



Article

<https://doi.org/10.1038/s41591-025-03834-0>

The Global Neurodegeneration Proteomics Consortium: biomarker and drug target discovery for common neurodegenerative diseases and aging

Received: 24 April 2025

A list of authors and their affiliations appears at the end of the paper

Accepted: 16 June 2025

Published online: 15 July 2025

 Check for updates

More than 57 million people globally suffer from neurodegenerative diseases, a figure expected to double every 20 years. Despite this growing burden, there are currently no cures, and treatment options remain limited due to disease heterogeneity, prolonged preclinical and prodromal phases, poor understanding of disease mechanisms, and diagnostic challenges. Identifying novel biomarkers is crucial for improving early detection, prognosis, staging and subtyping of these conditions. High-dimensional molecular studies in biofluids ('omics') offer promise for scalable biomarker discovery, but challenges in assembling large, diverse datasets hinder progress. To address this, the Global Neurodegeneration Proteomics Consortium (GNPC)—a public–private partnership—established one of the world's largest harmonized proteomic datasets. It includes approximately 250 million unique protein measurements from multiple platforms from more than 35,000 biofluid samples (plasma, serum and cerebrospinal fluid) contributed by 23 partners, alongside associated clinical data spanning Alzheimer's disease (AD), Parkinson's disease (PD), frontotemporal dementia (FTD) and amyotrophic lateral sclerosis (ALS). This dataset is accessible to GNPC members via the Alzheimer's Disease Data Initiative's AD Workbench, a secure cloud-based environment, and will be available to the wider research community on 15 July 2025. Here we present summary analyses of the plasma proteome revealing disease-specific differential protein abundance and transdiagnostic proteomic signatures of clinical severity. Furthermore, we describe a robust plasma proteomic signature of *APOE ε4* carriership, reproducible across AD, PD, FTD and ALS, as well as distinct patterns of organ aging across these conditions. This work demonstrates the power of international collaboration, data sharing and open science to accelerate discovery in neurodegeneration research.

Neurodegenerative diseases, including Alzheimer's disease (AD), Parkinson's disease (PD), amyotrophic lateral sclerosis (ALS), fronto-temporal dementia (FTD) and other related conditions, affect more than 57 million people worldwide¹. Until recently, treatment options were limited to managing symptoms, but approvals for disease-modifying drugs for AD and genetic forms of ALS point to considerable progress^{2,3}. Eventually, it may be possible to provide patients with targeted treatments, possibly in combination, that can prevent, slow, stop or reverse the progression of their disease^{4–6}. However, several major obstacles have delayed the realization of this vision. First, many neurodegenerative conditions have an extended preclinical or prodromal period where diagnosis using available symptom-based assessments is either not possible or extremely difficult due to subtle manifestations that are not detectable by current clinical tools. Second, heterogeneity in the concordance between molecular pathology and clinical syndrome as well as common co-occurrence of multiple pathologies ('co-pathology') contribute to misdiagnosis in clinical settings. Third, additional variability exists in the rate and pattern of symptom progression within conditions, impeding efforts to accurately prognosticate disease course. These diagnostic and prognostic challenges ultimately hinder the efficacy of clinical trials and make successful treatment of patients with any approved disease-modifying therapy challenging.

Biomarkers have the potential to resolve some of these obstacles by enabling earlier diagnosis linked to pathological processes, providing methods to subtype diseases, predicting outcomes and ultimately guiding effective intervention^{7,8}. They may also improve clinical trial design through precision recruitment and serve as pharmacodynamic or surrogate endpoints in experimental medicine. Recognizing this potential, the field has seen rapid advances in imaging and fluid biomarker research, leading to their growing incorporation into clinical trials and regulatory frameworks. Fluid biomarkers, in particular, offer a real-time window into brain pathology and may help bridge the long-standing disconnect between neuropathology and clinical symptoms in living patients. Reflecting this progress, clinical guidelines have begun to integrate fluid biomarkers into routine diagnostic workflows⁹. To date, fluid biomarker development has been most successful for AD, where markers of amyloid and tau pathology are now widely used. However, there remains an urgent need for reliable biomarkers of other neurodegenerative pathologies, including α -synuclein, TDP-43 and non-AD tauopathies. In addition, biomarkers that reflect non-specific but disease-relevant biological processes, such as neuroinflammation, metabolic dysregulation and vascular dysfunction, are essential to fully characterize the pathophysiological and molecular landscape of neurodegeneration.

The accelerating development of high-throughput molecular profiling technologies, combined with increasingly powerful computational tools applied to large, deeply phenotyped cohorts, is transforming the landscape of biomarker discovery¹⁰. Although multi-omics approaches, such as integration of genomics, transcriptomics and metabolomics, contribute to rich data-driven biomarker discovery, proteomics is uniquely positioned to impact both diagnosis and treatment of neurodegenerative disease. This is due to several key factors: (1) many clinically established biomarkers are protein based; (2) high-dimensional proteomic platforms such as SomaScan, Olink and mass spectrometry now offer sufficient depth to capture a sizable portion of the circulating proteome; and (3) protein-level changes often capture biological processes proximal to neurodegeneration, providing functional insights that are directly relevant to disease pathogenesis. Proteomic profiles derived from peripheral biofluids such as plasma and cerebrospinal fluid (CSF) not only hold promise for identifying biomarkers of disease presence and progression but also offer new avenues for therapeutic target discovery.

Robust high-dimensionality omics research in heterogeneous clinical groups necessitates the use of large datasets due to poor reproducibility of findings from single-site or smaller cohorts¹¹, but the

siloing of data among a fragmented research community has been a barrier to such biomarker discovery¹². Although many research institutions and initiatives have embarked on a variety of open data efforts, there is no standard model for providing researchers with easy access to data from multiple cohorts. Moreover, the use of such multicohort sources requires data aggregation and harmonization. The genetics research community has enabled huge consortia with joint data access and collaborative analysis¹³. In the field of neurodegenerative disease, large data-sharing efforts such as the ADNI¹⁴, AMP-AD¹⁵, PPMI¹⁶, ALS TDI's ARC¹⁷ and Answer ALS¹⁸ are examples of open datasets that facilitate cross-study collaborations. However, despite these highly productive examples of best practice, their disease-specific design limits identification of shared mechanisms of neurodegeneration and potential co-existing pathologies. In addition, most data have either never been shared or have been kept behind restrictive barriers to access¹². Reasons for this include a shortage of technology solutions, a range of challenging data governance rules and privacy regimes and cultural norms and misaligned incentives among researchers, research institutions, industry and research sponsors.

The GNPC was created to systematically address these challenges to large data analysis, accelerate biomarker discovery and advance the research and development of more precise treatments for neurodegenerative disease. Our goal was to generate a large proteomics resource using available samples from established cohort studies, accompanied by a harmonized clinical dataset, and to make these data available to the scientific community in a rapid and easily accessible manner. Having reached the planned point of public data release in July 2025, the GNPC has established what it thinks is the world's largest neurodegenerative disease-focused proteomics dataset for biomarker research, with 23 partners contributing more than 35,000 analyzed biosamples and approximately 250 million unique protein measurements with matched and harmonized clinical data. Here we summarize the GNPC version 1 (V1) dataset, together with key analysis vignettes. With the associated in-depth papers in this issue^{19–21}, this serves as the beginning of the explorations into this dataset and its contribution to the field of neurodegenerative disease.

Traditional 'on premises' data science analyses have become more challenging as datasets have increased in size, with a corresponding increase in resources required for moving data from one location to another. Moreover, local analysis presents a challenge in ensuring data integrity, safety and confidentiality. The GNPC's partnership with the AD Data Initiative provided the consortium with virtual access to the cloud-resident harmonized dataset with analysis workspaces via the AD Workbench²², a secure, cloud-based environment that is able to satisfy multiple different geographical data jurisdictions (for example, the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA))²³.

For GNPC V1, we opted to use SOMAmer technology (provided by SomaLogic) as the primary proteomics platform as it was one of the broadest discovery platforms available. However, we also analyzed a subset of samples with Olink and mass spectrometry methods to allow for cross-platform comparison. In general, the GNPC's approach is agnostic to platforms and is guided primarily by coverage, reproducibility and affordability. We also think that different platforms bring complementary information as they may measure different isoforms and/or posttranslational modifications of a given protein.

Developing the GNPC's large dataset required addressing several barriers to open data and data sharing. To bring together data from several countries required navigating legal regimes with different requirements, including the GDPR (Europe), the Data Protection Act (United Kingdom) and HIPAA (United States). The GNPC's legal team worked with institutions in each jurisdiction to address specific concerns and agree on a framework for data sharing that worked across the board, including collecting the data on servers physically located in Western Europe.

Table 1 | GNPC contributing cohort details (n=23)

Cohort name	Organization	Geography (country)	Disease area	Sample collection and study methodology
Amyloid-beta in CSF for PD (ABC-PD)	University of Tübingen	Germany	PD	Link 1 , Link 2
ALLFTD	University of California, San Francisco	United States	FTD	Link 1 , Link 2
ALS Therapy Development Institute	ALS Therapy Development Institute	United States	ALS	Link 1
Alzheimer and Families (ALFA)	Barcelonaβeta Brain Research Center	Spain	AD	Link 1 , Link 2
Amsterdam Dementia Cohort (ADC)	Amsterdam University Medical Center	The Netherlands	FTD and AD	Link 1 , Link 2
Answer-ALS	Answer ALS	United States	ALS	Link 1 , Link 2
Baltimore Longitudinal Study of Aging (BLSA)	NIA-NIH	United States	Aging	Link 1 , Link 2
Banner Health	Banner Health	United States	AD and PD	Link 1 , Link 2 , Link 3 , Link 4
BioFINDER	Lund University	Sweden	AD and PD	Link 1 , Link 2
Bio-Hermes	Global Alzheimer's Platform Foundation	United States	AD	Link 1
CHARIOT-PRO	Imperial College London	United Kingdom	AD	Link 1
Emory ADRC	Emory University	United States	AD	Link 1
European Medical Information Framework Multimodal Biomarker Discovery Study (EMIF-AD MBD)	Maastricht University	The Netherlands	AD	Link 1 , Link 2 , Link 3
Fundacio ACE	Fundacio ACE	Spain	Mixed ADRD	Link 1 , Link 2
Indiana ADRC	Indiana University	United States	Mixed ADRD	Link 1
Kansas ADRC	University of Kansas	United States	AD	Link 1
Knight ADRC	Washington University in St. Louis	United States	Mixed ADRD	Link 1 , Link 2
Mayo Clinic Study of Aging (MCSA)	Mayo Clinic	United States	AD	Link 1
Parkinson's Progression Markers Initiative (PPMI)	Michael J. Fox Foundation	United States	PD	Link 1 , Link 2
Religious Orders Study/Memory and Aging Project (ROSMAP)	Rush University	United States	AD	Link 1 , Link 2 , Link 3
Stanford ADRC	Stanford University	United States	AD	Link 1 , Link 2
Tracking PD	Oxford University	United Kingdom	PD	Link 1
Whitehall II	University College London	United Kingdom	Aging, Mixed ADRD	Link 1 , Link 2

All contributing cohorts to the V1 harmonized dataset include the name of the study, the contributing site or organization, the country where the study was conducted, the main disease area or focus of the study and the published sample collection protocols and study methodology. ADRC, Alzheimer's Disease Research Center; ADRD, Alzheimer's disease and related dementias.

The first version of the harmonized data was made available to consortium members in June 2024. Analysis of the harmonized dataset is organized in four workstreams to allow members of the consortium to collaborate on areas of related interest: longitudinal profiling, cross-sectional profiling, proteogenomics and prediction modeling. Here we present the first set of analyses of the GNPC dataset, including the overarching summary analyses and, in accompanying papers, the work conducted in the GNPC workstreams during the first year of data availability.

