Benchmarking the AI-based diagnostic potential of plasma proteomics for neurodegenerative disease in 17,170 people

Lijun An[1,*], Alexa Pichet-Binette[2,3,4], Ines Hristovska[2], Gabriele Vilkaite[1], Xiao Yu[1], Bart Smets[5], Rowan Saloner[6], Shinya Tasaki[7], Ying Xu[8,9], Varsha Krish[10], Farhad Imam[10], Shorena Janelidze[2], Danielle van Westen[11,12], the Global Neurodegeneration Proteomics Consortium (GNPC), Erik Stomrud[2,13], Christopher D. Whelan[14], Sebastian Palmqvist[2,13], Rik Ossenkoppele[2, 15,16], Niklas Mattsson-Carlgren[2,13], Oskar Hansson[2], Jacob W. Vogel[1,*]

[1] Department of Clinical Sciences Malmö, SciLifeLab, Lund University, Lund, Sweden;
[2] Clinical Memory Research Unit, Department of Clinical Sciences Malmö, Lund University, Lund, Sweden;
[3] Department of Physiology and Pharmacology, Université de Montréal, Montréal, QC, Canada;
[4] Centre de recherche de l'institut universitaire de gériatrie de Montréal (CRIUGM), Montréal, QC, Canada;
[5] Janssen Pharmaceutica NV, a Johnson & Johnson company, Beerse, Belgium;
[6] Memory and Aging Center, Department of Neurology, Weill Institute for Neurosciences, University of California, San Francisco, CA, USA;
[7] Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL, USA;
[8] Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, USA;
[9] NeuroGenomics and Informatics Center, Washington University School of Medicine, St. Louis, MO, USA;
[10] Gates Ventures, Seattle, WA, USA;
[11] Department of Diagnostic Radiology, Clinical Sciences, Lund University, Lund, Sweden;
[12] Image and Function, Skåne University Hospital, Lund, Sweden;
[13] Memory Clinic, Skåne University Hospital, Malmö, Sweden;
[14] Neuroscience Data Science, Janssen Research & Development, Cambridge, MA, USA;
[15] Amsterdam Neuroscience, Neurodegeneration, Vrije Universiteit Amsterdam, Amsterdam, Netherlands;
[16] Alzheimer Center Amsterdam, Neurology, Vrije Universiteit Amsterdam, Amsterdam UMC location VUmc, Amsterdam, Netherlands.

*Corresponding Authors: Jacob W. Vogel (jacob.vogel@med.lu.se) & Lijun An (lijun.an@med.lu.se)

Abstract

Co-pathology is a common feature of neurodegenerative diseases that complicates diagnosis, treatment and clinical management. However, sensitive, specific and scalable biomarkers for in vivo pathological diagnosis are not available for most neurodegenerative neuropathologies. Here, we present ProtAIDe-Dx, a deep joint-learning model trained on 17,170 patients and controls that uses plasma proteomics to provide simultaneous probabilistic diagnosis across six conditions associated with dementia in aging. ProtAIDe-Dx achieves cross-validated balanced classification accuracy of 69%-96% and AUCs > 79% across all conditions. The model's diagnostic probabilities highlighted subgroups of patients with co-pathologies, and were associated with pathology-specific biomarkers in an external sample, even among cognitively unimpaired people. Model interpretation revealed a suite of protein networks marking shared and specific biological processes across diseases, and identified novel and previously described proteins discriminating each diagnosis. ProtAIDe-Dx significantly improved biomarker-based differential diagnosis in a memory clinic sample, pinpointing proteins leading to diagnostic decisions at an individual level. Together, this work highlights the promise of plasma proteomics to improve patient-level diagnostic work-up with a single blood draw.

Introduction

The last five years have seen multiple breakthroughs in the treatment of neurodegenerative diseases. Early disease modifying therapies have emerged for Alzheimer's disease (AD) [1,2], and highly promising drug candidates are currently in clinical trials for AD [3], Parkinson's disease (PD) [4] and amyotrophic lateral sclerosis (ALS) [5]. However, differential diagnosis and disease comorbidity continue to pose considerable challenges in these treatment efforts. Misdiagnosis rates are around 25-30% even in specialized dementia clinics, and can exceed 50% in primary care [6–8]. Meanwhile, comorbidity is common in aging, with 70% of patients 80 years or older harboring multiple neurodegenerative pathologies simultaneously [9]. Misdiagnosis can make it difficult to select the right patients for a drug trial [10], while comorbid neuropathologies can mask the positive effects of a putative therapy [11,12]. Once such treatments do become available, misdiagnosis and comorbidity can both lead to treatment mismanagement [13]. With the rapid pace at which promising new drugs are being tested, there is an urgent need for powerful tools for diagnosis and precise identification of underlying comorbid pathologies.

The first step toward mitigating the issues presented by misdiagnosis and comorbidity is the development of biomarkers to identify underlying neurodegenerative pathology with high specificity. As an example, highly sensitive and specific fluid [14–18] and imaging [19,20] biomarkers for AD have become instrumental in the recent development of disease modifying therapies. AD diagnostic capabilities have expanded even further as sensitive tests for underlying AD pathology are now possible using only blood plasma [21,22]. Blood-based biomarkers have the potential to be highly accessible, inexpensive and minimally invasive, and the emergence of blood-based biomarkers for AD could facilitate accurate AD diagnosis even in primary care in the near future [7]. Despite the success of AD biomarkers, scalable, sensitive and specific biomarkers for other neurodegenerative diseases are lacking, and diagnosis can only be made with high confidence at autopsy.

