or tau biomarkers were chosen for replication in the Knight ADRC. These proteins included MDH1, ALDOA, PGK1, TPI1, PGAM1, PKM, GOT1, ALDOC, and ENO1. Of these 9 proteins, 4 of them (MDH1, ALDOA, PGK1, TPI1) were replicated in the Knight ADRC CSF case-control analysis, with MDH1 ($P = 1.07 \times 10^{-19}$) and ALDOA ($P = 7.43 \times 10^{-14}$) both meeting the Knight ADRC's Bonferroni-corrected threshold for significance (Table 3). The directions of effect were concordant except for PGK1. There was no significant association of these proteins with AD case status in the brain or plasma samples. Among the other AD-related outcomes in the Knight ADRC data, MDH1 and ALDOA were both statistically significantly associated with the CSF ptau/Aβ ratio ($P = 8.08 \times 10^{-28}$ and $P = 1.76 \times 10^{-27}$, respectively), matching what was seen for these proteins with this biomarker in the Wisconsin cohorts both in terms of strength of association ($P = 3.21 \times 10^{-10}$ and $P = 7.34 \times 10^{-8}$, respectively) and direction of effect (positive) (Supplementary Table 10).

Table 3. Replication of implicated glucose metabolism-related proteins in the Knight ADRC

	Discovery (WI ADRC)						Replication (Knight ADRC)					
	CSF AB42/40		CSF ptau/AB42		CSF ptau		CSF case-control		Brain case-control		Plasma case-control	
Protein	Beta	P value	Beta	P value	Beta	P value	Beta	P value	Beta	P value	Beta	P value
MDH1	-0.0088	1.44E-09	0.016	3.21E-10	6.78	3.29E-14	0.041	1.07E-19	.	.	.	.
ALDOA	-0.0070	1.75E-06	0.014	7.34E-08	6.59	1.26E-13	0.031	7.43E-14	.	.	.	.
PGK1	-0.0068	5.67E-06	0.013	5.08E-07	4.13	2.05E-05	-0.019	2.80E-05	-0.007	4.42E-01	.	.
TPI1	-0.0071	9.01E-07	0.012	4.51E-06	4.43	1.71E-06	0.018	9.34E-05	-0.004	7.50E-01	-0.018	2.15E-02
PGAM1	-0.0081	4.35E-08	0.017	4.72E-11	5.38	1.16E-08	0.000	3.17E-01	.	.	0.006	8.38E-01
ALDOC	-0.0048	1.44E-03	0.008	2.25E-03	4.42	2.53E-06	.	.	.	.	.	.
ENO1	-0.0068	4.72E-06	0.011	3.86E-05	2.86	3.12E-03	.	.	.	.	.	.
GOT1	-0.0072	5.67E-07	0.014	2.25E-08	6.58	5.88E-14	.	.	0.010	4.02E-01	0.005	4.99E-01
PKM	-0.0066	6.93E-06	0.014	5.75E-08	6.65	5.14E-14	.	.	0.019	3.43E-01	.	.

A "." indicates a protein that was not present in the QC'd Knight ADRC data set.

Secondary analysis of insulin-related proteins

Of the insulin-related proteins of interest, only IGF-1 and AKT1 were identified in the proteomics workflow. AKT1 was only quantified in one subject and was thus not suitable for further analysis, but IGF-1 was quantified in 82 samples (59.9%) and analyzed further using only the non-imputed measurements. A trend in missing values by AT category was noted: 44.6% of samples were missing IGF-1 in A-T-, 41.0% in A+T-, and 33.3% in A+T+. The ANCOVA analysis of IGF-1 did not show a statistically significant difference of the protein across AT categories (P = 0.170), though the distribution of the protein appeared to increase with amyloid positivity (Supplementary Figure 13). The association analysis between IGF-1 and the CSF biomarkers revealed a nominally significant negative association with Aβ42/Aβ40 (P = 0.011) and positive association with ptau/Aβ42 (P = 0.009) (Supplementary Table 11, Supplementary Figure 14).

Multiomic prediction models for amyloid and tau

The results of the multiomic amyloid and tau prediction models revealed a consistent pattern where the CSF proteome outperformed the other omic data sets in predicting positivity based on the core biomarkers of Aβ42/Aβ40, ptau/Aβ42, and ptau (Figure 4, Supplementary Table 12). The predictive model based on the CSF proteome (number of predictors selected ranged from 37-97) achieved a high AUC across for all three biomarkers (Aβ42/Aβ40 AUC = 0.924, ptau/Aβ42 AUC = 0.917, and ptau AUC = 0.891), performing slightly better than even the integrative model. For Aβ42/Aβ40 and ptau positivity, the sensitivity and specificity values further demonstrated the relative superiority of the proteomics model. For Aβ42/Aβ40 positivity, the sensitivity of the CSF proteome model was 0.947 compared to much lower values (0.316–0.579) from the other models. Similarly, for ptau positivity, the specificity of the CSF proteome model (0.571) was much higher than the other models (0.000–0.143). In all cases, the genome-based model performed poorly (AUCs ranged from 0.500–0.628) (Figure 4a). The 2D histograms showing the performance of the CSF proteome relative to the raw biomarker values highlighted the effective classification by the proteomic models with effective delineation between positive and negative amyloid statuses (Figure 4b–d).