Results

Initial findings from the GNPC harmonized dataset

The GNPC V1 harmonized dataset is focused on neurodegenerative diseases and, more specifically, on AD, PD, ALS, FTD and aging among 18,645 participants, drawn from 23 individual cohorts across a total of 31,083 unique peripheral plasma, serum and CSF samples, culminating in 35,056 unique proteomic assays and approximately 250 million individual protein measurements (Table 1 and Supplementary Table 1). Most of the proteomics characterization comes from the SOMAmer-based capture array (SomaScan version 4.1, version 4 and version 3 platforms), measuring approximately 7,000 ($n = 26,458$), 5,000 ($n = 4,528$) or 1,300 ($n = 95$) unique aptamers per biosample,

respectively. Additionally, 1,975 of the plasma samples characterized on the version 4.1 SomaScan platform had tandem mass tag mass spectrometry performed. The harmonized dataset additionally includes 40 clinical features, including demographic data, vital signs data and clinical features collected with each blood or CSF draw (Supplementary Table 3). These aggregated and harmonized data demonstrated their value to the consortium immediately, as they served to rapidly confirm signals originally identified in smaller datasets across the entirety of the GNPC V1 dataset, thereby serving as an 'instant validation' resource^{24,25}.

To evaluate the structure and comparability of the blood-based proteomics data, we conducted a principal component analysis on plasma and serum samples (Supplementary Fig. 1). Serum samples clustered distinctly from plasma, reflecting a clear matrix effect. Among plasma samples, a modest offset was observed between the 5K and 7K SomaScan platforms, whereas EDTA and citrate plasma samples appeared largely similar. All three vignettes described below focused exclusively on plasma proteomic data, and, where relevant, platform-related differences between 5K and 7K assays were addressed using scaling or predictive modeling approaches.

To highlight the breadth and utility of the GNPC V1 resource, we present three illustrative vignettes that showcase how this harmonized dataset can be applied to address key questions in neurodegenerative

disease and aging research: (1) disease-specific differential abundance profiling, (2) biological aging across organ systems and (3) protein markers of genetic risk as exemplified by the apolipoprotein E (*APOE*) genotype. As summarized below and described in more detail in the accompanying papers in this issue^{19–21}, these vignettes reflect the analytical depth enabled by the GNPC and are intended to catalyze further exploration by the broader research community upon public data release.

Vignette 1: Human blood proteomic profiles are robustly associated with neurodegenerative diseases and clinical severity. We examined the plasma proteome as measured using the SomaLogic 7K platform, among people with AD, PD, FTD and ALS (referred to hereafter as ‘Patients’, $n = 3,002$), and separately among people with no neurodegenerative disease diagnosis and cognitively normal test screenings (referred to hereafter as ‘Controls’, $n = 5,879$) (see Vignette 1 methods for sample selection criteria). First, we sought to identify proteins differentially abundant in the plasma of patients with different neurodegenerative diseases with cognitive effects—namely, AD, PD, FTD and ALS. Leveraging the breadth of cohorts included in the GNPC, we first performed cohort-stratified analyses to internally validate the most robust protein changes, focusing on those consistently altered across multiple study cohorts. Cohort-stratified results were subsequently combined using a meta-analysis for AD (Fig. 1a), PD (Fig. 1b), FTD (Fig. 1c) and ALS (Fig. 1d).

In AD ($n = 1,966$), 27 proteins from AD Patients robustly emerged as being significantly elevated compared to Controls across at least six of the 10 different cohorts, including ACHE, SPC25, LRRN1 and CTF1. Additionally, GDF2 and APOB also showed high meta-analytic effect sizes and were independently significant in five and four separate cohorts, respectively (Fig. 1a). In contrast, 130 proteins were consistently lower in AD plasma across at least six cohorts, including VAT1, GPD1, ARPC2 and PA2G4. Furthermore, we observed significant decreases in RPS12, NPTXR and NTSC across five cohorts. These top hits highlight both expected and underexplored targets consistently altered in AD plasma across cohorts, including those with established ties to lipid metabolism (APOB and GPD1), cholinergic signaling and/or treatment response (ACHE and VAT1) and synaptic integrity (NPTXR), as well as novel targets linked to cytoskeletal regulation (ARPC2) and RNA metabolism (PA2G4 and RPS12). Follow-up analysis (see Vignette 3) also indicated that elevation of some targets, such as SPC25, LRRN1 and CTF1, reflected underlying *APOE* ε4 genotype effects rather than AD diagnosis per se. Reactome pathway analyses ($n = 2,640$, Bonferroni-adjusted $P < 0.05$) revealed enrichment for terms related to sugar metabolism (‘glucose metabolism’ and ‘glycolysis’) and protein prenylation (‘RAB geranylgeranyltransferase’), reinforcing links to bioenergetics and vesicle trafficking (Fig. 1e).

In PD ($n = 607$), 40 proteins were significantly elevated across at least three of seven different cohorts, including SUMF1, PRR15, AARD3 and RDH16, which were elevated in at least four cohorts (Fig. 1b). Additional proteins such as PSMC5, DDX1 and VSIR exhibited strong effect sizes and replicated in two cohorts. In contrast, 15 proteins were significantly lower in PD plasma across at least three cohorts, including CLEC3B, GPD1 and SEMA4G. Meta-analytic effect sizes were very high for PRSS8, BAGE3, NPS, PRL and HEXB, all of which decreased in PD but were less reproducible across contributing cohorts, suggesting cohort-specific factors driving depleted abundance of these targets. These candidate PD-associated proteins include targets associated with proteostatic (SUMF1, HEXB and PSMC5), immune (VSIR and CLEC3B) and axonal guidance (SEMA4G) pathways, possibly reflecting both peripheral and brain-related pathophysiology. Similar to in AD, Reactome pathway analyses ($n = 2,251$, Bonferroni-adjusted $P < 0.05$) revealed enrichment for terms related to Ras superfamily/small GTPases and vesicle trafficking (‘vesicle-mediated transport’ and ‘signaling by RHO GTPases’), highlighting a plasma proteome pathway overlap between AD and PD (Fig. 1f).

Although FTD clinical syndromes are less common than AD or PD and have greater clinical and neuropathological diversity, nine targets exhibited decreased abundance in FTD plasma ($n = 175$) across multiple cohorts after Bonferroni correction (Fig. 1c). Strongly down-regulated hits included NPTXR, APLP1 and HS6ST3, which converge on processes critical for synaptic maintenance and neuronal support. Eleven proteins were significantly elevated in FTD plasma with a conventional false discovery rate (FDR) correction but did not survive Bonferroni correction. Despite limited power, Reactome pathway analyses ($n = 71$, FDR < 0.05) revealed two significantly enriched terms, ‘posttranslational protein phosphorylation’ and ‘regulation of insulin-like growth factor transport and uptake by IGFBPs’, highlighting conserved peripheral signatures of neurodegeneration even amid the clinical and pathological heterogeneity of FTD (Fig. 1g).

In ALS ($n = 254$), we analyzed plasma proteomic profiles from a single contributing cohort with Patients and Controls (Fig. 1d). After FDR correction, 44 targets exhibited significantly increased abundance in ALS, including a host of proteins related to skeletal muscle structure and function (PDLIM3, MYOM2, MYLPF and TNNI2). Thirty-eight targets exhibited significantly decreased abundance in ALS, including two aptamers targeting ART3, an ADP-ribosyltransferase enriched in skeletal muscle, as well as additional proteins linked to growth factor signaling and/or extracellular matrix composition (ANTXR2, CRTAC1 and RGMA). Reactome pathway analyses ($n = 82$, FDR < 0.05) confirmed this strong biological enrichment for skeletal muscle-related processes (‘muscle contraction’ and ‘collagen chain trimerization’) (Fig. 1h), underscoring a clear peripheral proteomic footprint of ALS consistent with its primary motor system pathology.

After identifying disease-related differential abundance patterns, we combined data across AD, PD and FTD to identify a global signature of dementia severity. Specifically, a 256-protein clinical impairment signature was derived using least absolute shrinkage and selection operator (LASSO)-based prediction of Clinical Dementia Rating (CDR) global scores, which was subsequently evaluated in a held-out test set (2,047 records; 30% of the dataset). Signature values correlated with CDR global scores (train: Pearson’s $r = 0.68$; test: $r = 0.58$) in a stepwise fashion, increasing at each level of clinical severity (Fig. 1i; see Supplementary Table 7 for individual-level scores). To demonstrate concordance with an orthogonal clinical outcome, we observed that higher signature values were also associated with lower cognitive test scores (standardized Montreal Cognitive Assessment/Mini-Mental State Examination (MoCA/MMSE); $r = -0.47$), consistent with expected inverse correlations between cognition and CDR. In disease-stratified analysis combined across training and test sets, the multivariate protein signature was reliably elevated with greater clinical severity in AD ($r = 0.55$), FTD ($r = 0.85$) and PD ($r = 0.70$), supporting its relevance as a transdiagnostic marker of clinical severity (Fig. 1j). Top proteins with high feature importance in the clinical impairment signature (Fig. 1k and Supplementary Table 8) again highlighted ACHE and NPTXR as well as additional targets linked to neuroplasticity (EPHA4 and CNTFR) and immune activation (MSMP and KLK3), underscoring their potential for transdiagnostic dementia staging.

Vignette 2: Organ age analysis reveals neurodegenerative disease-specific patterns of premature aging. One advantage of plasma proteomics is the ability to simultaneously query the health of distinct organ systems. We applied previously validated plasma proteomic organ aging models²⁶ to assess accelerated organ-specific aging across multiple Patient and Control data in the GNPC. Predicted organ ages showed moderate to strong correlations with chronological age (r range, 0.36–0.92; Fig. 2a), confirming model performance in the GNPC cohort²⁶. Figure 2b shows the association between organ age gaps, which capture person-specific differences between estimated organ age and actual age, and AD, FTD and PD, respectively. Elevated cognition-enriched brain age gaps, reflecting the subset