Given the diversity of underlying diseases that patients commonly present with at memory clinics [23,24], and given the substantial effects of comorbid pathology on prognosis [25,26], a multi-disease diagnostic test would be of great benefit to researchers, clinical neurologists and their patients. Plasma proteomics represent a promising tool toward this aim, allowing robust surveillance of thousands of potential biomarkers and relevant functional effectors with a single blood draw [27–29]. While still a new and rapidly evolving tool, high-throughput plasma proteomics has already shown promise in the simultaneous prediction of multiple health outcomes [30–32] and in the monitoring of the integrity of multiple organs and critical biological systems [33]. Beyond their capacity as potential biomarkers, circulating proteins further provide a view into complex molecular dynamics as they play out in living humans, providing novel biological insight into disease processes [34–37].

Despite great promise, plasma proteomics data is not without its challenges. Proteomics data is high rank [38], comes with burdensome technological artifacts [39,40], and likely represents complex nonlinear interactions [41]. In addition, the blood-brain barrier limits the number of brain-expressed proteins relevant in neurological conditions that can be detected in blood. Sophisticated modeling techniques such as artificial intelligence (AI) will likely be needed to synthesize clinically useful disease signatures from this rich and complex data [42]. Although some studies have leveraged population-based cohorts [43–46], the potential of AI-proteomics has not yet been benchmarked in large neurodegenerative cohorts. In the present study, we attempt to overcome these limitations by applying AI to the Global Neurodegeneration Proteomics Consortium (GNPC) v1.3 dataset, the largest neurodegenerative disease plasma proteomics dataset to date. We present a model called Proteomics-based AI for Dementia Diagnosis (ProtAIDe-Dx), a deep multi-task architecture to resolve differential and multiple neurodegenerative diagnoses with a single blood draw (Fig. 1A). ProtAIDe-Dx is based

on a joint-learning framework that simultaneously resolves multiple diagnostic tasks, thus learning shared signatures across diseases from a reduced set of disease-relevant proteins. ProtAIDe-Dx provides probabilistic diagnostic information across five of the most common incapacitating age-related neurological conditions in old age – AD, PD, neurovascular disease (represented here as stroke), ALS and frontotemporal dementia (FTD). Diagnostic predictions are based on nonlinear proteomic embeddings that can be extracted and used for few-shot transfer learning to related tasks and unseen datasets. We report the performance of ProtAIDe-Dx on multi-diagnosis prediction, evaluate its potential in identifying co-pathology, test its capability in differential diagnosis compared to other clinical markers, and explore putative molecular networks contributing to its predictions. Finally, we present a proof-of-concept using ProtAIDe-Dx for personalized and interpretable diagnostic testing in a clinical scenario. With this approach, we hope to set a benchmark for future plasma based multi-disease neurodegenerative diagnostic tools.

Results

A subsample of 17,170 participants with SomaLogic 7k proteomics, sampled across 19 contributing sites, was selected from the GNPC v1.3 dataset for subsequent analysis (see Methods). Table S1 describes sample and demographics across each contributing site, while Table S2 lists the frequency of six conditions across each site, namely AD, PD, FTD, ALS, previous stroke/transient ischemic attack (TIA), and cognitively unimpaired. The stroke/TIA group was chosen as a representation of patients with documented cerebrovascular disease. The total sample summed to 1,095 recruited controls and 5,399 patients. Of the patients, 1,889 were diagnosed with AD, 2,284 were diagnosed with PD, 435 with ALS, 176 with FTD and 335 with stroke/TIA. 280 patients had two or more diagnoses. 10,676 participants were neither diagnosed as Control, nor with any of the five above mentioned diseases, which were included in the “negative” group for each condition (see Methods). We applied the ProtAIDe-Dx model to this sample, an architecture capable of generating several features of interest simultaneously: binary diagnosis of each condition, probabilities of each diagnosis, and joint embeddings representing low-dimensional, non-linear protein combinations used by the model for diagnosis. Fig. 1A describes our overall workflow and how each of these model features is used. We specifically chose a multi-task, joint learning approach (as opposed to a multi-class classification task) to allow the model to signal disease co-pathology (i.e. positive for multiple disorders and probabilities for each disorder).