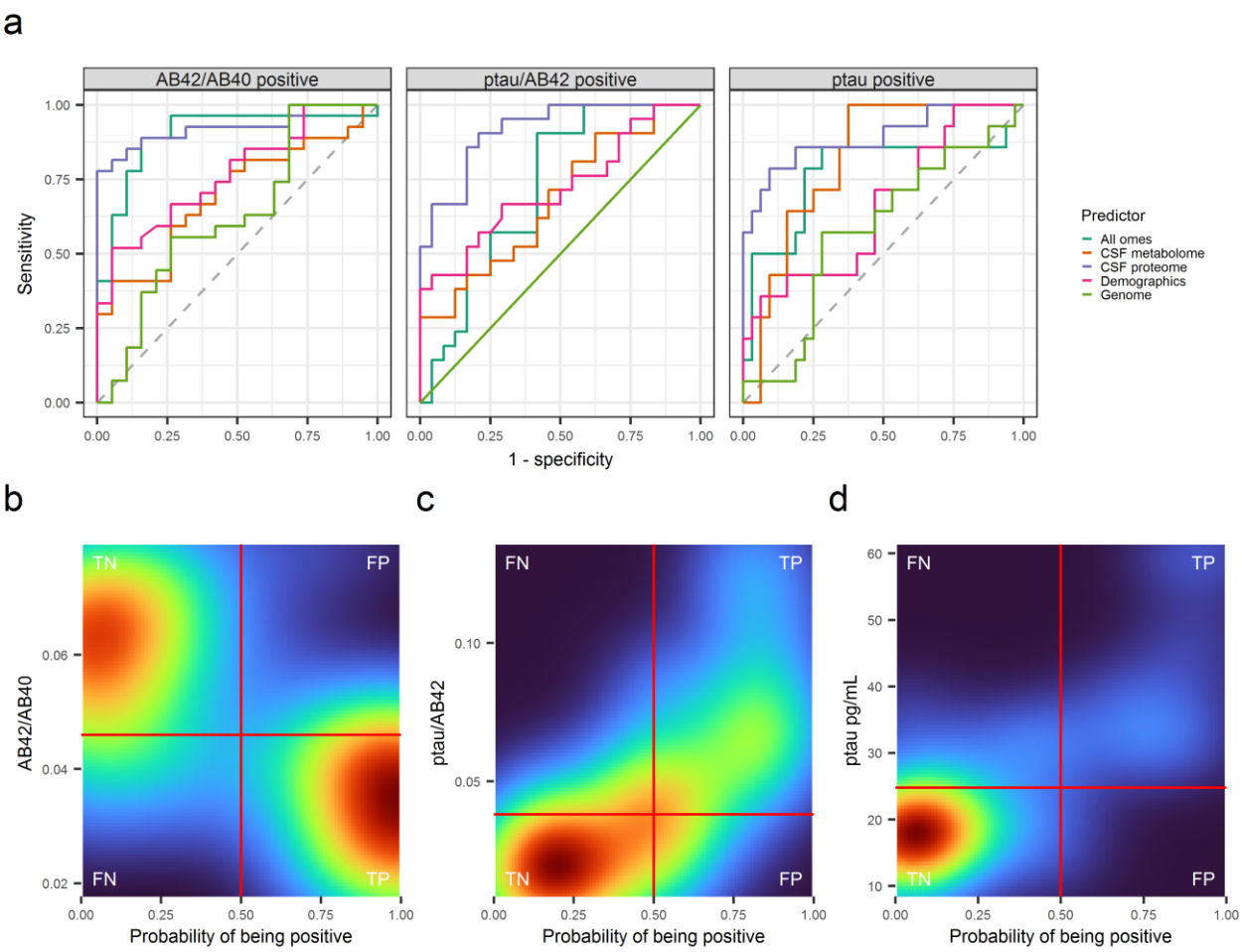

Figure 4: Multiomic predictive model performance for amyloid and tau

The ability of different omic data sets to predict the key AD CSF biomarkers for amyloid and tau is summarized. **a)** The receiver operator characteristic (ROC) curves for Aβ42/Aβ40, ptau/Aβ42, and ptau are shown for the held-out testing data set predictions. Each omic prediction set is plotted with a different line. **b-d)** A 2D density plot summarizes the actual values and predicted probability of positivity for each CSF biomarker (Aβ42/Aβ40, ptau/Aβ42, and ptau, respectively) from the proteomic predictor models applied to the held-out testing data set. The text labels in the corners refer to the prediction categories: TN = true negative, FP = false positive, TP = true positive, and FN = false negative. The vertical red line indicates the threshold for a hard classification of biomarker positivity from the prediction models. The horizontal red line indicates the binary threshold for the CSF biomarker determined from previous work (see Methods).

Discussion

Using the pipeline we developed from the pilot study, we successfully quantified 915 proteins that passed our QC metrics across 137 participant CSF samples. Our participant population covered the spectrum of amyloid and tau positivity in a largely preclinical cohort that, in combination with the rich set of standard and novel CSF biomarkers of neuroinflammation and neurodegeneration, gave us a window into the proteomic changes occurring as amyloid and tau accumulate. Among the CSF proteins we quantified, significantly enriched functional annotation included extracellular matrix, axonogenesis,

humoral immune system, complement system, and platelet pathways (Figure 1b), which was similar to previous work despite the difference in cohort (here, AD-focused, compared to typical or healthy CSF samples)[71–76]. Also similar to previous work was our finding that there was no discernible difference by sex among the top 2 PCs of our CSF proteome (Supplementary Figure 6), echoing a previous study where unsupervised hierarchical clustering failed to distinguish samples by sex[73], though we note that such results do not preclude sex differences in individual AD-related proteins. When our CSF proteome was examined for significantly enriched disease-related proteins, we identified enriched clusters of proteins related to AD and tauopathy. This finding underscored results from the Zhang et al. study of 2,513 proteins from 14 CSF samples that showed an enrichment of proteins related to neurological disease in the CSF[74], though we also identified enriched clusters of proteins related to cardiovascular disease among the CSF proteome (Figure 1c), which could potentially reflect differences in the studied population.