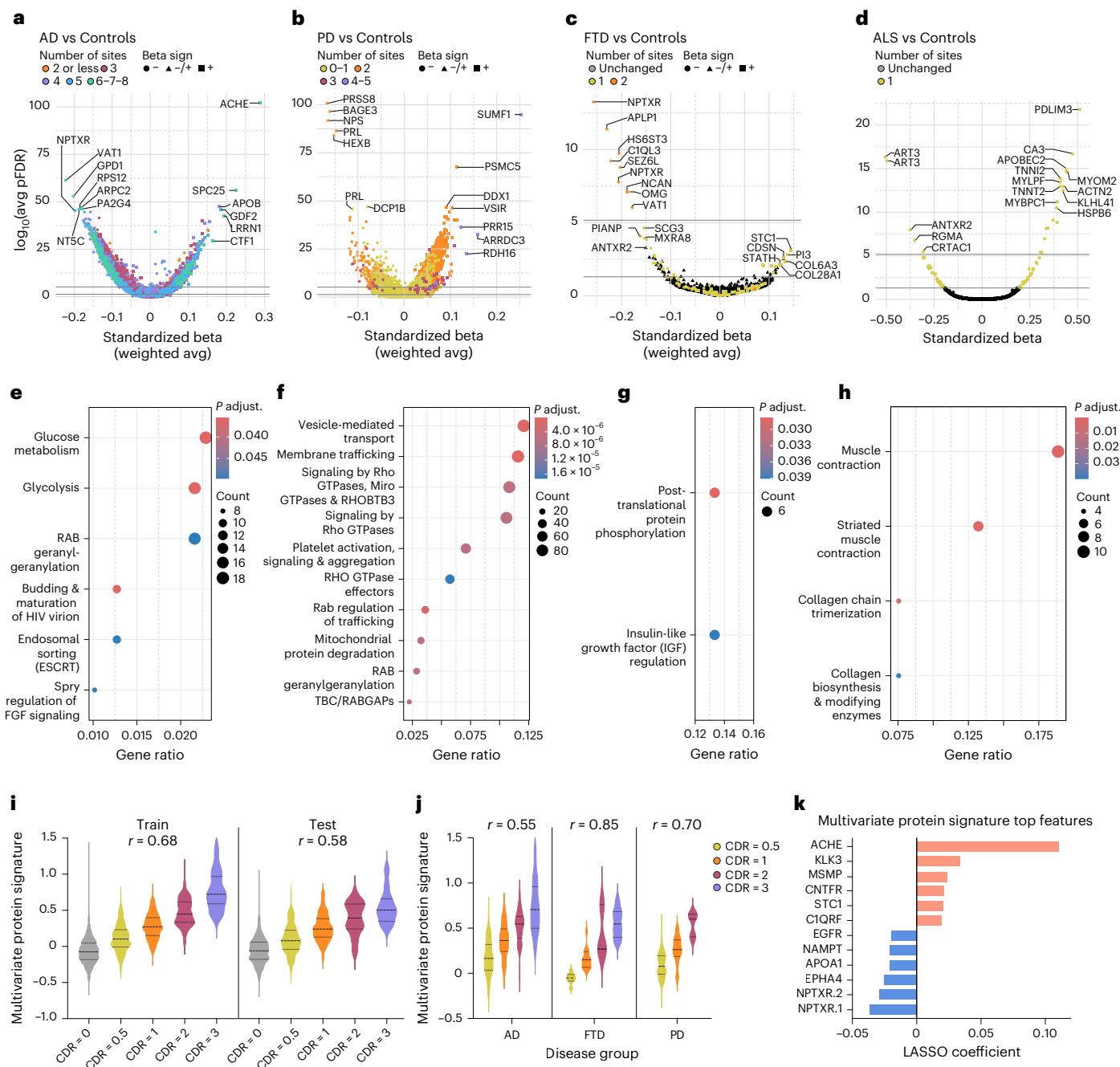


Fig. 1 | Circulating blood proteome specifies neurodegenerative disease type, mechanism and clinical severity. **a–d**, Meta-analytic differential abundance analysis showing changes in relative protein expression of AD (a), PD (b), FTD (c) and ALS (d) compared to Controls. Each dot represents a protein. The x axis shows the direction and effect size of protein changes relative to Controls, from linear regression models including age and sex as covariates; the y axis shows the $-\log_{10}$ FDR-adjusted P value. P values from two-sided tests and after adjustment from FDR are reported. The parallel line at the bottom of each plot shows which proteins are significant after FDR correction for multiple comparisons. The line above shows proteins further surviving Bonferroni correction. Dots are colored based on the number of cohorts where the protein was found to be independently significant after (within-cohort) FDR correction and changed in the same direction relative to Controls (that is, increased or decreased compared to Controls). **e–h**, Significant proteins from the differential

abundance analyses were fed into Reactome enrichment analysis for AD (e), PD (f), FTD (g) and ALS (h), using unique SomaScan 7K proteins as background. Enriched Reactome pathway terms for each condition are visualized as dot plots, with dot size corresponding to the number of differentially abundant proteins assigned to a given pathway (one-sided Fisher's test with FDR adjustments). Full Reactome enrichment summary statistics are reported in Supplementary Table 6. HIV, human immunodeficiency virus; FGF, fibroblast growth factor; TBC/RABGAPs, Tre2-Bub2-Cdc16 (TBC) domain-containing RAB-specific GTPase-activating proteins. **i**, Violin plots displaying LASSO-derived clinical severity protein signatures across CDR global level in training and test sets. **j**, Violin plots displaying LASSO-derived clinical severity protein signatures across CDR global level (0.5 and higher) in AD, FTD and PD, using the combined training and test sets. **k**, LASSO coefficients for the top 12 protein aptamers selected in the clinical severity protein signature. avg, average; pFDR, FDR-corrected P value.

of brain-specific proteins that previously enhanced model age gap prediction of cognitive impairment, were associated with higher odds of AD (odds ratio = 1.33 per 1-s.d. age gap increase (95% confidence

interval: 1.25–1.41)) and FTD (odds ratio = 1.26 (95% confidence interval: 1.06–1.48)). Non-cognition-enriched brain age gap was weakly associated with AD risk (odds ratio = 1.08 (95% confidence interval: 1.02–1.14))

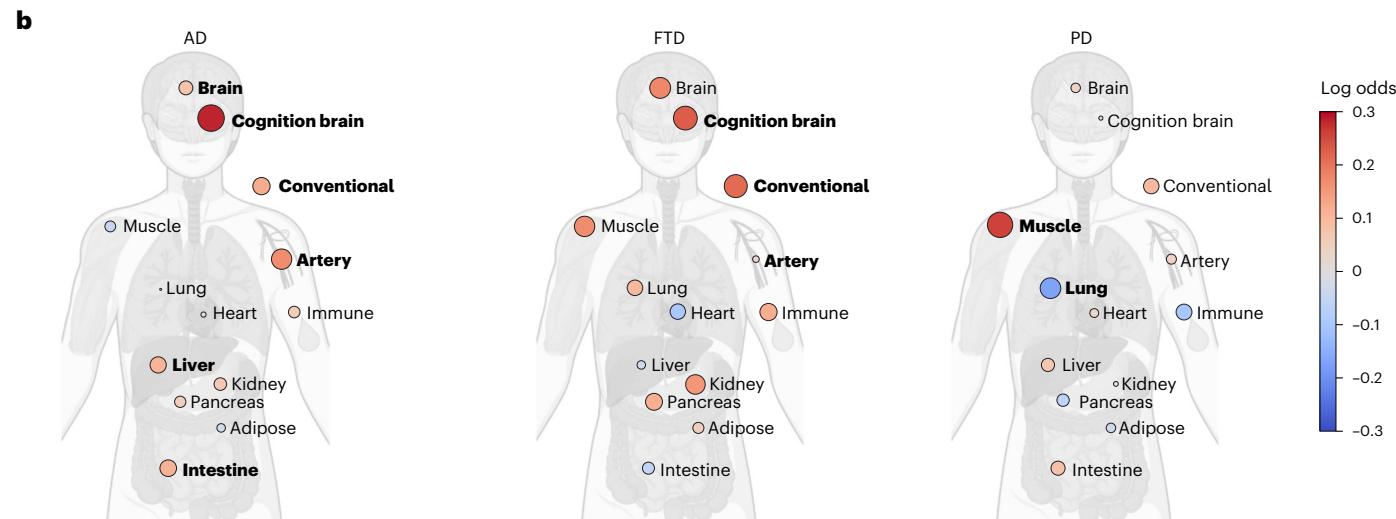
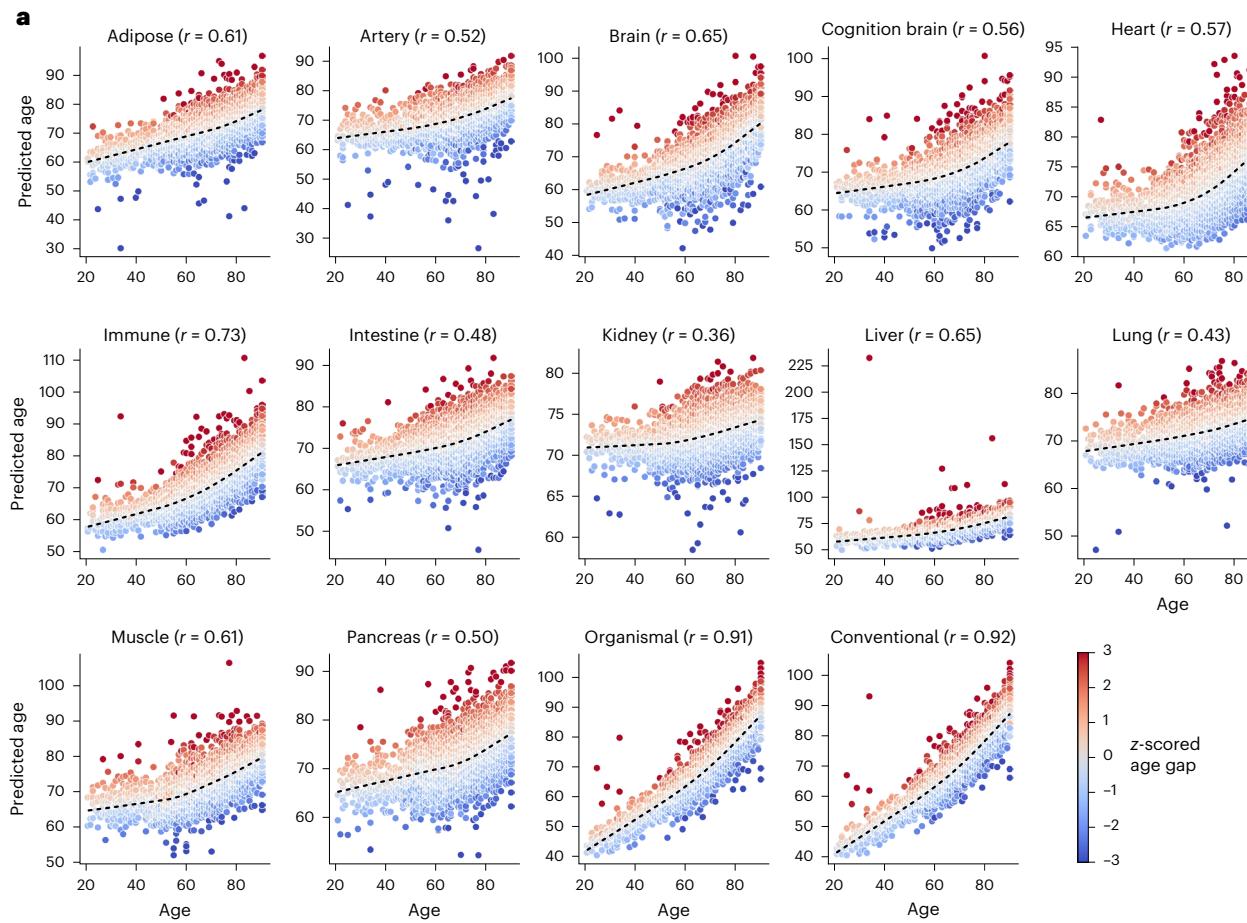


Fig. 2 | Organ age patterns characterize distinct neurodegenerative disease types. **a**, Scatterplots of chronological age versus predicted age for each organ aging clock in clinically normal individuals. Black dashed line indicates the LOWESS regression estimate of the population mean. Pearson's correlation coefficient r is reported for each clock. **b**, Body plots showing associations of standardized organ age gaps with neurodegenerative disease based on binary

logistic regression models. P values are from two-sided tests. Red dots indicate positive associations (higher age gap with disease); blue dots indicate negative associations (lower age gap with disease). Bold labels highlight organ ages associated with organ age gap with $P < 0.05$ after FDR correction. The body plots were created in BioRender: Oh, H. (2025): <https://BioRender.com/afoqtzw>.

but not other conditions. Beyond brain, we observed contributions of artery (odds ratio = 1.18 (95% confidence interval: 1.11–1.25)), liver (odds ratio = 1.11 (95% confidence interval: 1.05–1.17)) and intestine aging to AD (odds ratio range, 1.12–1.18) as well as a unique link between muscle aging and PD (odds ratio = 1.12 (95% confidence interval:

1.05–1.19)). These findings extend previous work by demonstrating shared and distinct patterns of blood-detectable accelerated organ aging across AD, FTD and PD, underscoring connections between systemic health and neurodegenerative disease that may be related as a cause, correlate or consequence.

Vignette 3: Human blood proteomic signatures reflect *APOE* genetic status and uncouple systemic AD and *APOE* effects. Using a combination of machine learning and biological enrichment approaches, we sought to isolate the molecular signatures of *APOE* ε4, the main genetic risk factor for sporadic AD, independent of AD and other conditions. Several proteins, including SPC25, LRRN1, S100A13 and NEFL, were strongly associated with *APOE* ε4 versus other alleles (Fig. 3a), paralleling prior observations for some of these targets in serum²⁷. Some proteins, such as SPC25 and LRRN1, showed no difference between AD and Controls but exhibited dose-dependent associations with *APOE* ε4 (Fig. 3b), suggesting that previously identified links to AD from Vignette 1's differential abundance analysis were driven by *APOE* ε4 enrichment in AD cases. Conversely, well-known AD-associated neuromodulatory proteins such as NPTXR and GDF2 were robustly associated with AD diagnosis, irrespective of *APOE* ε4 allelic dose (Fig. 3c).

Notably, the effects of *APOE* ε4 genotype on the plasma proteome were so robust that a machine learning model with only five proteins (SPC25, NEFL, S100A13, TBCA and LRRN1) predicted *APOE* ε4 carrier status in unseen patients with high accuracy, both within AD and within non-AD Patients (area under the curve (AUC) range, 0.90–0.96; Fig. 3d). Leveraging protein–protein interaction libraries and single brain cell RNA sequencing data from the Human Protein Atlas²⁸, we observed that three of these proteins (SPC25, TBCA and S100A13) were central nodes in the protein–protein interaction network (Fig. 3e) and brain cell type expression patterns (Fig. 3f)²⁸. Ubiquitin-C, however, was found to be a key connection point across the central node proteins, suggesting potential convergence on proteostatic pathways.