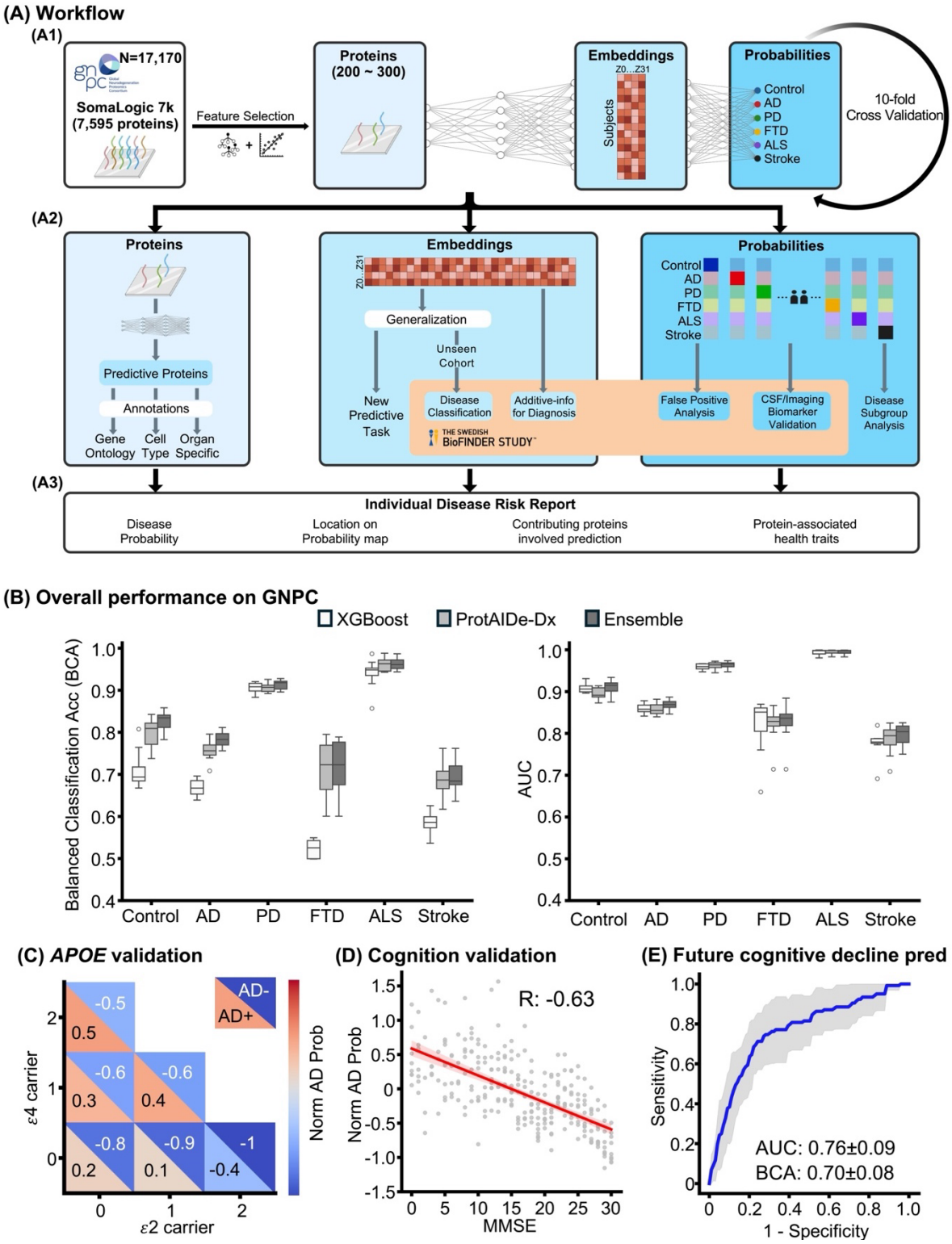


Figure 1. Workflow and overall performances of ProtAIDe-Dx on GNPC. (A) Project workflow. (A1) Model development on GNPC with 10-fold cross-validation procedure. (A2) Model evaluation on GNPC (blue panels) and BioFINDER-2 cohort (orange panel). (A3) Individual neurological disease risk report based on developed model. (B) Overall model performances on GNPC. Left: Balanced classification accuracy (BCA) scores for each diagnostic task; Right: AUC score for each diagnostic task. (C) Normalized AD probabilities stratified by different *APOE* ε2/ε4 groups, shown in participants with (lower diagonal) and without (upper diagonal) AD diagnosis. (D) Correlation between normalized AD probabilities and MMSE. (E) ROC curve of model generalization to a new task for predicting longitudinal clinical progression.

Joint learning improves multi-diagnostic prediction of neurodegenerative diagnosis from blood in unbalanced samples.

We applied ProtAIDe-Dx to the GNPC sample, using 10-fold cross-validation stratified for each contributing site. Importantly, we used only proteomic information in the model – no site, demographic, cognitive or diagnostic information was used. We compared the diagnostic performance of ProtAIDe-Dx to XGBoost, a classical but high-performing [47] machine learning model trained separately on each task. Given that tree-based models have been shown to match deep learning models when trained on tabular data [48], XGBoost performance provides a baseline indicating whether our deep, joint architecture improves diagnostic performance. Due their highly disparate approaches, we anticipated that ProtAIDe-Dx and XGBoost might employ different classification strategies, and therefore might provide complementary information in an ensemble framework. To test this hypothesis, we also tested an ensemble model including aspects of both models.

ProtAIDe-Dx achieved a median balanced accuracy performance above 90% for ALS (96%) and PD (91%) classification, 81% for Control, 78% for AD, 72% for FTD, and 69% for stroke/TIA (Fig. 1B). This performance significantly outperformed that of the XGBoost baseline model for all tasks except ALS and PD prediction (Table S3, SuppData 1). Both models reached AUCs >0.8 for all tasks except for stroke/TIA prediction and showed comparable performance (Fig. 1B, Table S3, SuppData 1), though AUC is notably a less robust performance indicator for imbalanced data. We found that the ensemble model produced balanced accuracy scores and AUC that significantly outperformed ProtAIDe-Dx for AD diagnosis. The ensemble model also significantly improved balanced accuracy scores over XGBoost for all tasks except ALS and PD prediction (Fig. 1B, Table S3, SuppData 1). In all, AI-based proteomic models showed impressive multi-disease diagnostic performance within the GNPC, and the deep, joint-learning framework provided by ProtAIDe-Dx afforded performance advantages over a canonical tree-based, single-task baseline.