We identified a total of 61 AT-associated CSF proteins after multiple testing correction (Supplementary Table 6), with 43 of these proteins having been previously implicated in AD. The protein here with the strongest association with AT category was the SMOC1 protein, which increased in the CSF with increasing pathology along the AT categories. The increase of SMOC1 with increasing pathology or disease severity has been noted before[77–79], and the protein has also been found to partly colocalize with amyloid plaques[80]. Fatty acid binding protein 3 (FABP3), which was found to increase across AT groups in this study, is another example. FABP3 has been commonly associated with AD across CSF proteomics studies[79,81,82]. However, several proteins commonly associated with AD were not significantly associated with AT group here, including APOE, clusterin, and secretogranin[82]. Although the reason for this lack of association is unclear, it could be due to the present cohort being largely preclinical. On the other hand, we identified several novel protein associations, including ectonucleotide pyrophosphatase/phosphodiesterase family member 5 (ENPP5*,* noted to possibly be involved in neuronal cell communications[83]), heparin cofactor2 (SERPIND1, previously associated with multiple sclerosis[84]), extracellular matrix protein 2 (ECM2, jointly associated with iron along with APOE[85]), and glycoprotein endo-alpha-1,2-mannosidase-like protein (MANEAL, where variants in both *MANEAL* and *OSTM1* have been observed in connection with a neurodegenerative disorder[86]). The enriched KEGG pathways among the AT-associated proteins were carbon metabolism (hsa01200), biosynthesis of amino acids (hsa01230), and glycolysis/gluconeogenesis (hsa00010), and all three of these enriched pathways were enriched whether the full human proteome or only our 915 CSF proteins were used as the background distribution. The enriched GO terms varied but included multiple pathways related to the regulation of peptidases, some of which have been shown to be related to AD and amyloid metabolism[87,88]. The differential expression of proteins related to metabolism has been seen in other CSF proteomics cohorts[24,78,81,89], but although protein peptidases are known to be present in a large number in the CSF proteome[72], there is comparatively less proteomics work highlighting the role of peptidases beyond a study by Whelan et al. that found up-regulated endopeptidases in AD patients relative to controls[23].

One of the major trends in AD research has been the movement toward a biomarker-based definition of AD[11]. Recent CSF proteomics work in AD has explored the relationship between CSF protein levels and various AD biomarkers, especially measures of amyloid and tau pathology[79,90]. We replicated numerous

previously reported protein associations with CSF levels of ptau (e.g., SMOC1, BASP1, and GAP43)[79,90], but failed to replicate any of the proteins significantly associated with CSF amyloid, though we note that our analysis used Aβ42/Aβ40 as the outcome rather than Aβ42 since Aβ42/Aβ40 is considered to be a better biomarker for AD than Aβ42 alone[91].

Unique to this study was the inclusion of a more comprehensive set of CSF biomarkers relevant to AD, neurodegeneration, and neuroinflammation. Each biomarker had its own unique set of significantly associated proteins, but the network community analysis revealed that the affected proteins for the biomarkers tended to separate between the traditional AD biomarkers of amyloid and tau, neurogranin, and the remaining biomarkers. Notably, IL-6—a cytokine that has been explored as a possible marker of AD-related neuroinflammation—was alone in lacking statistically significant protein associations, supporting a meta-analysis that found no significant difference in peripheral IL-6 levels between AD cases and controls[92]. Collectively, our results identify broadly different protein networks associated with the classical AD biomarkers compared to more general markers of neurodegeneration and neuroinflammation.

The pathway enrichment analysis of biomarker-associated proteins underscored the differences among these protein groups (Supplementary Table 9). Among the amyloid and tau biomarkers, the associated proteins shared a common theme of enrichment of glucose metabolism pathways, including two proteins (MDH1 and ALDOA) that showed evidence of association with AD diagnosis in the Knight ADRC replication data set. MDH1 levels have been associated with AD in the past, although the direction of effect has not been consistent[93–95]. Here, MDH1 was observed to increase with amyloid and tau pathology. ALDOA has been previously positively associated with CSF tau levels[79], as was the case here, and also associated with the *APOE* ε4 allele[96]. Glucose metabolism, which is the major source of energy for the brain, has long been known to show signs of dysfunction in AD even before the emergence of symptoms[97–101]. Our findings here, where 74% of participants were cognitively unimpaired, support the emergence of glucose metabolic dysregulation presymptomatically as amyloid and tau begin to show alterations. Further underscoring potential abnormalities in energy metabolism as AD develops is the observation that 18 of the 61 proteins associated with AT here have been previously connected with insulin resistance, including pyruvate kinase (*PKM*), alpha-enolase (*ENO1*), and triosephosphate isomerase (*TPI*), which in a previous study were found to be elevated in participants with type 1 diabetes[102]. These proteins were statistically significantly associated with one or more of the three amyloid and tau biomarkers. The secondary analysis of IGF-1 further suggests possible abnormalities in insulin signaling. IGF-1, which can bind insulin receptors[103], has been previously implicated in AD in studies that found decreased *IGF1* expression in brain tissue and evidence of IGF-1 resistance[104–106]. Here, we found increased levels of CSF IGF-1 with increasing amyloid pathology, though the lower sample size in the IGF-1 analysis warrants caution in the interpretation. Importantly, the enrichment of glucose metabolic pathways among the amyloid and tau-related proteins in this study was not seen among the other two communities of biomarkers, which instead tended to be enriched for extracellular matrix, cell junctions and adhesion, and other pathways. Taken together, these results provide evidence for a preclinical dysregulation of the glucose metabolic proteome that is more specific to amyloid and tau than to biomarkers of neuroinflammation or neurodegeneration.