Lastly, to identify potential genotype–phenotype links, we compared proteins associated with *APOE* ε4 in cognitively unimpaired individuals ($n = 2,817$; 215 proteins) to those associated with AD in *APOE* ε3 homozygotes ($n = 1,843$; 2,150 proteins). Forty-four overlapping proteins showed consistent directionality (Fig. 3g), including targets elevated in *APOE* ε4 and AD. Patients involved in immunovascular signaling (MMP8), synaptic vesicle fusion (SNAP23) and lipid trafficking (APOB) pathways relevant to *APOE* biology and cognitive decline. This overlap highlights potential early molecular footprints of AD pathophysiology present even in asymptomatic *APOE* ε4 carriers and underscores the utility of contextualizing by *APOE* genotype when seeking AD-relevant proteomic signals. These shared proteins may reflect core features of *APOE*-related biology that are also prominent in AD, highlighting potential mechanisms through which *APOE* ε4 contributes to disease vulnerability.

Discussion

The availability of high-dimensional molecular datasets has led to an increasing number of large-scale collaborative programs to share and use these data—a trend forged by the genetics community. Following these data-sharing collaborations, numerous initiatives have led the way in data sharing with the wider scientific community. We would like to highlight in particular two programs focused on neurodegeneration or proteomics. The first large-scale open data-sharing program in neurodegeneration was the Alzheimer's Disease Neuroimaging Initiative (ADNI): an immensely productive public–private partnership that continues today and has spawned many followers. These include AddNeuroMed/InnoMed, an ADNI-like program in Europe that served as a pilot for the European Union Innovative Medicines Initiative (IMI) funding scheme that itself has generated many data-sharing and sample-sharing programs, including, for example, IMI-EMIF and IMI-EPND in neurodegeneration. In the proteomics arena, the GNPC was preceded by the UK Biobank Pharma Proteomics Project (UKBB-PPP) that generated extensive protein data on 50,000 research participants and is now planning analysis on an additional 250,000 participants, to accompany the extensive clinical, imaging and genomic data available

to the scientific community. The GNPC complements both initiatives and many others in providing a disease-focused dataset, as in the ADNI, but at scale, as in the UKBB-PPP.

Notably, a key point of emphasis of the GNPC's mission is the strong intent to share this dataset with the global research community early in its life cycle. We do not present these analyses as definitive but, rather, as the pilot experiments by a subset of researchers whose datasets contributed to the construction of GNPC V1. We hope this summary paper and more in-depth papers serve as an invitation of collaboration and/or independent analysis from the global community. Only then will we be able to maximize disease insights from GNPC V1 and its combination with other datasets to accelerate translation of insights into the next generation of diagnostics and therapeutics for neurodegenerative diseases.

The initial analyses presented here underscore the versatility and translational potential of the GNPC dataset. First, disease-specific differential abundance and disease-shared clinical severity analyses revealed both established and novel protein targets in plasma across AD, PD and FTD, highlighting shared and distinct biological processes such as vesicle trafficking, synaptic integrity and metabolic dysregulation. These results not only validate previously reported protein markers but also nominate new candidates for mechanistic follow-up and blood biomarker development—an urgent clinical need across diseases. Second, organ aging clocks applied to the GNPC dataset uncovered disease-specific patterns of accelerated aging across brain and peripheral organs, offering a systems-level view of proteomic aging that bridges central and systemic health. These findings extend previous work on biological aging by demonstrating that distinct conditions are associated with unique organ-specific age gaps, supporting their relevance to age-related neurodegenerative diseases. Finally, proteome-wide analysis of the *APOE* genotype revealed a robust and disease-independent *APOE* ε4 signature, with potential mechanistic relevance to proteostasis and lipid transport. These vignettes, each explored further in companion publications^{19–21}, illustrate the range of insights made possible by GNPC V1 and set the stage for future hypothesis-driven and exploratory research by the broader scientific community, with the potential to better support trial design, monitoring and subtyping of clinical patients.

These data, as well as these vignette analyses and those in the accompanying papers, suggest that very large protein datasets have potential to add value to drug discovery. Hitherto, the value of genetic data has been increasingly recognized in supporting effective drug discovery. Notably, targets with genetic support are more likely to progress through the drug discovery pipeline²⁹ with probability of success recently being calculated to be 2.6 times greater than in targets lacking genetic support, even for genes with small effect sizes³⁰. In contrast to the inherited traits represented by genetic variants, proteomics represents biological states. Although genetic factors are intrinsically causal in their relationship with disease, a proteomic association with disease might be consequential of the disease, a factor associated with disease (including response to a therapeutic) or reflect a causal process. Although these observations might suggest a role for proteomics more in supporting drug discovery through biomarker discovery rather than for target identification, the vignettes reported here and in the accompanying papers illustrate the potential for more direct drug target identification/validation. For example, in Vignette 1, the proteomic profiles identified include strong support for synaptic dysfunction with proteins identified that are already clearly part of a mechanism targeted for neurodegeneration drug discovery, such as NPTXR³¹, whereas, in Vignette 3, proteins are identified that are very strongly associated with *APOE* status and with disease state. These proteins will surely now be considered as possible targets for drug discovery, especially considering that the *APOE* genotype is a striking example of a very strong genetic risk factor that has not been the source of equally strong drug discovery programs. The GNPC data also support

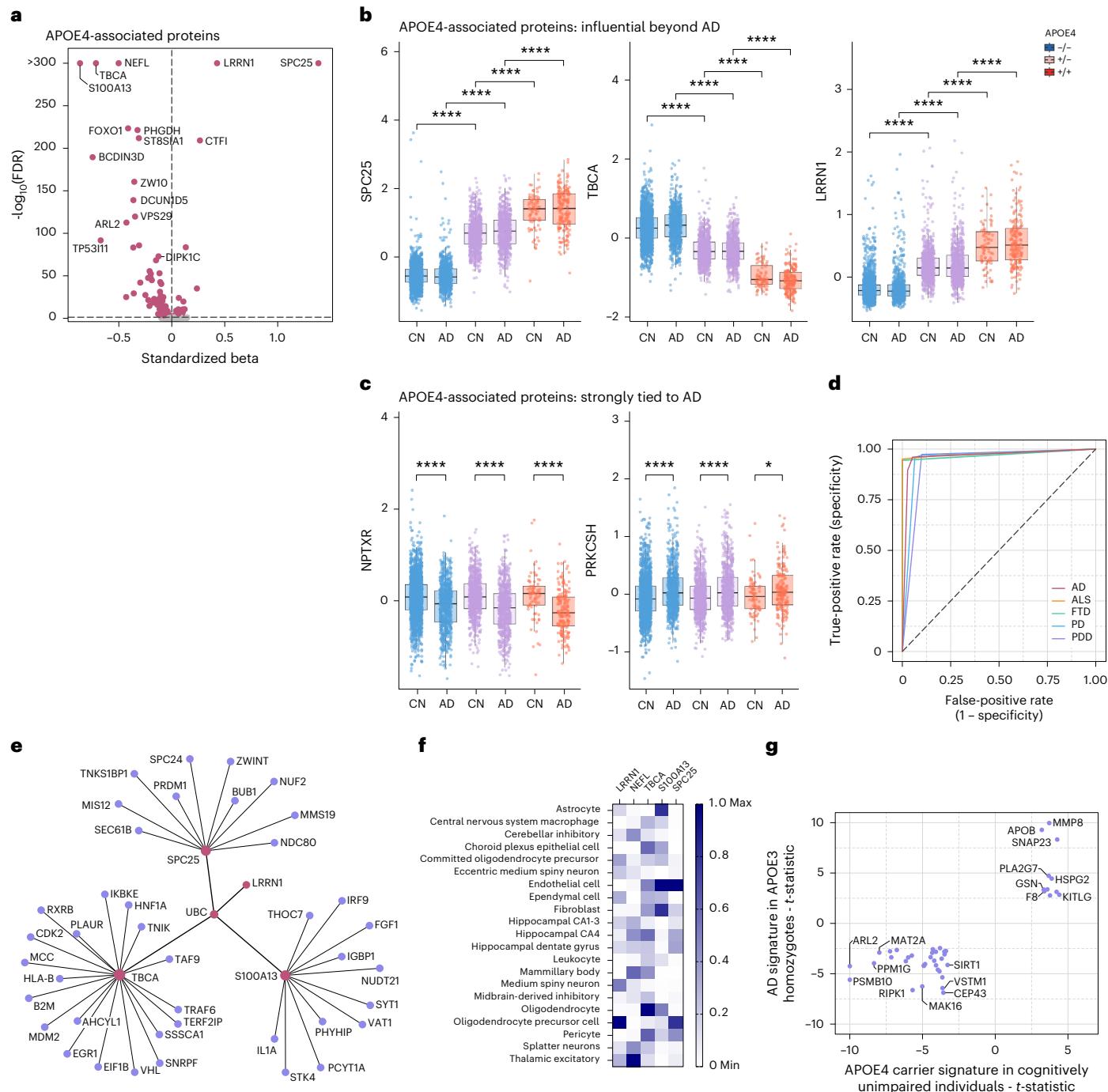


Fig. 3 | Disease-dependent and disease-independent of APOE ε4 on the human proteome. **a**, Volcano plot shows the protein association profile of APOE ε4 after adjusting for AD dementia diagnosis, with red representing significant associations (after FDR correction). At the y axis, the $-\log_{10}$ (FDR-adjusted P values) > 300 were set to 300 for better visualization. This was done for S100A13, TBCA, NEFL, LRRN1 and SPC25. **b,c**, Box plots show plasma protein level changes of the proteins with the strongest APOE ε4 associations (**b**) and for APOE ε4-associated proteins strongly tied to AD dementia diagnosis (**c**). For **b** and **c**, the y axis represents residual protein levels after adjusting for age, sex, mean protein level and contribution site. The center line of each box indicates the median, with lower and upper edges representing the 25th and 75th percentiles. Whiskers extend to the most extreme values within 1.5 times the interquartile range; data points beyond this range were excluded as outliers. The x axis represents AD diagnosis. The color indicates APOE ε4 carrier status; ‘-/-’ indicates APOE ε4 non-carriers; ‘±’ indicates ε3/ε4; and ‘+/+’ indicates ε4/ε4. Welch’s t -test was used to compare residual protein levels between groups. Two-sided P values are reported. *** $P < 0.0001$ and * $P < 0.05$. P values were not

adjusted for multiple comparisons, as only prespecified group contrasts are shown. Results marked with *** remain significant ($p\text{FDR} < 0.0001$) even after adjustment for multiple comparisons with the Benjamini–Hochberg method, whereas those marked with * do not. **d**, Receiver operating characteristic area under the curve (ROC-AUC) showing the performance of a machine learning model using only five proteins to predict APOE ε4 status across different diagnostic groups, in a held-out sample. **e**, Protein interaction network including four of those five proteins (red). **f**, Neural cell type expression of RNA transcripts encoding the five APOE ε4-predictive proteins. Plot shows mix-max scaling of protein-coding transcripts per million for each identified APOE ε4 protein. **g**, Correlation of effect sizes for proteins associated with APOE ε4 in cognitively unimpaired samples (x axis) and AD associated with AD diagnosis in APOE ε3/ε3 homozygotes. Limma t -statistic is shown for both contrasts; only proteins associated with both APOE ε4 and AD (adjusted $P < 0.05$ for both analyses) with the same direction of effect are visualized. For visibility purposes, t -statistic values higher than 10 were capped. PDD, Parkinson’s disease dementia; pFDR, FDR-corrected P value.

target validation, an important component of target identification in driving drug discovery. Preclinical models might be used to generate signatures of targets or interventions, and, using GNPC data, these signatures now can be used to predict outcomes. In reverse, signatures or candidate protein targets identified in the GNPC could be validated in such preclinical models. Whether in conjunction with genetic data or with preclinical models, the proteomics data now being made available are likely to become a strong additional component of effective target identification and validation. As additional datasets similar to the GNPC and UKBB-PPP become available, it is possible that proteomics will become as important an element of drug discovery as genetics is today.