As a sanity check, we extracted the probability of AD diagnosis across all individuals and compared them to factors known to be altered in AD, namely cognitive performance and *APOE* ε2 or ε4 allele carriage. In participants both with and without an AD diagnosis, presence of more copies of the *APOE* ε4 allele was associated with higher AD probabilities, and more copies of the ε2 allele was associated with lower AD probabilities (Fig. 1C). Additionally, a strong negative correlation was observed between AD probabilities and MMSE (mini-mental status examination) score – a measurement of global cognition – indicating worse cognition associated with higher AD probabilities (Fig. 1D). These analyses suggest model-derived diagnostic probabilities can serve as continuous proteomic scores associated with indicators of disease progression. We further test this hypothesis in an external dataset with complete biomarkers below.

Diagnostic prediction model generalizes to new disease-relevant tasks.

Low-dimensional nonlinear proteomic embeddings were extracted from the last layer of the ProtAIDe-Dx model (Fig. 1A). Given ProtAIDe-Dx's joint-learning architecture, these embeddings should represent a compressed representation of plasma proteomic data optimized toward tasks related to neurological diseases. To test this hypothesis, we used these embeddings to generalize the ProtAIDe-Dx model to a task that ProtAIDe-Dx was not trained specifically for, namely prediction of longitudinal clinical progression in healthy controls (see Methods). The performance of this model in predicting individuals who progressed from unimpaired to impaired (i.e. from clinical dementia rating [CDR] 0 to CDR 0.5 or 1; N=107), compared to those that remained stably at CDR 0 over time (N=1,945), reached a balanced classification accuracy of 70% and an AUC of 76% (Fig. 1E). Notably, this model generalization was achieved simply by transferring the embeddings with a logistic

regression model, a process that requires minimal CPU power and does not require any GPUs. These results support ProtAIDe-Dx as a flexible and extensible model for neurodegenerative disease-related tasks.

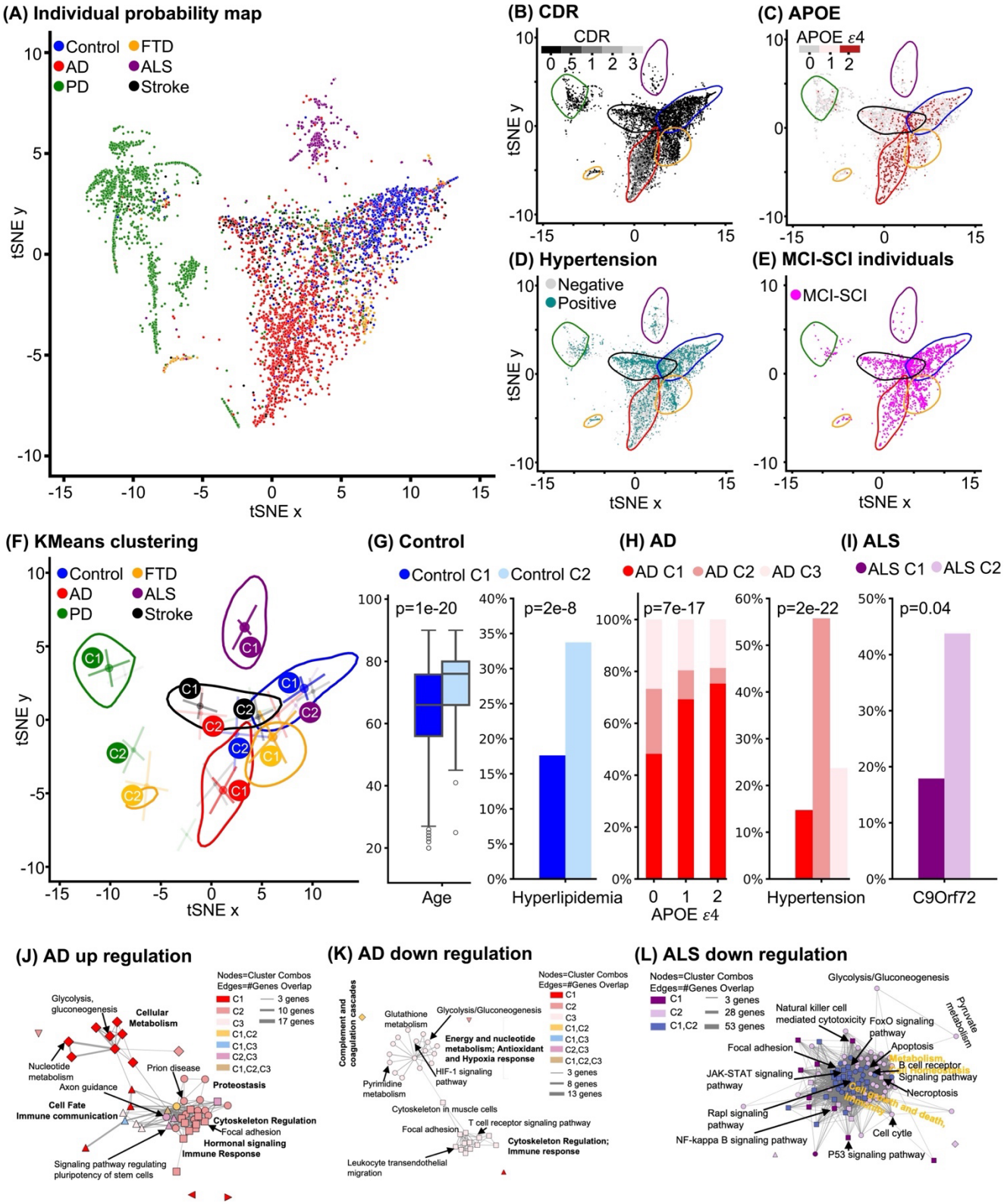


Figure 2. **Diagnostic probability map derived by ProtAIDe-Dx reveal disease heterogeneity**. (A) Participants from the test set were selected for tSNE to reduce dimensionality from six to two, colored by clinical diagnosis. (B) - (E) Participants were projected onto the fitted 2-dimensional diagnostic probability map, colored by variable values. (F) KMeans clustering within each diagnostic category, where the top two clusters were annotated. The contour was drawn based on Gaussian kernel density estimation with a threshold of 0.01. (G) - (I) Distribution of variables varied across the Control, AD, and ALS clusters. (J) - (L) KEGG

terms (pFDR < 0.05) enriched among up- or down-regulated proteins in AD and ALS clusters. Node shape denotes KEGG community; color indicates cluster. All P values were FDR corrected.