A potential consequence of altered glucose metabolism is a disruption in autophagy and proteostasis[97], which has been observed in AD[107–110]. Although many proteasome-relevant proteins were not quantified in our CSF data set, several heat shock proteins, which often help regulate protein folding and degradation, were, including the chaperone protein HSP90AA1. We observed a statistically significant negative association of HSP90AA1 levels with Aβ42/Aβ40 and a significant positive association with ptau/Aβ42, consistent with worse pathology. Interestingly, cathepsin D (CTSD), a lysosomal protease that has been targeted by drugs seeking to modulate autophagy in AD mouse models[110,111], was statistically significantly associated with sTREM2, alpha-synuclein, neurogranin, NFL, and YKL-40, but not with amyloid or tau (Supplementary Table 8). Moreover, we observed an enrichment of peptidase-related pathways among proteins associated with AT group, ptau, and neurogranin (Supplementary Table 7, Supplementary Table 9). Further potential evidence implicating proteostasis was the substantial enriched association signal across the proteome, with 54.2% (496/915) of CSF proteins being nominally associated with AT category and substantial deviation from the expected null distribution across the proteome (Figure 2b). These nominally associated proteins included many proteins known to form intracellular or extracellular deposits in disease[108], including APP, TTR, B2M, APOA1, APOA2, APOA4, APOC2, APOC3, LYZ, CST3, SOD1, IGHG1, IGHG2, IGHG3, and HBB (Supplementary Table 6). Although individual protein associations and pathway enrichment have tended to be the focus of previous work, several studies have reported similar widespread enrichment among the proteome, including a case-control study of AD diagnosis with 487 of 1,968 (24.7%) of proteins nominally associated with AD diagnosis[112] and a study of protein correlations with CSF total tau, ptau, and Aβ42 where 63 out of 106 proteins (59.4%) were associated with at least one of the biomarkers[90]. Indeed, one core feature of AD proteomics work historically has been the identification of numerous AD-associated proteins that nonetheless do not replicate upon further study[81,82]. There are technical and study design reasons why such protein associations may not replicate easily[81], and unmeasured confounding could certainly explain some of the signal enrichment, but another potential reason could be dysregulated proteostasis which leads to greater variation in the levels of many proteins. Finally, the relative importance of the CSF proteome over the CSF metabolome, genome, and demographic information in predicting relevant AD biomarkers, seen not just here but in multiple studies where high prediction performance was achieved from the CSF proteome[23,75,113], further supports the unique relevance of the proteome in AD pathophysiology.

A few limitations deserve mention. First, the sample size and studied population were both limited. Though our sample size of 137 was comparable to other CSF proteomics work in AD, analysis in a larger sample would provide more precision. Our study was also limited to individuals of European ancestry, which limits its generalizability to broader populations. Another limitation of our study was the lack of the N (neurodegeneration) category in our main analyses. Expanding the categories from AT only to all relevant combinations of ATN would allow for more nuanced analysis. In a similar vein, having comparison groups for other non-AD causes of dementia would allow for better triangulation of AD-specific proteomic changes. Finally, instead of using the entire omic data sets (or outcome-blind predictor reductions, as was the case with the genetic data set), additional filtering steps that narrowed down to predictors more likely to be associated with AD would provide even better prediction for the amyloid and tau measures.

Nevertheless, our study provides a thorough investigation of the CSF proteome and its relationship to AD, AD biomarkers, and other omics. We replicated previous work in the general CSF proteome that showed an enrichment of extracellular and immune system process proteins. We identified numerous proteins associated with AD and CSF biomarkers of neurodegeneration and neuroinflammation, including several novel protein associations and distinct associated proteomes between amyloid and tau measures and the other biomarkers. Furthermore, we demonstrated that the CSF proteome associated with amyloid and tau was enriched for glucose metabolic pathways in contrast to the other biomarkers whose associated proteins were enriched for more extracellular and structural pathways. We then highlighted the importance of the CSF proteome as a whole in predicting amyloid and tau, above and beyond the predictive capabilities of CSF metabolomics, genomics, and demographic information. In total, this study highlights the importance and biomarker-specific associations of the proteome in AD, with potential implications of altered glucose metabolism.

Acknowledgments

We would like to thank WRAP and ADRC participants and the Wisconsin Alzheimer's Institute (WAI) and ADRC staff for their contributions to the WRAP and ADRC studies. Without their efforts this research would not be possible. This research is supported by National Institutes of Health (NIH) grants R01AG27161 (Wisconsin Registry for Alzheimer Prevention: Biomarkers of Preclinical AD), R01AG054047 (Genomic and Metabolomic Data Integration in a Longitudinal Cohort at Risk for Alzheimer's Disease), P41GM108538 (National Center for Quantitative Biology of Complex Systems), R01AG037639 (White Matter Degeneration: Biomarkers in Preclinical Alzheimer's Disease), R01AG021155 (The Longitudinal Course of Imaging Biomarkers in People at Risk of AD), and P50AG033514 and P30AG062715 (Wisconsin Alzheimer's Disease Research Center Grant), the Clinical and Translational Science Award (CTSA) program through the NIH National Center for Advancing Translational Sciences (NCATS) grant UL1TR000427, and the University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation. Computational resources were supported by a core grant to the Center for Demography and Ecology at the University of Wisconsin-Madison (P2CHD047873). We also acknowledge use of the facilities of the Center for Demography of Health and Aging at the University of Wisconsin-Madison, funded by NIA Center grant P30AG017266. Author DJP was supported by NLM training grants to the Bio-Data Science Training Program (T32LM012413) and the Interdisciplinary Training Program in Cardiovascular and Pulmonary Biostatistics (5T32HL83806). Author YKD was supported by a training grant from the National Institute on Aging (T32AG000213). Author GEE was supported by an Alzheimer's Association Research Fellowship (AARF-643973). Author HZ is a Wallenberg Scholar supported by grants from the Swedish Research Council (#2018-02532), the European Research Council (#681712), Swedish State Support for Clinical Research (#ALFGBG-720931), the Alzheimer Drug Discovery Foundation (ADDF), USA (#201809-2016862), the AD Strategic Fund and the Alzheimer's Association (#ADSF-21-831376-C, #ADSF-21-831381-C and #ADSF-21-831377-C), the Olav Thon Foundation, the Erling-Persson Family Foundation, Stiftelsen för Gamla Tjänarinnor, Hjärnfonden, Sweden (#FO2019-0228), the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No