Despite the evident success in establishing a substantial dataset, the GNPC has some limitations. These include the relative lack of diversity, reflecting much of past observational bioresources where most research participants have been individuals of European ancestry living in the Global North. The dataset would also be enhanced with other proteomics platforms, including complementary molecular biomarker data including genomics, transcriptomics and metabolomics as well as imaging and clinical data and with other disease types related to neurodegeneration. Data harmonization challenges in a post hoc meta-analysis such as the GNPC include site differences in sample processing, clinical methodologies for diagnoses and patient demographics. For example, incomplete medication information limited the ability to identify and separate definitive drug-induced changes to the proteome from disease-relevant changes and/or to correlate proteomic signatures comprehensively with amyloid and synuclein biomarkers. However, despite the cross-site heterogeneity, we were nonetheless able to identify clear and novel signals in the plasma proteome as presented in the vignettes.

The GNPC has been built to accommodate growth and aims to increase diversity in terms of patients and measurements over time. As a next step, the GNPC is poised to incorporate additional cohorts, samples and data platforms in a V2 dataset and will seek to rectify some of these limitations as we build out the collaboration. The flexibility of the platform allows for secure, iterative data releases to adapt to new discoveries made in the field. The GNPC's aspiration is to facilitate neurodegenerative disease research and development, driving advances toward better outcomes for people with neurodegenerative diseases through precision, combination therapy optimized for a patient's disease subtypes.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41591-025-03834-0>.

References

1. GBD 2019 Dementia Forecasting Collaborators. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. *Lancet Public Health* **7**, e105–e125 (2022).
2. Huang, L., Kuan, Y., Lin, H.-W. & Hu, C. Clinical trials of new drugs for Alzheimer disease: a 2020–2023 update. *J. Biomed. Sci.* **30**, 80 (2023).
3. Miller, R. G., Mitchell, J. D., Lyon, M. & Moore, D. H. Riluzole for amyotrophic lateral sclerosis (ALS)/motor neuron disease (MND). *Cochrane Database Syst. Rev.* **2012**, CD001447 (2002).
4. Boxer, A. L. & Sperling, R. Accelerating Alzheimer's therapeutic development: the past and future of clinical trials. *Cell* **186**, 4757–4772 (2023).
5. Cummings, J. et al. Alzheimer's disease drug development pipeline: 2023. *Alzheimers Dement.* **9**, e12385 (2023).
6. van der Flier, W. M., de Vugt, M. E., Smets, E. M. A., Blom, M. & Teunissen, C. E. Towards a future where Alzheimer's disease pathology is stopped before the onset of dementia. *Nat. Aging* **3**, 494–505 (2023).
7. Blennow, K. & Zetterberg, H. Biomarkers for Alzheimer's disease: current status and prospects for the future. *J. Intern. Med.* **284**, 643–663 (2018).
8. Hendl, T. J. et al. Proteomics approaches for biomarker and drug target discovery in ALS and FTD. *Front. Neurosci.* **13**, 548 (2019).
9. Jack, C. R. et al. Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's Association Workgroup. *Alzheimers Dement.* **20**, 5143–5169 (2024).
10. Vilkaite, G., Vogel, J. & Mattsson-Carlsgren, N. Integrating amyloid and tau imaging with proteomics and genomics in Alzheimer's disease. *Cell Rep. Med.* **5**, 101735 (2024).
11. Mattsson-Carlsgren, N., Palmqvist, S., Blennow, K. & Hansson, O. Increasing the reproducibility of fluid biomarker studies in neurodegenerative studies. *Nat. Commun.* **11**, 6252 (2020).
12. McHugh, C., Clement, M. & Phatak, M. AD Workbench: transforming Alzheimer's research with secure, global, and collaborative data sharing and analysis. *Alzheimers Dement.* **21**, e70278 (2025).
13. Alzheimer's Disease Sequencing Project (ADSP) Data Discovery Portal (National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS), accessed 2 July 2025); <https://adsp.niagads.org/>
14. Alzheimer's Disease Neuroimaging Initiative (ADNI) (ADNI, accessed 2 July 2025); <https://adni.loni.usc.edu/>
15. Accelerating Medicines Partnership® Program for Alzheimer's Disease (AMP® AD) (National Institute on Aging, US Department of Health and Human Services, National Institutes of Health, accessed 2 July 2025); <https://www.nia.nih.gov/research/amp-ad>
16. PPMI: A Landmark Study to Identify Biomarkers of Parkinson's Disease Progression (Parkinson's Progression Markers Initiative, 2 July 2025); <https://www.ppmi-info.org/>
17. ARC Data Commons (ALS Therapy Development Institute, 2025); <https://www.als.net/arc/data-commons/>
18. Answer ALS: The Most Comprehensive ALS Research Initiative (Answer ALS, accessed 2 July 2025); <https://www.answersals.org/>
19. Ali, M. et al. Shared and disease-specific pathways in fronto-temporal dementia and Alzheimer's and Parkinson's diseases. *Nat. Med.* <https://doi.org/10.1038/s41591-025-03833-1> (2025).
20. Farinas, A. et al. Disruption of the cerebrospinal fluid–plasma protein balance in cognitive impairment and aging. *Nat. Med.* <https://doi.org/10.1038/s41591-025-03831-3> (2025).
21. Shvetsov, A. et al. APOE ε4 carriers share immune-related proteomic changes across neurodegenerative diseases. *Nat. Med.* <https://doi.org/10.1038/s41591-025-03835-z> (2025).
22. AD Discovery Portal (Alzheimer's Disease Data Initiative, AD Workbench, accessed 2 July 2025); <https://discover.alzheimersdata.org/>
23. Toga, A. W. et al. The pursuit of approaches to federate data to accelerate Alzheimer's disease and related dementia research: GAAIN, DPUK, and ADDI. *Front. Neuroinform.* **17**, 1175689 (2023).
24. Oh, H. S.-H. et al. A cerebrospinal fluid synaptic protein biomarker for prediction of cognitive resilience versus decline in Alzheimer's disease. *Nat. Med.* **31**, 1592–1603 (2025).
25. Cruchaga, C. et al. Large-scale plasma proteomic profiling unveils novel diagnostic biomarkers and pathways for Alzheimer's disease. *Nat. Aging* **5**, 1114–1131 (2025).
26. Frick, E. A. et al. Serum proteomics reveal APOE-ε4-dependent and APOE-ε4-independent protein signatures in Alzheimer's disease. *Nat. Aging* **4**, 1446–1464 (2024).
27. Uhlen, M. et al. A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. *Science* **366**, eaax9198 (2019).

28. Sjöstedt, E. et al. An atlas of the protein-coding genes in the human, pig, and mouse brain. *Science* **367**, eaay5947 (2020).
29. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* **12**, 581–594 (2013).
30. Minikel, E. V. & Nelson, M. R. Human genetic evidence enriched for side effects of approved drugs. *PLoS Genet.* **21**, e1011638 (2025).
31. Zhou, J. et al. The neuronal pentraxin Nptx2 regulates complement activity and restrains microglia-mediated synapse loss in neurodegeneration. *Sci. Transl. Med.* **15**, eadfo141 (2023).

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

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

© The Author(s) 2025

Farhad Imam  ¹, **Rowan Saloner**  ², **Jacob W. Vogel**  ³, **Varsha Krish** ¹, **Gamal Abdel-Azim** ⁴, **Muhammad Ali**  ^{5,6}, **Lijun An**  ³, **Federica Anastasi**  ^{7,8,9}, **David Bennett** ¹⁰, **Alexa Pichet Binette** ^{11,12,13}, **Adam L. Boxer**  ², **Martin Bringmann**  ⁴, **Jeffrey M. Burns**  ^{14,15}, **Carlos Cruchaga**  ^{5,6,16}, **Jeff L. Dage** ^{17,18}, **Amelia Farinas**  ^{19,20,21}, **Luigi Ferrucci** ²², **Caitlin A. Finney**  ^{23,24}, **Mark Frasier** ²⁵, **Oskar Hansson**  ¹¹, **Timothy J. Hohman**  ^{26,27}, **Erik C. B. Johnson**  ^{28,29}, **Mika Kivimaki**  ^{30,31}, **Roxanna Korologou-Linden** ³², **Agustin Ruiz Laza**  ^{33,34,35}, **Allan I. Levey**  ^{28,29}, **Inga Liepelt-Scarfone** ^{36,37,38}, **Lina Lu** ¹¹, **Niklas Mattsson-Carlgren**  ^{11,39}, **Lefkos T. Middleton**  ³², **Kwang sik Nho** ⁴⁰, **Hamilton Se-Hwee Oh** ^{20,21,41}, **Ronald C. Petersen** ⁴², **Eric M. Reiman**  ⁴³, **Oliver Robinson**  ^{32,44}, **Jeffrey D. Rothstein** ⁴⁵, **Andrew J. Saykin**  ^{17,18}, **Artur Shvetsov** ^{23,24}, **Chad Slawson**  ^{14,46}, **Bart Smets** ⁴⁷, **Marc Suárez-Calvet** ^{7,8,48}, **Betty M. Tijms**  ^{49,50}, **Maarten Timmers** ⁴⁷, **Fernando Vieira**  ⁵¹, **Natalia Viloz-Tejedor** ^{7,9,52}, **Pieter Jelle Visser** ^{49,50,53}, **Keenan A. Walker**  ⁵⁴, **Laura M. Winchester** ⁵⁵, **Tony Wyss-Coray**  ^{20,21,56}, **Chengran Yang**  ^{5,6}, **Niranjan Bose**  ¹, **Simon Lovestone** ⁵⁷ & **The Global Neurodegeneration Proteomics Consortium (GNPC)***

¹Gates Ventures, Seattle, WA, USA. ²Department of Neurology, University of California, San Francisco, San Francisco, CA, USA. ³Department of Clinical Sciences Malmö, SciLifeLab, Lund University, Lund, Sweden. ⁴Johnson & Johnson, Spring House, PA, USA. ⁵Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, USA. ⁶NeuroGenomics and Informatics Center, Washington University School of Medicine, St. Louis, MO, USA. ⁷Barcelonaβeta Brain Research Center (BBRC), Pasqual Maragall Foundation, Barcelona, Spain. ⁸Hospital del Mar Research Institute, Barcelona, Spain. ⁹Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. ¹⁰Department of Neurological Sciences, Rush Alzheimer's Disease Center, Chicago, IL, USA. ¹¹Clinical Memory Research Unit, Department of Clinical Sciences Malmö, Lund University, Lund, Sweden. ¹²Department of Physiology and Pharmacology, Université de Montréal, Montreal, Quebec, Canada. ¹³Montreal Geriatrics Institute Research Center, Montreal, Quebec, Canada. ¹⁴University of Kansas Alzheimer's Disease Research Center, Kansas City, KS, USA. ¹⁵Department of Neurology, University of Kansas Medical Center, Kansas City, KS, USA. ¹⁶Department of Neurology, Washington University School of Medicine, St. Louis, MO, USA. ¹⁷Indiana Alzheimer's Disease Research Center, Indianapolis, IN, USA. ¹⁸Department of Neurology, Indiana University School of Medicine, Indianapolis, IN, USA. ¹⁹Graduate Program in Neuroscience, Stanford University, Stanford, CA, USA. ²⁰The Phil and Penny Knight Initiative for Brain Resilience, Stanford University, Stanford, CA, USA. ²¹Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA, USA. ²²Translational Gerontology Branch, National Institute on Aging, Bethesda, MD, USA. ²³Neurodegeneration and Precision Medicine Research Group, Westmead Institute for Medical Research, Westmead, New South Wales, Australia. ²⁴Faculty of Medicine and Health, University of Sydney School of Medical Sciences, Westmead, New South Wales, Australia. ²⁵Michael J. Fox Foundation, New York, NY, USA. ²⁶Vanderbilt Memory & Alzheimer's Disease, Department of Neurology, Vanderbilt University Medical Center, Nashville, TN, USA. ²⁷Vanderbilt Genetics Institute, Vanderbilt Medical Center, Nashville, TN, USA. ²⁸Emory University School of Medicine, Atlanta, GA, USA. ²⁹Department of Neurology, Emory University School of Medicine, Atlanta, GA, USA. ³⁰UCL Brain Sciences, University College London, London, UK. ³¹University of Helsinki, Clinicum, Helsinki, Finland. ³²Ageing & Epidemiology (AGE) Research Unit, School of Public Health, Imperial College London, London, UK. ³³Ace Alzheimer Center Barcelona, Universitat Internacional de Catalunya, Barcelona, Spain. ³⁴Biomedical Research Networking Centre in Neurodegenerative Diseases (CIBERNED), National Institute of Health Carlos III, Madrid, Spain. ³⁵Glenn Biggs Institute for Alzheimer's & Neurodegenerative Diseases and Department of Microbiology, Immunology and Molecular Genetics, Long School of Medicine, The University of Texas Health Science Center, San Antonio, TX, USA. ³⁶Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, Tübingen, Germany. ³⁷Department of Neurodegenerative Diseases, German Center of Neurodegenerative Diseases, Tübingen, Germany. ³⁸IB Hochschule für Gesundheit und Soziales, Standort Stuttgart, Germany. ³⁹Memory Clinic, Skåne University Hospital, Malmö, Sweden. ⁴⁰Department of Radiology & Imaging Sciences, Indiana University School of Medicine, Indianapolis, IN, USA. ⁴¹Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ⁴²Department of Neurology, Mayo Clinic, Rochester, MN, USA. ⁴³Banner Alzheimer's Institute, Phoenix, AZ, USA. ⁴⁴Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK. ⁴⁵Robert Packard Center for ALS Research, Johns Hopkins University, Baltimore, MD, USA. ⁴⁶Department of Biochemistry and Molecular Biology, University of Kansas Medical Center, Kansas City, KS, USA. ⁴⁷Johnson & Johnson, Beerse, Belgium. ⁴⁸Department of Neurology, Hospital del Mar, Barcelona, Spain. ⁴⁹Department of Neurology, Alzheimer Center Amsterdam, Amsterdam, The Netherlands. ⁵⁰Amsterdam Neuroscience, Amsterdam, The Netherlands. ⁵¹ALS Therapy Development Institute, Cambridge, MA, USA. ⁵²Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands. ⁵³Alzheimer Center Limburg, School for Mental Health and Neuroscience, Maastricht University, Maastricht, The Netherlands. ⁵⁴Laboratory of Behavioral Neuroscience, National Institute on Aging, Bethesda, MD, USA. ⁵⁵Department of Psychiatry, Oxford University, Oxford, UK. ⁵⁶Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Stanford, CA, USA. ⁵⁷Johnson & Johnson, London, UK. *A list of authors and their affiliations appears at the end of the paper.  e-mail: farhad.imam@gatesventures.com