Diagnostic probabilities reveal disease heterogeneity and co-pathology.

ProtAIDe-Dx provides probabilities of each condition for each individual. To visualize the overall landscape of how individuals are distributed based on probabilities, we projected all individuals into a two-dimensional nonlinear embedding based on their vector of disease probabilities (Fig. 2A). As expected, individuals naturally clustered based on their clinical diagnosis. PD and ALS cases formed their own separate islands (reflecting the ease with which these conditions were classified). In contrast, a central island emerged with AD and cognitive normal individuals on either end, and offshoots from the center showing higher density of FTD cases on one side and stroke/TIA cases on the other. Importantly, these islands were driven primarily by etiology, and not by contributing site (Fig. S1). Common phenotypic data in the GNPC distributed in expected patterns across the embeddings, with worse cognitive impairment and more *APOE* ε4 carriers in AD regions, less ε4 carriers in PD and ALS regions, and more hypertension in the stroke/TIA region (Fig. 2B, C, D).

Notably, patients diagnosed with subjective cognitive decline (SCD) or mild cognitive impairment (MCI) were not used to train ProtAIDe-Dx, due to the etiology being less clear in these early phases of impairment. We therefore used ProtAIDe-Dx to predict etiological diagnoses for these SCD and MCI cases (N=1,899) in GNPC and projected these cases onto the diagnostic embedding (Fig. 2E). Interestingly, the cases were distributed throughout the embedding, with cases falling neatly into regions corresponding to different conditions. This signals the potential for ProtAIDe-Dx to aid in the diagnosis of patients in early phases of impairment (formally tested in an external sample below).

There were many cases distributed on the embedding into regions inconsistent with their clinical diagnosis, i.e. AD cases distributed into stroke/TIA regions. This observation may have many interpretations – such cases could simply reflect failed model predictions, or even incorrect clinical diagnoses. However, these cases could also reflect cases with multiple overlapping pathologies, or conditions with overlapping molecular etiologies or systemic presentations. Fig. 2F projects contours onto the embeddings representing the highest density of cases for each diagnosis, outlining the space in the embedding that these individuals typically occupy. In generating contours, we also noted sub-clusters of case densities distributed outside of the primary density. To investigate these trends, we clustered cases with the same diagnosis based on embedding coordinates (Fig. S2). Fig. 2F visualizes the top two clusters for each diagnosis. The second (i.e. first non-dominant) cluster of healthy controls emerged in the intersection of shallow extremes of the AD and stroke/TIA regions. This group was older than the dominant cognitively unimpaired group (73.3±10.3 versus 64.8±14.1, p=1e-20), and had higher rates of hyperlipidemia (Fig. 2G, 34% versus 18%, p=2e-8) and other cardiovascular risk factors, more frequent history of smoking and alcoholism, higher cancer rates and worse cognition (SuppData 2). Differential abundance analyses comparing the two control clusters revealed a set of proteins related to metabolism, growth and cofactors, longevity and cognition that were more abundant in the dominant cluster and less abundant in the minor cluster (Fig S3A, SuppData 3, SuppData 4). The minor cluster additionally showed increased abundance of cancer- and inflammation-related proteins. These two clusters may resemble healthy (dominant) vs. unhealthy (minor) aging.

Two minor AD clusters also emerged, with one colocalizing in the stroke/TIA region, and the other in the cognitively unimpaired region (Fig. 2F). Patients in these minor clusters showed lower rates of *APOE* ε4 carriage (p=7e-17). Patients in the minor cluster in the stroke/TIA region were older and showed increased frequency of hypertension (Fig. 2H, 56% versus 15% and 24%, p=2e-22) and

231 other vascular risk factors (SuppData 2), likely signaling mixed AD and vascular co-pathology or
232 misdiagnosis. The dominant cluster meanwhile showed worse cognition (MMSE), while cognition in
233 the "normal" cluster was significantly better. Compared to the other clusters the dominant AD cluster
234 had a higher abundance of proteins involved in metabolism, cell death and stress response, but lower
235 abundance of proteins involved in immune and defense response (Fig 2J, Fig 2K). The minor cluster
236 in the stroke/TIA region meanwhile had more abundance of proteins involved in immune response,
237 proteostasis and cytoskeletal regulation, but lower abundance of proteins involved in synaptic
238 signaling and cell adhesion (SuppData 3, SuppData 4). Finally, the minor cluster in the cognitively
239 unimpaired region showed less enrichment of proteins involved in immune response and energetic
240 metabolism, but a greater abundance of proteins in cell migration, growth factors and plasticity
241 (SuppData 3, SuppData 4). One interpretation of cases in the "cognitively unimpaired" cluster is
242 misdiagnosis of AD, or earlier stages of pathology.
243 Perhaps most interestingly, a minor cluster for ALS emerged distant from the dominant ALS
244 cluster, but in closer proximity to the primary FTD region (Fig. 2F). This minor cluster showed a
245 higher rate of C9orf72 mutations (Fig. 2I, 44% versus 18%, $p=0.04$) and a higher rate of mild
246 cognitive impairment (SuppData 2). Differential abundance between these two groups revealed
247 upregulation of proteins relating to cell death in the C9orf72 cluster, and downregulation of proteins
248 relating to cell growth, motility and adhesion, as well as protein metabolism (Fig. 2L, S3B; SuppData
249 3, SuppData 4). The dominant cluster meanwhile showed downregulation of proteins relating to
250 metabolism, cell fate and muscle use. Differential abundance and characteristics for PD, stroke/TIA
251 and FTD can be found in SuppData 3, and SuppData 4. The KEGG terms enriched among up-
252 regulated proteins in FTD clusters can be found in Fig S3C.