860197 (MIRIADE), and the UK Dementia Research Institute at UCL. Author KB was supported by the Swedish Research Council (#2017-00915), the Alzheimer Drug Discovery Foundation (ADDF), USA (#RDAPB-201809-2016615), the Swedish Alzheimer Foundation (#AF-742881), Hjärnfonden, Sweden (#FO2017-0243), the Swedish state under the agreement between the Swedish government and the County Councils, the ALF-agreement (#ALFGBG-715986), the European Union Joint Program for Neurodegenerative Disorders (JPND2019-466-236), and the NIH, USA, (grant #1R01AG068398-01). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Author CC receives support from the National Institutes of Health (R01AG044546, R01AG064877, RF1AG053303, R01AG058501, U01AG058922, RF1AG058501, R01AG064614), and the Chuck Zuckerberg Initiative (CZI). The recruitment and clinical characterization of research participants at Washington University were supported by NIH P30AG066444, and P01AG003991. This work was supported by access to equipment made possible by the Hope Center for Neurological Disorders, the NeuroGenomics and Informatics Center (NGI: https://neurogenomics.wustl.edu/) and the Departments of Neurology and Psychiatry at Washington University School of Medicine.

ELECSYS, COBAS and COBAS E are trademarks of Roche. The Roche NeuroToolKit robust prototype assays are for investigational purposes only and are not approved for clinical use.

We thank the University of Wisconsin Madison Biotechnology Center Gene Expression Center for providing Illumina Infinium genotyping services.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Conflicts of interest

Author CC receives research support from Biogen, EISAI, Alector, GSK and Parabon; these funders of the study had no role in the collection, analysis, or interpretation of data; in the writing of the report; or in the decision to submit the paper for publication. Author CC is a member of the advisory board of Vivid Genomics, Halia Therapeutics and ADx Healthcare. Author HZ has served at scientific advisory boards and/or as a consultant for Alector, Eisai, Denali, Roche Diagnostics, Wave, Samumed, Siemens Healthineers, Pinteon Therapeutics, Nervgen, AZTherapies, CogRx and Red Abbey Labs, has given lectures in symposia sponsored by Cellectricon, Fujirebio, Alzecure and Biogen, and is a co-founder of Brain Biomarker Solutions in Gothenburg AB (BBS), which is a part of the GU Ventures Incubator Program. Author KB has served as a consultant, at advisory boards, or at data monitoring committees for Abcam, Axon, Biogen, JOMDD/Shimadzu. Julius Clinical, Lilly, MagQu, Novartis, Roche Diagnostics, and Siemens Healthineers, and is a co-founder of Brain Biomarker Solutions in Gothenburg AB (BBS), which is a part of the GU Ventures Incubator Program. Author GK is a full-time employee of Roche Diagnostics GmbH. Author IS is a full-time employee and shareholder of Roche Diagnostics International Ltd. Author AB is a full-time employee and shareholder of Roche Diagnostics GmbH. Author SCJ serves as a consultant to Roche Diagnostics and receives research funding from Cerveau Technologies.

Other authors have no competing interests to declare.

Data availability

The data sets generated and analyzed in this study from the Wisconsin ADRC may be requested at https://www.adrc.wisc.edu/apply-resources. The Knight-ADRC proteomic data is available at NIAGADS: NG00102 collection and can be interactively explored at http://ngi.pub:3838/ONTIME_Proteomics/.

Figures

Figure 1: CSF proteomics descriptive analyses

Figure 2: Proteins associated with AT group

Figure 3: Network analysis of protein-biomarker associations

Figure 4: Multiomic predictive model performance for amyloid and tau

Supplementary Figure 1: Pilot study power analysis

Supplementary Figure 2: Protein missingness overall

Supplementary Figure 3: Protein missingness by AT group

Supplementary Figure 4: Protein imputation examples

Supplementary Figure 5: CSF proteomics PCA scree plot

Supplementary Figure 6: CSF proteomics PC plots

Supplementary Figure 7: Clustering BIC values

Supplementary Figure 8: Overlap of ANCOVA and *APOE*-controlled results

Supplementary Figure 9: Overlap of ANCOVA and logistic regression results

Supplementary Figure 10: Significantly associated proteins with A+T+ vs A-T-

Supplementary Figure 11: Protein-biomarker association Q-Q plot

Supplementary Figure 12: Overlap in the proteins associated among the biomarkers

Supplementary Figure 13: Distribution of IGF-1 by AT category

Supplementary Figure 14: Relationships between IGF-1 and CSF biomarkers

Tables

Table 1: Summary of sample demographics and CSF biomarkers

Table 2: Top 3 significantly associated proteins for each biomarker

Table 3: Replication of implicated glucose metabolism-related proteins in the Knight ADRC

Supplementary Table 1: Protein quality control steps

Supplementary Table 2: Protein information

Supplementary Table 3: Nominally significant pairwise protein correlations

Supplementary Table 4: Significantly enriched pathways among the CSF proteome compared to the human proteome

Supplementary Table 5: Significantly enriched pathways across CSF proteome clusters

Supplementary Table 6: Protein-AT association test results

Supplementary Table 7: Significantly enriched pathways among proteins associated with AT group

Supplementary Table 8: Protein-biomarker association test results

Supplementary Table 9: Significantly enriched pathways among proteins significantly associated with biomarkers

Supplementary Table 10: Associations of glucose metabolism-related proteins with AD outcomes in the Knight ADRC

Supplementary Table 11: IGF-1-biomarker association test results

Supplementary Table 12: Summary of mean multiomic biomarker prediction model performance

References

1. Alzheimer's Disease International. *World Alzheimer Report 2019: Attitudes to dementia*. 160 (2019).
2. 2019 Alzheimer's disease facts and figures. *Alzheimers Dement.* **15**, 321–387 (2019).
3. Kumar, A., Singh, A. & Ekavali. A review on Alzheimer's disease pathophysiology and its management: an update. *Pharmacol. Rep.* **67**, 195–203 (2015).
4. Vinters, H. V. Emerging Concepts in Alzheimer's Disease. *Annu. Rev. Pathol. Mech. Dis.* **10**, 291–319 (2015).