The Global Neurodegeneration Proteomics Consortium (GNPC)

Farhad Imam¹, Rowan Saloner², Jacob W. Vogel³, Varsha Krish¹, Gamal Abdel-Azim⁴, Muhammad Ali^{5,6}, Lijun An³, Federica Anastasi^{7,8,9}, David Bennett¹⁰, Alexa Pichet Binette^{11,12,13}, Adam L. Boxer², Martin Bringmann⁴, Jeffrey M. Burns^{14,15}, Carlos Cruchaga^{5,6,16}, Jeff L. Dage^{17,18}, Amelia Farinas^{19,20,21}, Luigi Ferrucci²², Caitlin A. Finney^{23,24}, Mark Frasier²⁵, Oskar Hansson¹¹, Timothy J. Hohman^{26,27}, Erik C. B. Johnson^{28,29}, Mika Kivimaki^{30,31}, Roxanna Korologou-Linden³², Agustin Ruiz Laza^{33,34,35}, Allan I. Levey^{28,29}, Inga Liepelt-Scarfone^{36,37,38}, Lina Lu¹¹, Niklas Mattsson-Carlgren^{11,39}, Lefkos T. Middleton³², Kwangsik Nho⁴⁰, Hamilton Se-Hwee Oh^{20,21,41}, Ronald C. Petersen⁴², Eric M. Reiman⁴³, Oliver Robinson^{32,44}, Jeffrey D. Rothstein⁴⁵, Andrew J. Saykin^{17,18}, Artur Shvetsov^{23,24}, Chad Slawson^{14,46}, Bart Smets⁴⁷, Marc Suárez-Calvet^{7,8,48}, Betty M. Tijms^{49,50}, Maarten Timmers⁴⁷, Fernando Vieira⁵¹, Natalia Vilor-Tejedor^{7,9,52}, Pieter Jelle Visser^{49,50,53}, Keenan A. Walker⁵⁴, Laura M. Winchester⁵⁵, Tony Wyss-Coray^{20,21,56}, Chengran Yang^{5,6}, Niranjan Bose¹ & Simon Lovestone⁵⁷

A full list of members and their affiliations appears in the Supplementary Information.

Methods

SomaScan proteomics data processing

Proteomic profiling within the GNPC was primarily performed using the SomaScan platform (SomaLogic). Biofluid specimens were independently shipped by each contributing cohort to SomaLogic, with coordination support provided by Gates Ventures. The SomaScan platform uses slow off-rate modified aptamers (SOMAmers) to quantify thousands of proteins in human biofluids, including plasma, serum and CSF. Samples were analyzed within each contributing cohort using versions 3 (~1,300 targets), 4 (~5,000 targets) or 4.1 (~7,000 targets) of the SomaScan assay.

All participating cohorts confirmed that all contributed clinical and generated biosample data were in compliance with the individual patient consents prior to contributing data to the GNPC. Proteomic data from each contributing cohort were processed separately, following SomaLogic's standardized adaptive normalization by maximum likelihood (ANML) pipeline for hybridization normalization, signal calibration and quality control. These procedures adjust for systematic variation using internal reference standards and buffer controls included on each assay plate. Between-sample normalization was performed using median signal intensities and adaptive procedures to reduce batch and run-to-run variation. Samples with signal intensities that substantially deviated from expected ranges are flagged by SomaLogic for quality concerns but were not removed from the dataset. After cohort-level processing and normalization, datasets were combined to form the harmonized GNPC V1 dataset.

At the time of release, we have included 53 clinical variables across all datasets and will continue to increase this number toward a target of over 150 harmonizable variables in the next version of the harmonized dataset. The clinical variables included in this first release of data include demographic data, harmonized cognitive data from research assessments and comorbidity data (Supplementary Table 3).

Clinical and phenotypic data harmonization

To harmonize the clinical metadata across the 23 cohorts, we started with a minimum set of 13 required features that were requested from each cohort: age, gender, years of education, date of visit, diagnosis, date of diagnosis, medication use, comorbidities, vital signs, at least one psychiatric measure, at least one cognitive dementia or functional rating score, at least one cognitive test score and disease-specific genotype data. After initial data contribution, five key vital signs (height, weight, body mass index, resting heart rate and blood pressure) and 14 comorbidities (alcohol use, smoking/tobacco use, stroke, transient ischemic attack, traumatic brain injury, cancer, congestive heart failure, chronic obstructive pulmonary disease, myocardial infarction, atrial fibrillation, angina, hyperlipidemia and hypertension) were identified as common features across at least five cohorts for harmonization. The 19 identified features were then identified, where available, from each of the cohort's data contributions. Demographic and biometric information was normalized to a common scale. Biologically impossible values were cleaned from the dataset. To handle outlier values in years of education, height, weight, body mass index, resting heart rate, blood pressure and total years smoked, an additional variable was created for each feature to indicate if a value was within 2 or 3 s.d. from the mean. Variables were aligned through a mapping schema, ensuring that equivalent tests and demographic categories were matched correctly.

Diagnosis information and control and cognitive impairment data were included, if provided. Diagnosis data for AD, PD, ALS and FTD were captured from the provided clinical data. Due to the variability in AD diagnosis methodology across the 23 cohorts, the method for diagnosis was provided. Using the provided diagnosis data from each site and CDR test scores, a harmonized Clinical Diagnosis variable was created categorizing each participant into one of four categories: Cognitively Normal (CN) (CDR = 0 or confirmed recruited control participant),

MCI (CDR = 0.5), Dementia (CDR ≥ 1 or a confirmed diagnosis of AD or FTD) or Other Neurodegenerative Disease (a confirmed case of PD or ALS). For participants with no formal diagnosis information or CDR reported, cognitive test scores (MMSE or MoCA) were used to categorize an individual's cognitive impairment as Not Impaired (MMSE ≥ 24 and MoCA ≥ 17) or Impaired (MMSE < 24 and MoCA < 17) (Supplementary Tables 2 and 3). Datasets were merged using unique participant identifiers, followed by quality control checks to rectify inconsistencies. Finally, validation was performed by comparing results from the integrated dataset with original study findings.

Vignette approach

Each of the three vignettes presented here was conducted by separate workstreams within the GNPC, each using distinct criteria and methodological frameworks tailored to their specific research questions. These analyses were intentionally designed to highlight the breadth of analytic approaches enabled by the GNPC dataset—from disease-specific differential abundance and transdiagnostic clinical severity modeling to biological aging clocks and genotype-based signatures. As such, they serve as illustrative examples rather than a unified analytic pipeline, showcasing the flexibility and depth of the GNPC V1 resource for diverse scientific inquiries.

Vignette 1 methods: disease-specific differential abundance

Sample selection and inclusion criteria. For AD, PD, FTD and ALS, we used protein data from each participant's first available plasma sample, based on clinical data from the cohorts. To reduce diagnostic ambiguity, we excluded participants with conflicting or overlapping clinical labels. This included individuals labeled with both AD and mild cognitive impairment (MCI)/subjective cognitive impairment (SCI) ($n = 271$) and those assigned two distinct neurodegenerative disease diagnoses ($n = 90$). We opted to include only individuals diagnosed with AD at the dementia stage, thus excluding those in prodromal stages. CN Control participants were included if they were explicitly labeled as Controls in their respective cohorts and/or had a CDR score of 0. We also removed participants missing age or sex data ($n = 54$). After all exclusions, the final dataset comprised 5,879 Controls, 1,966 patients with AD, 607 patients with PD, 175 patients with FTD and 254 patients with ALS (Supplementary Table 4).

Protein data processing. Protein aptamer abundance levels (relative fluorescence units) were \log_2 transformed prior to analysis. Extreme outliers, defined as values more than 5 s.d. above or below the mean across the full dataset, were removed for each aptamer. To ensure biological relevance, only aptamers targeting human proteins were retained for analysis, resulting in a final analytic set of 7,289 unique aptamers (Supplementary Table 5).

Cohort-specific differential abundance and meta-analysis. Given the inherent heterogeneity across contributing GNPC cohorts, all differential abundance analyses were first conducted within each cohort. For each neurodegenerative disease, Patient samples were compared to Control samples using linear regression with the protein used as the dependent variable and disease diagnosis as the independent variable, with age and sex included as covariates. Only cohorts with at least five patients in each diagnostic category (AD, PD, FTD or ALS) were analyzed. In the few cohorts that lacked internal control samples (for example, Cohorts R and G for AD, Cohorts R and T for PD and Cohort S for ALS), patient data were compared to pooled Controls from the remaining cohorts. See Supplementary Table 3 for a breakdown of Controls and Patients across cohorts.

After estimating effect sizes within each cohort, we conducted fixed-effects meta-analyses to identify proteins with reproducible disease associations across cohorts, thereby providing internal replication of top signals. For each aptamer, a weighted average of effect

sizes was calculated, taking into account the sample size of patients at each cohort, and meta-analytic *P*values were computed using the weighted *z*-score method via the ‘metapro R’ package. These *P*values were then adjusted for multiple comparisons using both FDR and Bonferroni correction (separately). To further assess reproducibility, we calculated the number of cohorts in which each aptamer was significantly differentially abundant (FDR < 0.05) as well as the consistency of directionality (that is, the number of cohorts showing concordant upregulation or downregulation for a given target.)

Pathway enrichment analysis. Proteins identified in the meta-analytic differential abundance analysis were filtered into gene set enrichment analysis using the R libraries ‘ReactomePA’ and ‘clusterProfiler’. Proteins were selected based on Bonferroni cutoffs where possible (AD and PD comparisons). For FTD and ALS, where statistical power was more limited, we applied a less stringent threshold of FDR < 0.05. Enrichment analysis was performed with the Reactome library against a background universe consisting of 6,404 unique human proteins measured on the SomaScan 7K platform.