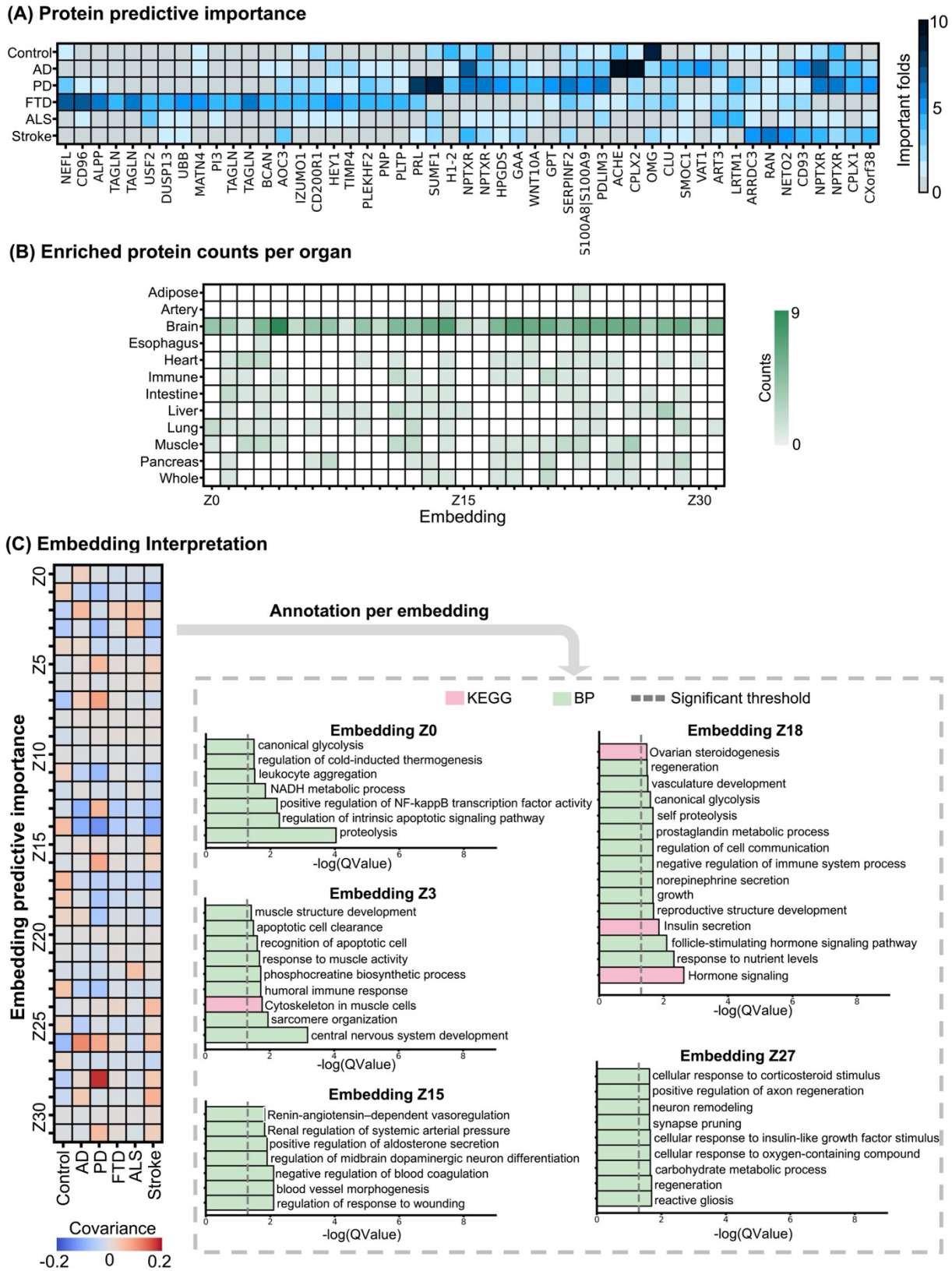


Figure 3. Model interpretations reveal proteomic content underlying model diagnostic predictions. (A) Predictive importance of proteins by fold counts. Proteins were visualized if they showed significant feature importance in more than 4 folds for at least one of the diagnostic tasks. (B) - (C) Interpretations for model-derived embeddings. (B) Frequency of organ-specific enrichment among embedding-specific proteins. (C) Left: Importance of embeddings for diagnostic tasks using covariance, which is a reliable feature importance metric for linear models proposed by Haufe et al., 2014; Right: Gene ontology enrichment analysis on selected embedding-specific proteins. Z0, Z3 and Z15 show signatures specific to AD, ALS and stroke/TIA, respectively, while Z18 and Z27 may signal brain health or resilience.