5. Blennow, K., Hampel, H., Weiner, M. & Zetterberg, H. Cerebrospinal fluid and plasma biomarkers in Alzheimer disease. *Nat. Rev. Neurol.* **6**, 131–144 (2010).

6. McKhann, G. *et al.* Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* **34**, 939–944 (1984).

7. Jack, C. R. *et al.* Introduction to Revised Criteria for the Diagnosis of Alzheimer's Disease: National Institute on Aging and the Alzheimer Association Workgroups. *Alzheimers Dement.* **7**, 257–262 (2011).

8. Sperling, R. A. *et al.* Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* **7**, 280–292 (2011).

9. Albert, M. S. *et al.* The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* **7**, 270–279 (2011).

10. McKhann, G. M. *et al.* The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* **7**, 263–269 (2011).

11. Jack, C. R. *et al.* NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimers Dement.* **14**, 535–562 (2018).

12. Karczewski, K. J. & Snyder, M. P. Integrative omics for health and disease. *Nat. Rev. Genet.* **19**, 299–310 (2018).

13. Neuner, S. M., Tcw, J. & Goate, A. M. Genetic architecture of Alzheimer's disease. *Neurobiol. Dis.* **143**, 104976 (2020).

14. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414 (2019).

15. Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).

16. Bellenguez, C. *et al.* New insights on the genetic etiology of Alzheimer's and related dementia. *medRxiv* 2020.10.01.20200659 (2020) doi:10.1101/2020.10.01.20200659.

17. Kaddurah-Daouk, R. *et al.* Metabolomic changes in autopsy-confirmed Alzheimer's disease. *Alzheimers Dement.* **7**, 309–317 (2011).

18. Trushina, E., Dutta, T., Persson, X.-M. T., Mielke, M. M. & Petersen, R. C. Identification of Altered Metabolic Pathways in Plasma and CSF in Mild Cognitive Impairment and Alzheimer's Disease Using Metabolomics. *PLOS ONE* **8**, e63644 (2013).

19. Khoonsari, P. E. *et al.* Analysis of the Cerebrospinal Fluid Proteome in Alzheimer's Disease. *PLOS ONE* **11**, e0150672 (2016).

20. Paterson, R. W. *et al.* A targeted proteomic multiplex CSF assay identifies increased malate dehydrogenase and other neurodegenerative biomarkers in individuals with Alzheimer's disease pathology. *Transl. Psychiatry* **6**, e952 (2016).

21. Spellman, D. S. *et al.* Development and Evaluation of a Multiplexed Mass Spectrometry-Based Assay for Measuring Candidate Peptide Biomarkers in Alzheimer's Disease Neuroimaging Initiative (ADNI) CSF. *Proteomics Clin. Appl.* **9**, 715–731 (2015).

22. Wildsmith, K. R. *et al.* Identification of longitudinally dynamic biomarkers in Alzheimer's disease cerebrospinal fluid by targeted proteomics. *Mol. Neurodegener.* **9**, 22 (2014).

23. Whelan, C. D. *et al.* Multiplex proteomics identifies novel CSF and plasma biomarkers of early Alzheimer's disease. *Acta Neuropathol. Commun.* **7**, 169 (2019).

24. Johnson, E. C. B. *et al.* Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. *Nat. Med.* **26**, 769–780 (2020).

25. McKetney, J. *et al.* Pilot proteomic analysis of cerebrospinal fluid in Alzheimer's disease. *Proteomics Clin. Appl.* e2000072 (2021) doi:10.1002/prca.202000072.

26. Johnson, S. C. *et al.* The Wisconsin Registry for Alzheimer's Prevention: A review of findings and current directions. *Alzheimers Dement. Diagn. Assess. Dis. Monit.* **10**, 130–142 (2017).

27. Melah, K. E. *et al.* CSF markers of Alzheimer's pathology and microglial activation are associated with altered white matter microstructure in asymptomatic adults at risk for Alzheimer's disease. *J. Alzheimers Dis.* **50**, 873–886 (2016).

28. von Elm, E. *et al.* The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J. Clin. Epidemiol.* **61**, 344–349 (2008).

29. Van Hulle, C. A. *et al.* An examination of a novel multipanel of CSF biomarkers in the Alzheimer's disease clinical and pathological continuum. *Alzheimers Dement.* **17**, 431–445 (2021).

30. Komsta, L. & Novomestky, F. *moments: Moments, cumulants, skewness, kurtosis and related tests*. (2015).

31. *MATLAB*. (The MathWorks Inc.).

32. Hilden, J. The Area under the ROC Curve and Its Competitors. *Med. Decis. Making* **11**, 95–101 (1991).

33. Darst, B. F. *et al.* Pathway-Specific Polygenic Risk Scores as Predictors of Amyloid-β Deposition and Cognitive Function in a Sample at Increased Risk for Alzheimer's Disease. *J. Alzheimers Dis.* **55**, 473–484 (2017).

34. Darst, B. F., Lu, Q., Johnson, S. C. & Engelman, C. D. Integrated analysis of genomics, longitudinal metabolomics, and Alzheimer's risk factors among 1,111 cohort participants. *Genet. Epidemiol.* **43**, 657–674 (2019).