Development of transdiagnostic signature of cognitive decline. Independent of diagnostic category, we also sought to identify a plasma proteomic signature of clinical severity across the GNPc disease continuum by leveraging the subset of Patients with global CDR scales ($N = 6,187$) and cognitive test scores ($N = 5,969$). To model the full spectrum of clinical severity, inclusion criteria were intentionally broadened to include all AD, PD and FTD Patients ranging from CDR = 0.5 to CDR = 3. CDR global scores were modeled based on a five-level ordinal stage from no impairment (CDR = 0) to severe impairment (CDR = 3). A proteome-wide association analysis for CDR global scores was conducted by age (linear and quadratic), sex, their interactions and smoking/alcohol status. Significant proteins (FDR < 0.05) were used to construct a multivariate protein signature via LASSO regression, with data split into 70% training and 30% test sets. Diagnosis-stratified models also tested the multivariate protein signature’s ability to track cognitive severity across each clinical condition. To demonstrate robustness of prediction to orthogonal clinical measures, the CDR-derived protein signature was modeled as a predictor of cognitive performance using cross-walked, harmonized MoCA/MMSE scores (0–30 scale).

Vignette 2 methods: organ aging across diseases

A series of recent studies has generated compelling evidence for the use of molecular ‘clocks’ to estimate biological aging at the organismal and organ-specific level using plasma proteomics. Building upon this foundation, we applied validated organ age models to the GNPc dataset to evaluate disease-specific patterns of accelerated organ aging³².

Sample selection and inclusion criteria. This analysis was restricted to nine GNPc cohorts that included both Patients and CN Control participants, as control data were required to compute normative organ aging estimates. Diagnostic groups included individuals with AD ($n = 1,973$), FTD ($n = 151$) and PD ($n = 334$).

Protein data processing. Protein levels from the SomaScan version 4.1 platform were normalized to match the version 4.0 reference using internal reference-based scaling methods, enabling consistent application of the organ aging models across assays.

Organ age estimation and disease associations. Organ age estimates were computed for brain, liver, kidney, muscle, adipose, immune, lung, intestine, artery, pancreas and heart tissues, using established organ aging models that use proteins specific to each organ. Cognition-enriched brain age estimates were derived using the CognitionBrain model, which additionally limits the brain-specific proteins used for age estimation to only those that are important for

the association of the model age gap with cognitive impairment, as determined by the Feature Importance for Biological Aging (FIBA) algorithm. Additionally, whole-organism (‘organismal’) age estimates, using only proteins common across organs, and ‘conventional’ age estimates, using all measured proteins regardless of tissue enrichment, were also calculated. The organ age gap was derived as the difference between predicted age and the cohort-specific locally weighted scatterplot smoothing (LOWESS) regression estimate for individuals with a normal cognitive clinical diagnosis (Supplementary Table 9). Associations between *z*-score normalized organ age gaps and diagnosis (AD, PD and FTD) were determined using logistic regression models, adjusting for age, sex and cohort.

Vignette 3 methods: *APOE* proteome

The *APOE* ε4 allele is the leading genetic risk factor of late-onset AD, yet its high frequency in patients with AD complicates efforts to disentangle *APOE* from AD-related proteomic changes. To address this, we implemented a stepwise analytic framework to isolate proteomic signatures linked to *APOE* ε4 carriership, irrespective of AD and other neurodegenerative conditions.

Sample selection and inclusion criteria. Participants were selected from GNPc cohorts based on strict diagnostic and cognitive performance criteria to ensure well-defined AD and Control groups. We first identified individuals labeled with an AD dementia diagnosis or recruited as Controls. Participants with multiple neurodegenerative disease diagnoses were excluded. Patients with AD dementia were required to have either MMSE < 24 or MoCA < 17 and CDR ≥ 1, to ensure alignment with established diagnostic thresholds for dementia. CN participants did not meet criteria for any clinical diagnosis (for example, AD, MCI, FTD, ALS and PD) and had CDR = 0 and, at a minimum, cognitive test scores above a dementia threshold (MMSE ≥ 24 and/or MoCA ≥ 17) (Supplementary Table 10). To minimize confounding from opposing genotype effects, *APOE* ε2/ε4 heterozygotes were excluded from the analysis.

After applying the above criteria, a total of 3,934 participants were identified for further analysis. Among them, $n = 2,357$ were *APOE* ε4 non-carriers (ε4⁻) and $n = 1,577$ were *APOE* ε4 carriers (ε4⁺, either ε3/ε4 or ε4/ε4). At baseline, 1,438 individuals were diagnosed with AD, and the remaining 2,496 were non-impaired (CN) Controls.

***APOE* ε4 association analysis.** Consistent with differential abundance analyses (see Vignette 1 methods), log₂-transformed aptamer values that deviated more than 5 s.d. from the mean were defined as outliers and excluded. Proteins associated with *APOE* ε4 carriers (carrier versus non-carrier) were identified using linear regression models, adjusting for AD dementia diagnosis (CN versus AD dementia diagnosis), age, sex, contributing cohort code and mean overall protein level (Supplementary Table 11). Multiple testing correction was performed using FDR adjustment (Benjamini–Hochberg method, $\alpha = 0.05$). Residualized protein values were used for visualization. The Wilcoxon rank-sum test in the ‘ggpubr’ R package (version 0.6.0) was used to test and mark the changes in protein residuals in the *APOE* ε4 carrier group and the AD dementia diagnosis group (Supplementary Table 12)³³.

Machine learning classification of *APOE* ε4 status. To identify a disease-agnostic proteomic signature of *APOE* ε4 allele status, a machine learning classification model was developed. Proteomic data were split into training (70%) and test (30%) sets, with standardization performed independently within each set. A classification and regression trees (CART) approach was applied to the data using cross-validation and fine tuning. All final metrics reported were derived from application of the model to the left-out testing dataset.

Model performance was stratified by CN and AD as well as FTD, PD and Parkinson’s disease dementia. First-order functional networks

were built from proteins of interest identified by the CART model. In addition, brain cell subtype-specific enrichment analysis was performed using single-nuclei brain RNA sequencing data from the Human Protein Atlas.

Overlap of APOE-associated and AD-associated proteins. To explore proteins influenced independently by *APOE ε4* carriership and AD diagnosis, two additional comparisons were performed. First, proteins associated with *APOE ε4* in cognitively unimpaired individuals were identified. Second, proteins associated with AD compared to Controls were identified specifically in *APOE ε3* homozygote AD patients. Analyses were performed using limma (limma_3.62.2) and adjusted for sex, age at visit and contributor code. Multiple testing correction was performed using FDR adjustment (Benjamini–Hochberg method, $\alpha = 0.05$).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The harmonized GNPC data used to generate these findings were provided to consortium members in June 2024 and will be made available for public request by the AD Data Initiative in July 2025. Members of the global research community will be able to access the metadata and place a data use request via the AD Discovery Portal (<https://discover.alzheimersdata.org/>). Access is contingent upon adherence to the GNPC Data Use Agreement and the Publication Policies.

Code availability

All custom code and data that were employed for this paper are saved within the GNPC AD Workbench. All custom code central to the findings in this paper will be made available outside of private workspaces upon publication.

References

32. Oh, H. Organ aging signatures in the plasma proteome track health and disease. *Nature* **624**, 164–172 (2023).
33. Shvetcov, A. et al. Proteome profiling of cerebrospinal fluid using machine learning shows a unique protein signature associated with APOE4 genotype. *Aging Cell* **24**, e14439 (2025).

Acknowledgements

We thank all GNPC V1 consortium members who helped support the creation of the V1 harmonized dataset. We thank the participants, their families as well as the cohorts, institutions and their dedicated staff who helped create this consortium. Funding for the majority of biosample analyses was provided by Gates Ventures and Johnson & Johnson. Funding for cohort creation, participant biosample collection and additional biosample analysis was provided by the following sources. ALLFTD data collection and dissemination was supported by the ALLFTD Consortium (U19: AGO63911, funded by the National Institute on Aging (NIA) and the National Institute of Neurological Disorders and Stroke (NINDS)) and the former ARTFL & LEFFTDS Consortia (ARTFL: U54 NS092089, funded by NINDS and the National Center for Advancing Translational Sciences; LEFFTDS: U01 AG045390, funded by the NIA and the NINDS). Additional funding support was provided by the National Institutes of Health (NIH)-NIA for Frontotemporal Dementia: Genes, Images, and Emotions: P01 AG019724 and the UCSF Alzheimer's Disease Research Center: P30 AG062422. Samples from the National Centralized Repository for Alzheimer's Disease and Related Dementias (NCRAD), which receives government support under a cooperative agreement grant (U24 AG21886) awarded by the NIA, were used in this study. The ALLFTD proteomic data included in this paper were funded by a grant from

Gates Ventures. Alzheimer and Families (ALFA) was supported by funding from the MCIN/AEI/10.13039/501100011033/FEDER, European Union, through the PID2023-153312OB-I00 project; the CaixaResearch Institute under the project titled DETEC-PK; and the RYC2023-043831-I grant, funded by MCIN/AEI/10.13039/501100011033 and the FSE+. Additional researchers receive funding from the JDC2022-049347-I grant, funded by MCIU/AEI/10.13039/501100011033, the European Union NextGenerationEU/PRTR, the Spanish Research Agency MCIU/AEI/10.13039/501100011033 (grant RYC2022-038136-I co-funded by the European Union FSE+, grant PID2022-143106OA-I00 co-funded by the European Union FEDER), the William H. Gates Sr. Fellowship from the Alzheimer's Disease Data Initiative, the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 948677), ERA PerMed-ERA NET and the Generalitat de Catalunya (Departament de Salut) through project SLD077/21/000001, Project 'PI19/00155' and 'PI22/00456', funded by Instituto de Salud Carlos III (ISCIII) and co-funded by the European Union, and from a fellowship from 'la Caixa' Foundation (ID 100010434) and from the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie grant agreement no. 847648 (LCF/BQ/PR21/11840004). The ALFA proteomic data included in this paper were funded by a grant from Gates Ventures. Stichting Alzheimer Nederland and Stichting VUmc provided funds for supporting Alzheimer Center Amsterdam infrastructure. Additionally, ZonMw Memorabel program project 'PRODIA', as part of the Deltaplan Dementie, and 2bike4alzheimer (Alzheimer Nederland) are acknowledged for funding this study for the generation of proteomics data for Amsterdam UMC. The baseline ABC-PD study (study period 2014–2017) was funded by a pre-competitive grant from Janssen Research and Development, a division of Janssen Pharmaceutica N.V. The follow-up study (study period 2018–2020) was funded by a grant from the Michael J. Fox Foundation for Parkinson's Research (grant no. 15227). Neither study sponsor was involved in collection, analysis and interpretation of data, writing of the report or decision to submit the manuscript for publication. Janssen Research and Development was only involved in study design of the baseline visit. Answer ALS was supported by NEALS Biofluid Repository, and Answer ALS collected, annotated and granted permission to use ALS biofluid samples. The proteomic data included in this paper were funded by a grant from Gates Ventures. The Baltimore Longitudinal Study of Aging (BLSA) was supported by the Intramural Research Program (IRP) of the NIH-NIA. Banner Health was supported by Arizona ADRC and Arizona APOE grants from the NIA (P30 AG019610 and R01 AG069453), the state of Arizona, the Banner Alzheimer's Foundation, the Sun Health Foundation, the Michael J. Fox Foundation and Gates Ventures. Work at Lund University was supported by the NIA (R01AG083740), the ERC (ADG-101096455), the Alzheimer's Association (ZEN24-1069572 and SG-23-1061717), the GHR Foundation, the Swedish Research Council (2022-00775 and 2021-02219), ERA PerMed (ERAPERMED2021-184), the Knut and Alice Wallenberg Foundation (2022-0231), Strategic Research Area MultiPark (Multidisciplinary Research in Parkinson's Disease) at Lund University, the Swedish Alzheimer Foundation (AF-980907 and AF-994229), the Swedish Brain Foundation (FO2021-0293 and FO2023-0163), the Parkinson Foundation of Sweden (1412/22), the Cure Alzheimer's Fund, the Rönström Family Foundation, Konung Gustaf V:s och Drottning Victorias Frimurarestiftelse, the Skåne University Hospital Foundation (2020-O000028), the WASP and DDLS joint call for research projects (WASP/DDLS22-066), the Lilly Research Award Program, Regionalt Forskningsstöd (2022-1259) and the Swedish federal government under the ALF agreement (2022-Projekt0080 and 2022-Projekt0107). Additionally, this work was supported by the SciLifeLab & Wallenberg Data Driven Life Science Program (grant no. KAW 2020.0239), the Swedish Research Council (2024-03642), the NIA (U01 AG079847-02) and Gates Ventures. The CHARIOT-PRO study