Model interpretation highlights disease-specific networks and key discriminative proteins.

While deep learning models have a reputation of being difficult to interpret, understanding the underlying biological trends driving predictions made by ProtAIDe-Dx are essential for clinical adoption and biological insight. We employed state-of-the-art approaches in model interpretation [49] to identify the most discriminative proteins used by our model. We employed a conservative approach, robust to multiple comparisons, highlighting proteins that were consistently associated with specific conditions across cross-validation folds (Fig. 3A). Several expected and previously described proteins emerged from this analysis, such as NEFL for FTD and ALS [50–53], CLU [54,55] and SMOC1 [56–58] for AD, and multiple NPTXR aptamers for multiple neurodegenerative diseases [59,60]. ACHE was unsurprisingly an important discriminator for AD, given the common treatment of this condition with ACHE-inhibitors, and given the propensity of systemic medications to affect levels of circulating proteins [61]. SUMF1 and PRL have both been reported to be associated with PD [62,63], and PLTP in the CSF with FTD [64], while AARDC3 (associated here with stroke/TIA) is known to be related to obesity. However, some novel and interesting proteins were also found to be highly discriminative of certain conditions. OMG (oligodendrocyte myelin glycoprotein) was highly discriminative of cognitively unimpaired controls compared to other diseases, being used across all ten folds, while H1-2 (histone H1-2) was also discriminative for this group. The appearance of UBB as a discriminative protein for FTD was also interesting, given the frequency of ubiquitin inclusions in FTD at autopsy [65–67].

Another method for better understanding the ProtAIDe-Dx model is to probe the proteomic composition of the various embeddings that have been optimized toward these multiple diagnostic tasks. The embeddings should represent proteins that express unique nonlinear relationships in relation to neurodegenerative/neurological conditions, and therefore may represent isolated disease-relevant molecular networks or processes. We examine this underlying biology by observing the contribution of embeddings to diagnostic prediction, isolation and enrichment analysis of proteins involved in each embedding (Fig. 3C, SuppData 5, SuppData 6), and association between the embeddings and disease-specific biomarkers in an external dataset (Fig. 4C; see below). We employed a novel method for discovering the selected 738 proteins involved in each embedding (see Methods), but not the directionality of the effects. Contributing protein sets therefore represent a combination of co-abundant proteins and proteins with anti-correlated abundance within the embedding.

Given that proteins in blood come from organs throughout the body, we expected these embeddings to represent processes stemming from multiple organs. However, brain-specific proteins were highly prevalent across all embeddings (Fig. 3B). We therefore tested for enrichment of specific neural cell types (SuppData 7, SuppData 8). Embedding Z4 showed strong specificity to the brain, and to excitatory neurons and astrocytes in particular. Z4 also showed some relationship to cognitively unimpaired and AD diagnosis (Fig. 3C), enrichment of terms relating to apoptosis and inflammation (SuppData 5, SuppData 6), and was associated with older age, worse cognition and unhealthier levels across multiple markers in an external sample (Fig. 4C). This embedding may represent signs of damage to excitatory neurons selectively vulnerable in both aging and AD. Embedding Z11 differentiated controls from various neurodegenerative conditions, and was enriched for proteins expressed by microglia and other immune processes. This embedding was unsurprisingly enriched for proteins associated with inflammatory and immune processes, but also neural differentiation, and was associated with younger age, better cognition and healthier levels across all biomarkers. Z29 differentiated AD and stroke/TIA from controls and was enriched for oligodendrocyte precursor cells and terms relating to cell adhesion, possibly indicating a shared role of these proteins/cells in mixed AD and vascular disease. Finally, Z31 was enriched for astrocytic proteins, was discriminative for PD and was associated with PD-specific markers in the external datasets.

Other embeddings emerged with interpretable annotations helping to understand proteomic underpinnings of specific neurodegenerative diagnosis (Fig. 3C, SuppData 5, SuppData 6). Embedding Z0 showed specificity for AD and was associated with worse cognition and AD biomarkers in an external sample (Fig. 4C). This embedding was enriched for proteins involved in several AD-related processes, such as proteolysis, altered metabolism, immune response and cell death (Fig. 3C, SuppData 5, SuppData 6). Z9 also showed specificity to AD diagnosis and AD-specific biomarkers, and was also enriched for proteolytic processes. Z3 and Z22 both showed specificity for ALS. Z3, in particular, was enriched for both brain and skeletomuscular system processes, as well as cell death, summarizing the degeneration of these systems in ALS. Embedding Z2 was the only embedding without brain-specific proteins and also showed some specificity to ALS (as well as FTD and AD), and was notably associated with neurodegeneration and inflammation biomarkers in the external sample. Embedding Z15 showed some specificity with stroke/TIA, and was enriched with terms relating to wound healing and blood pressure, but also some brain-related terms, highlighting it as a possible signature of neurovascular dysfunction. Notably, while Z15 was associated with unhealthy levels across multiple biomarkers, these relationships were particularly strong with indicators of neurovascular disease in a separate sample, namely white matter lesions and ventricular volume.