35. Panyard, D. J. *et al.* Cerebrospinal fluid metabolomics identifies 19 brain-related phenotype associations. *Commun. Biol.* **4**, 1–11 (2021).

36. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

37. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

38. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).

39. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

40. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

41. Champely, S. *pwr: Basic Functions for Power Analysis*. (2020).

42. Shishkova, E., Hebert, A. S., Westphall, M. S. & Coon, J. J. Ultra-High Pressure (>30,000 psi) Packing of Capillary Columns Enhancing Depth of Shotgun Proteomic Analyses. *Anal. Chem.* **90**, 11503–11508 (2018).

43. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).

44. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).

45. Carlson, M. *org.Hs.eg.db: Genome wide annotation for Human*. (2020).

46. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

47. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

48. Tweedie, S. *et al.* Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Res.* **49**, D939–D946 (2021).

49. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

50. Kassambara, A. & Mundt, F. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. (2020).

51. The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2020).

52. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27 (2000).

53. Schriml, L. M. *et al.* Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.* **47**, D955–D962 (2019).

54. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics J. Integr. Biol.* **16**, 284–287 (2012).

55. Yu, G., Wang, L.-G., Yan, G.-R. & He, Q.-Y. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* **31**, 608–609 (2015).

56. Yu, G. & He, Q.-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst.* **12**, 477–479 (2016).

57. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J.* **8**, 289–317 (2016).

58. Gao, C.-H. *ggVennDiagram: A 'ggplot2' Implement of Venn Diagram*. (2019).

59. Pedersen, T. L. *tidygraph: A Tidy API for Graph Manipulation*.

60. Clauset, A., Newman, M. E. J. & Moore, C. Finding community structure in very large networks. *Phys. Rev. E* **70**, 066111 (2004).

61. Csardi, G. & Nepusz, T. *The igraph software package for complex network research*. (2006).

62. Gehlenborg, N. *UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets*. (2019).

63. Morris, J. C. The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology* **43**, 2412–2414 (1993).

64. Hu, W. T. *et al.* Plasma multianalyte profiling in mild cognitive impairment and Alzheimer disease. *Neurology* **79**, 897–905 (2012).

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

66. Yang, C. *et al.* Genomic atlas of the proteome from brain, CSF and plasma prioritizes proteins implicated in neurological disorders. *Nat. Neurosci.* 1–11 (2021) doi:10.1038/s41593-021-00886-6.

67. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).

68. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 301–320 (2005).

69. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).

70. Kuhn, M. & Wickham, H. *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles*. (2020).

71. Pan, S. *et al.* A combined dataset of human cerebrospinal fluid proteins identified by multi-dimensional chromatography and tandem mass spectrometry. *PROTEOMICS* **7**, 469–473 (2007).

72. Zougman, A. *et al.* Integrated Analysis of the Cerebrospinal Fluid Peptidome and Proteome. *J. Proteome Res.* **7**, 386–399 (2008).

73. Schutzer, S. E. *et al.* Establishing the Proteome of Normal Human Cerebrospinal Fluid. *PLoS ONE* **5**, (2010).

74. Zhang, Y. *et al.* A comprehensive map and functional annotation of the normal human cerebrospinal fluid proteome. *J. Proteomics* **119**, 90–99 (2015).

75. Begcevic, I., Brinc, D., Drabovich, A. P., Batruch, I. & Diamandis, E. P. Identification of brain-enriched proteins in the cerebrospinal fluid proteome by LC-MS/MS profiling and mining of the Human Protein Atlas. *Clin. Proteomics* **13**, (2016).

76. Macron, C. *et al.* Exploration of human cerebrospinal fluid: A large proteome dataset revealed by trapped ion mobility time-of-flight mass spectrometry. *Data Brief* **31**, (2020).

77. Wang, H. *et al.* Integrated analysis of ultra-deep proteomes in cortex, cerebrospinal fluid and serum reveals a mitochondrial signature in Alzheimer's disease. *Mol. Neurodegener.* **15**, 43 (2020).

78. Higginbotham, L. *et al.* Integrated proteomics reveals brain-based cerebrospinal fluid biomarkers in asymptomatic and symptomatic Alzheimer's disease. *Sci. Adv.* **6**, eaaz9360 (2020).

79. Dayon, L. *et al.* Alzheimer disease pathology and the cerebrospinal fluid proteome. *Alzheimers Res. Ther.* **10**, 66 (2018).

80. Bai, B. *et al.* Deep Multilayer Brain Proteomics Identifies Molecular Networks in Alzheimer's Disease Progression. *Neuron* **105**, 975-991.e7 (2020).

81. Korolainen, M. A., Nyman, T. A., Aittokallio, T. & Pirttilä, T. An update on clinical proteomics in Alzheimer's research. *J. Neurochem.* **112**, 1386–1414 (2010).

82. Portelius, E. *et al.* Proteomic studies of cerebrospinal fluid biomarkers of Alzheimer's disease: an update. *Expert Rev. Proteomics* **14**, 1007–1020 (2017).

83. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).

84. Comabella, M. *et al.* Cerebrospinal fluid chitinase 3-like 1 levels are associated with conversion to multiple sclerosis. *Brain* **133**, 1082–1093 (2010).

85. Britton, L. J. *et al.* Iron Inhibits the Secretion of Apolipoprotein E in Cultured Human Adipocytes. *Cell. Mol. Gastroenterol. Hepatol.* **6**, 215-217.e8 (2018).

86. Herebian, D. *et al.* Coexisting variants in OSTM1 and MANEAL cause a complex neurodegenerative disorder with NBIA-like brain abnormalities. *Eur. J. Hum. Genet.* **25**, 1092–1095 (2017).

87. Nalivaeva, N. N., Belyaev, N. D., Zhuravin, I. A. & Turner, A. J. The Alzheimer's Amyloid-Degrading Peptidase, Neprilysin: Can We Control It? *Int. J. Alzheimers Dis.* **2012**, e383796 (2012).