was funded by Johnson & Johnson, Merck, Takeda and Gates Ventures, and analysis was supported by the UKRI Future Leaders Fellowship (MR/S03532X/1 and MR/Y02012X/1) and the NIHR ARC Dementia Fellowship. The Emory ADRC participation was supported by the Emory Goizueta Alzheimer's Research Disease Center (P50AG025688) and grants from Gates Ventures for the generation of proteomics data, and data collection was supported through funding by NIA grants R01AG053960, R01AG057911, R01AG061800, RF1AG057471, RF1AG057470, R01AG061800, R01AG057911 and R01AG057339. The Kansas ADRC participation was supported by the NIH (P30 AG072973, P30 AG072974 and R01 AG064227) and further supported by the Neil & Norma Hill Foundation, the Annemarie & Arturo Gondioli-Fumagalli Foundation, the Perpetual Foundation, the John Williams Endowment and the Hillcrest Foundation. The proteomic data included in this paper were funded by a grant from Gates Ventures. The Knight ADRC participation was supported by grants from the NIH (R01AG044546, RF1AG053303, RF1AG058501 and U01AG058922), the Michael J. Fox Foundation, the Alzheimer's Association Zenith Fellows Award (ZEN-22-848604), Gates Ventures and an anonymous foundation. The recruitment and clinical characterization of research participants at Washington University were supported by NIH P30AG066444, P01AG03991 and P01AG026276. Additionally, this work was supported by access to equipment made possible by the Hope Center for Neurological Disorders, the Neurogenomics and Informatics Center (<https://neurogenomics.wustl.edu/>) and the Departments of Neurology and Psychiatry at Washington University School of Medicine. The Whitehall II study was supported by the Wellcome Trust (221854/Z/20/Z), the UK Medical Research Council (Y014154, R024227, S011676 and K013351) and the NIA (R01AG056477 and R01AG062553).

Author contributions

All authors contributed substantially to the development of the GNPC V1 harmonized dataset. S.L. and N.B. envisaged and initiated the GNPC as a public–private consortium and, together with F.I., V.K., M.T. and M.B., guided the organization and operation of the consortium. R.S., J.W.V., V.K., F.I. and S.L. were responsible for the writing and consolidation of vignette analysis into the overarching manuscript. T.J.H., E.C.B.J. and P.J.V. advised on data harmonization across the multiple disease types. Specific vignette analyses were provided by G.A.-A., M.A., L.A., A.L.B., A.F., C.A.F., T.J.H., R.K.-L., L.L., N.M.-C., A.S. and B.S. All authors contributed to review and refinement of the manuscript.

Competing interests

A.L.B. receives research support from the NIH, the Tau Research Consortium, the Association for Frontotemporal Degeneration,

Bluefield Project to Cure Frontotemporal Dementia, the GHR Foundation and the Alzheimer's Association. He has been a consultant for Alchemab, Alector, Alexion, Amylyx, Arrowhead, Arvinas, Eli Lilly, Muna, Neurocrine, Ono, Oscotec, Pfizer, Switch, Transposon and UnlearnAI. C.C. has received research support from GSK and Eisai. C.C. is a member of the scientific advisory board of Circular Genomics and owns stocks. C.C. is a member of the scientific advisory board of ADmit. J.L.D. has a patent pending for compounds and methods targeting human tau. L.F. has given unpaid seminars and/or webinars sponsored or co-sponsored by SomaLogic. O.H. has received nonfinancial support from Roache and Lilly and is currently employed by Lilly. E.M.R. has received grants from National Institute on Aging and the state of Arizona; receives philanthropic funding from the Banner Alzheimer's Foundation, Sun Health Foundation and Roche/Roche Diagnostics; receives personal fees from Alkahest, Alzheon, Aural Analytics, Denali, Green Valley, MagQ, Takeda/Zinfandel and United Neuroscience; has since submission of manuscript become a cofounder of ALZpath, which aims to further develop P-tau217 and fluid biomarkers and advance their use in research, drug development and clinical settings; holds a patent owned by Banner Health for a strategy to use biomarkers to accelerate evaluation of Alzheimer prevention therapies; and is a principal investigator of prevention trials that include research agreements with Genentech/Roche and Novartis/Amgen, PET studies that include research agreements with Avid/Lilly and several NIH and Foundation-supported research studies. T.W.-C. and H.S.-H.O. are co-founders and scientific advisors of Teal Omics Inc. and have received equity stakes. T.W.-C. is a co-founder and scientific advisor of Alkahest Inc. and Qinotto Inc. and has received equity stakes in these companies. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41591-025-03834-0>.

Correspondence and requests for materials should be addressed to Farhad Imam.

Peer review information *Nature Medicine* thanks Boris Decourt, Jin-Tai Yu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Jerome Staal, Saheli Sadanand and Liam Messin in collaboration with the *Nature Medicine* team.

Reprints and permissions information is available at www.nature.com/reprints.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection *The GNPC harmonized dataset exists in a cloud-resident workspace in the AD Workbench which reviewers will be provided access to the complete set of harmonized and anonymized clinical and proteomic data.*

Data analysis *The code for each analysis included in the manuscript is saved within the GNPC AD Workspace*

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

All custom code and data that was employed for this manuscript is saved within the GNPC AD Workbench. The harmonized GNPC data used to generate these findings was provided to Consortium Members in June 2024 and will be made available for public request by the AD Data Initiative in July, 2025. Members of the global research community will be able to access the metadata and place a data use request via the AD Discovery Portal (<https://discover.alzheimersdata.org/>).

Access is contingent on adherence to the GNPC Data Use Agreement and the Publication Policies
The Human Protein Atlas (<https://www.proteinatlas.org/>) was used for protein-protein interaction libraries and single brain cell RNA seq data.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

Neither sex nor gender are utilized in individual analysis. The harmonized dataset reports out on sex as part of the summary demographics from the HDS (Supp. Table 2)

Reporting on race, ethnicity, or other socially relevant groupings

Race, ethnicity, or other socially relevant groupings are utilized in individual analysis. The harmonized dataset reports out on race and ethnicity as part of the summary demographics from the HDS (Supp. Table 2). The features shared include the standard US Census race/ethnicity groupings and UDS-3 coding groupings.

Population characteristics

27 population characteristics were collated from the 23 participating cohorts including common demographic features, vital sign data, and co-morbidity data. Age > 90 was capped and any characteristic with fewer than 10 participants would be hidden from the Harmonized dataset. No individual feature had fewer than 10 participants represented.

Recruitment

No individual participants were recruited for the GNPC HDS. Existing published datasets were harmonized to create the HS.

Ethics oversight

No new participants were recruited for the GNPC HDS. All individual published studies are listed in Table 1. Each participating cohort confirmed ethical compliance with the signed consents prior to participation in the GNPC per the terms of the GNPC Consortium Agreement.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences

Behavioural & social sciences

Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Sample selection for the 3 analyses were chosen to maximize data availability for each approach while maintaining consistent proteomic assay type or sample matrix. The maximum number of samples that would be available for given exclusion criteria.

Data exclusions

Data were excluded from specific analyses if the features did not align with the intended goal of the analysis (e.g. in Vignette 3 samples without APOE genotype data were excluded from the APOE association analysis). No data were excluded from the Harmonized Dataset if the features matched the Harmonized Dataset schema.

Replication

To ensure replicability of these findings, the dataset will be provided to the public on July 1 and all methods for the analyses have been thoroughly described in the manuscript. Each of the experiments were performed independently.

Randomization

The generated proteomics data was contributed to the GNPC for harmonization from 23 groups. Analyses considered potential batch-to-batch variation and cross-cohort variability as the total set of samples were not available to randomize prior to data generation. The data contributed in came from each individual cohort.

Blinding

Investigators are blinded to which sample data IDs link back to original sample collection. Each sample provided in was pre-blinded by the contributing members and only identified through a unique ID.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).

Research sample

State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.

Sampling strategy

Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a

	<i>rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.</i>
Data collection	<i>Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.</i>
Timing	<i>Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.</i>
Data exclusions	<i>If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.</i>
Non-participation	<i>State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.</i>
Randomization	<i>If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.</i>

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	<i>Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.</i>
Research sample	<i>Describe the research sample (e.g. a group of tagged <i>Passer domesticus</i>, all <i>Stenocereus thurberi</i> within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.</i>
Sampling strategy	<i>Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.</i>
Data collection	<i>Describe the data collection procedure, including who recorded the data and how.</i>
Timing and spatial scale	<i>Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken</i>
Data exclusions	<i>If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.</i>
Reproducibility	<i>Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.</i>
Randomization	<i>Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.</i>
Blinding	<i>Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.</i>

Did the study involve field work? Yes No

Field work, collection and transport

Field conditions	<i>Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).</i>
Location	<i>State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).</i>
Access & import/export	<i>Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).</i>
Disturbance	<i>Describe any disturbance caused by the study and how it was minimized.</i>

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	Antibodies
<input checked="" type="checkbox"/>	Eukaryotic cell lines
<input checked="" type="checkbox"/>	Palaeontology and archaeology
<input checked="" type="checkbox"/>	Animals and other organisms
<input checked="" type="checkbox"/>	Clinical data
<input checked="" type="checkbox"/>	Dual use research of concern
<input checked="" type="checkbox"/>	Plants

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	ChIP-seq
<input checked="" type="checkbox"/>	Flow cytometry
<input checked="" type="checkbox"/>	MRI-based neuroimaging

Antibodies

Antibodies used

Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.

Validation

Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)

State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.

Authentication

Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.

Mycoplasma contamination

Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.

Commonly misidentified lines (See [ICLAC](#) register)

Name any commonly misidentified cell lines used in the study and provide a rationale for their use.

Palaeontology and Archaeology

Specimen provenance

Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.

Specimen deposition

Indicate where the specimens have been deposited to permit free access by other researchers.

Dating methods

If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.

Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about [studies involving animals; ARRIVE guidelines](#) recommended for reporting animal research, and [Sex and Gender in Research](#)

Laboratory animals

For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.

Wild animals

Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.

Reporting on sex

Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.

Field-collected samples

For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration

Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.

Study protocol

Note where the full trial protocol can be accessed OR if not available, explain why.

Data collection

Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.

Outcomes

Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.

Dual use research of concern

Policy information about [dual use research of concern](#)

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

- | | |
|-------------------------------------|---|
| No | Yes |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Public health |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> National security |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Crops and/or livestock |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Ecosystems |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Any other significant area |

Experiments of concern

Does the work involve any of these experiments of concern:

- | | |
|-------------------------------------|--|
| No | Yes |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Demonstrate how to render a vaccine ineffective |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Confer resistance to therapeutically useful antibiotics or antiviral agents |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Enhance the virulence of a pathogen or render a nonpathogen virulent |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Increase transmissibility of a pathogen |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Alter the host range of a pathogen |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Enable evasion of diagnostic/detection modalities |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Enable the weaponization of a biological agent or toxin |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Any other potentially harmful combination of experiments and agents |

Plants

Seed stocks	Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.
Novel plant genotypes	Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.
Authentication	Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.

ChIP-seq

Data deposition

- Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links

May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session (e.g. [UCSC](#))

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.

Instrument

Identify the instrument used for data collection, specifying make and model number.

Software

Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.

Cell population abundance

Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.

Gating strategy

Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.

Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type

Indicate task or resting state; event-related or block design.

Design specifications

Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.

Behavioral performance measures

State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).

Acquisition

Imaging type(s)

Specify: functional, structural, diffusion, perfusion.

Field strength

Specify in Tesla

Sequence & imaging parameters

Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.

Area of acquisition

State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.

Diffusion MRI

Used

Not used

Preprocessing

Preprocessing software

Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).

Normalization

If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.

Normalization template

Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.

Noise and artifact removal

Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: Whole brain ROI-based Both

Statistic type for inference

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

(See [Eklund et al. 2016](#))

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a Involved in the study

Functional and/or effective connectivity

Graph analysis

Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Differential abundance analyses (FDR adjusted p-value and Bonferroni correction).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

LASSO models and LOWESS regression estimates