88. Goldhardt, O. *et al.* Kallikrein-related peptidases 6 and 10 are elevated in cerebrospinal fluid of patients with Alzheimer's disease and associated with CSF-TAU and FDG-PET. *Transl. Neurodegener.* **8**, 25 (2019).

89. Tijms, B. M. *et al.* Pathophysiological subtypes of Alzheimer's disease based on cerebrospinal fluid proteomics. *Brain J. Neurol.* **143**, 3776–3792 (2020).

90. Remnestål, J. *et al.* Association of CSF proteins with tau and amyloid β levels in asymptomatic 70-year-olds. *Alzheimers Res. Ther.* **13**, 54 (2021).

91. Hansson, O., Lehmann, S., Otto, M., Zetterberg, H. & Lewczuk, P. Advantages and disadvantages of the use of the CSF Amyloid β (Aβ) 42/40 ratio in the diagnosis of Alzheimer's Disease. *Alzheimers Res. Ther.* **11**, 34 (2019).

92. Ng, A. *et al.* IL-1β, IL-6, TNF- α and CRP in Elderly Patients with Depression or Alzheimer's disease: Systematic Review and Meta-Analysis. *Sci. Rep.* **8**, 12050 (2018).

93. Sonntag, K.-C. *et al.* Late-onset Alzheimer's disease is associated with inherent changes in bioenergetics profiles. *Sci. Rep.* **7**, 1–13 (2017).

94. Canchi, S. *et al.* Integrating Gene and Protein Expression Reveals Perturbed Functional Networks in Alzheimer's Disease. *Cell Rep.* **28**, 1103-1116.e4 (2019).

95. Ryu, W.-I. *et al.* Brain cells derived from Alzheimer's disease patients have multiple specific innate abnormalities in energy metabolism. *Mol. Psychiatry* 1–13 (2021) doi:10.1038/s41380-021-01068-3.

96. Berger, M. *et al.* APOE4 Copy Number-Dependent Proteomic Changes in the Cerebrospinal Fluid. *J. Alzheimers Dis.* **79**, 511–530 (2021).

97. Butterfield, D. A. & Halliwell, B. Oxidative stress, dysfunctional glucose metabolism and Alzheimer disease. *Nat. Rev. Neurosci.* **20**, 148–160 (2019).

98. Bosco, D., Fava, A., Plastino, M., Montalcini, T. & Pujia, A. Possible implications of insulin resistance and glucose metabolism in Alzheimer's disease pathogenesis. *J. Cell. Mol. Med.* **15**, 1807–1821 (2011).

99. Reiman, E. M. *et al.* Preclinical evidence of Alzheimer's disease in persons homozygous for the epsilon 4 allele for apolipoprotein E. *N. Engl. J. Med.* **334**, 752–758 (1996).

100. Reiman, E. M. *et al.* Functional brain abnormalities in young adults at genetic risk for late-onset Alzheimer's dementia. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 284–289 (2004).

101. Perkins, M. *et al.* Altered Energy Metabolism Pathways in the Posterior Cingulate in Young Adult Apolipoprotein E ε4 Carriers. *J. Alzheimers Dis. JAD* **53**, 95–106 (2016).

102. Iori, E. *et al.* Glycolytic enzyme expression and pyruvate kinase activity in cultured fibroblasts from type 1 diabetic patients with and without nephropathy. *Biochim. Biophys. Acta BBA - Mol. Basis Dis.* **1782**, 627–633 (2008).

103. Arnold, S. E. *et al.* Brain insulin resistance in type 2 diabetes and Alzheimer disease: concepts and conundrums. *Nat. Rev. Neurol.* **14**, 168–181 (2018).

104. Steen, E. *et al.* Impaired insulin and insulin-like growth factor expression and signaling mechanisms in Alzheimer's disease – is this type 3 diabetes? *J. Alzheimers Dis.* **7**, 63–80 (2005).

105. Rivera, E. J. *et al.* Insulin and insulin-like growth factor expression and function deteriorate with progression of Alzheimer's disease: Link to brain reductions in acetylcholine. *J. Alzheimers Dis.* **8**, 247–268 (2005).

106. Talbot, K. *et al.* Demonstrated brain insulin resistance in Alzheimer's disease patients is associated with IGF-1 resistance, IRS-1 dysregulation, and cognitive decline. *J. Clin. Invest.* **122**, 1316–1338 (2012).

107. Morawe, T., Hiebel, C., Kern, A. & Behl, C. Protein Homeostasis, Aging and Alzheimer's Disease. *Mol. Neurobiol.* **46**, 41–54 (2012).

108. Chiti, F. & Dobson, C. M. Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade. *Annu. Rev. Biochem.* **86**, 27–68 (2017).

109. Höhn, A., Tramutola, A. & Cascella, R. Proteostasis Failure in Neurodegenerative Diseases: Focus on Oxidative Stress. *Oxid. Med. Cell. Longev.* **2020**, (2020).

110. Festa, B. P., Barbosa, A. D., Rob, M. & Rubinsztein, D. C. The pleiotropic roles of autophagy in Alzheimer's disease: From pathophysiology to therapy. *Curr. Opin. Pharmacol.* **60**, 149–157 (2021).

111. Mahajan, A. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* **46**, 234–244 (2014).

112. Zhang, Q. *et al.* Integrated proteomics and network analysis identifies protein hubs and network alterations in Alzheimer's disease. *Acta Neuropathol. Commun.* **6**, (2018).

113. Khoonsari, P. E. *et al.* Improved Differential Diagnosis of Alzheimer's Disease by Integrating ELISA and Mass Spectrometry-Based Cerebrospinal Fluid Biomarkers. *J. Alzheimers Dis.* **67**, 639–651 (2019).