Finally, some embeddings appeared to capture proteomic features associated with brain health or resilience (Fig. 3C, SuppData 5, SuppData 6). Embedding Z27 discriminated cognitively unimpaired cases from patients (Fig. 3C), and was associated with younger age, better cognition and healthier biomarker levels in the external sample (Fig. 4C). This embedding was composed of proteins enriched for processes relating to neuronal and synaptic remodeling, regeneration, metabolism and growth factors. Embedding Z1 showed qualities of resistance specifically to non-AD disorders, while Z14 was also highly discriminative of controls from all other conditions. Embedding Z18 presented an interesting case, discriminating controls from PD. This embedding was unintuitively associated with more abnormal levels across almost all biomarkers, but was not associated with cognition, perhaps indicating it as a signature of resilience. Proteins contributing to this embedding were enriched for terms relating to hormonal signaling, regeneration, glycolysis, norepinephrine and female sex, the latter of which was corroborated by much higher expression of this embedding in women from the external dataset compared to men.

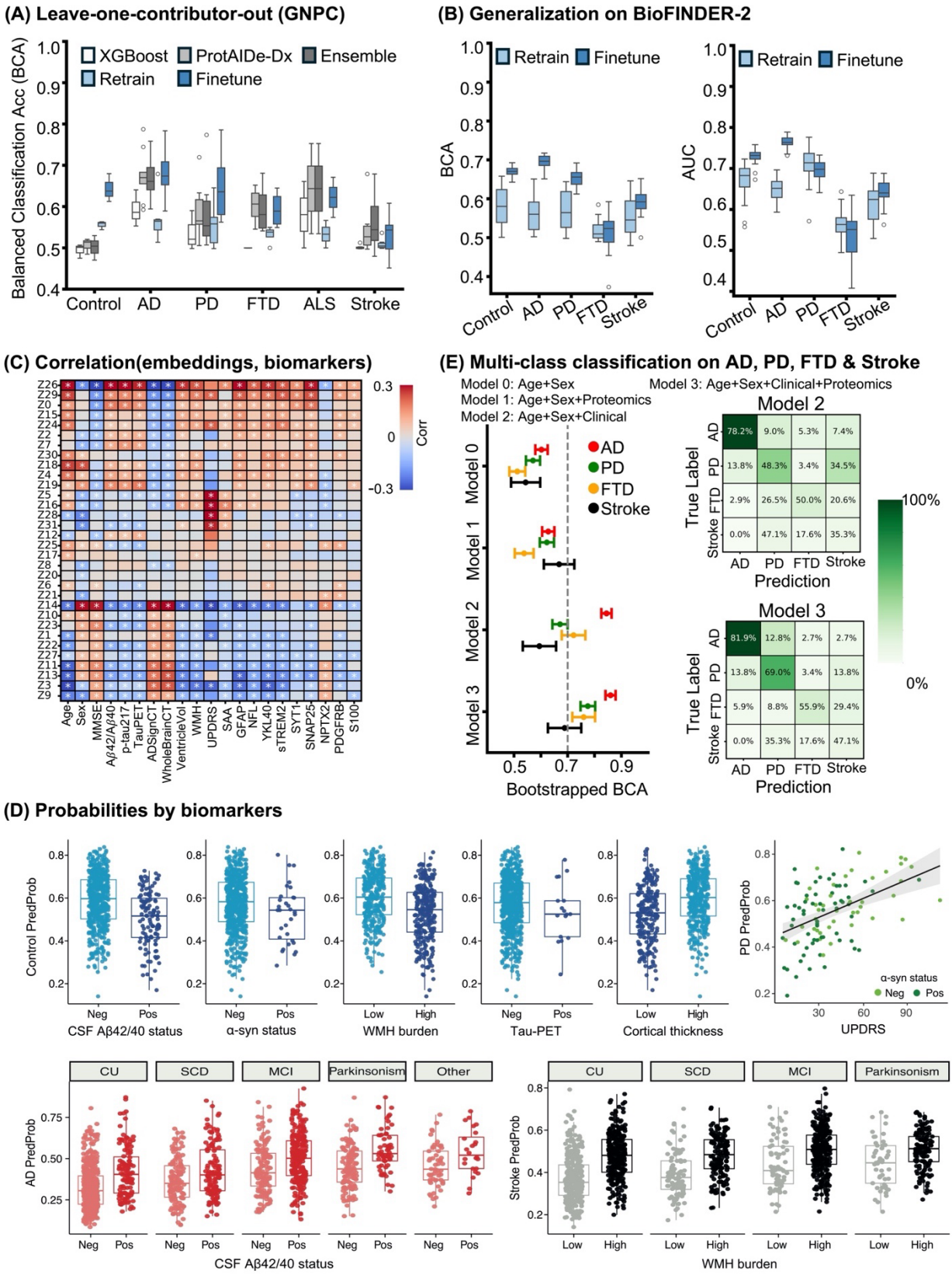


Figure 4. Model validation on external BioFIDNER-2 cohort. (A) Leave-one-site-out BCAs of ProtAIDe-Dx on GNPC. Retrain: Retrain logistic regression model on K participants' proteins, FineTune: Retrain lostic regression model on K participants' proteomics embeddings. (B) Model generalization performance on BioFINDER-2 cohort, Left: BCA; Right: AUC. (D) Correlation between embedding and biomarkers, white star means indicates a significant relationship after FDR correction across all comparisons. (D) Diagnostic predictive probabilities for control (blue), AD (red), PD (green) and stroke/TIA (black) diagnoses, associated with different biomarkers, stratified by biomarker-confirmed clinical diagnosis. (E) Multi-class classification on AD, PD, FTD, and Stroke patients. Left: One-versus-rest BCA computed by 1000 bootstraps on testing patients, Right: Confusion matrix on testing patients for Model 2 (top) and Model 3 (bottom).

Probabilistic disease information maps onto biomarkers of disease-specific neuropathology.

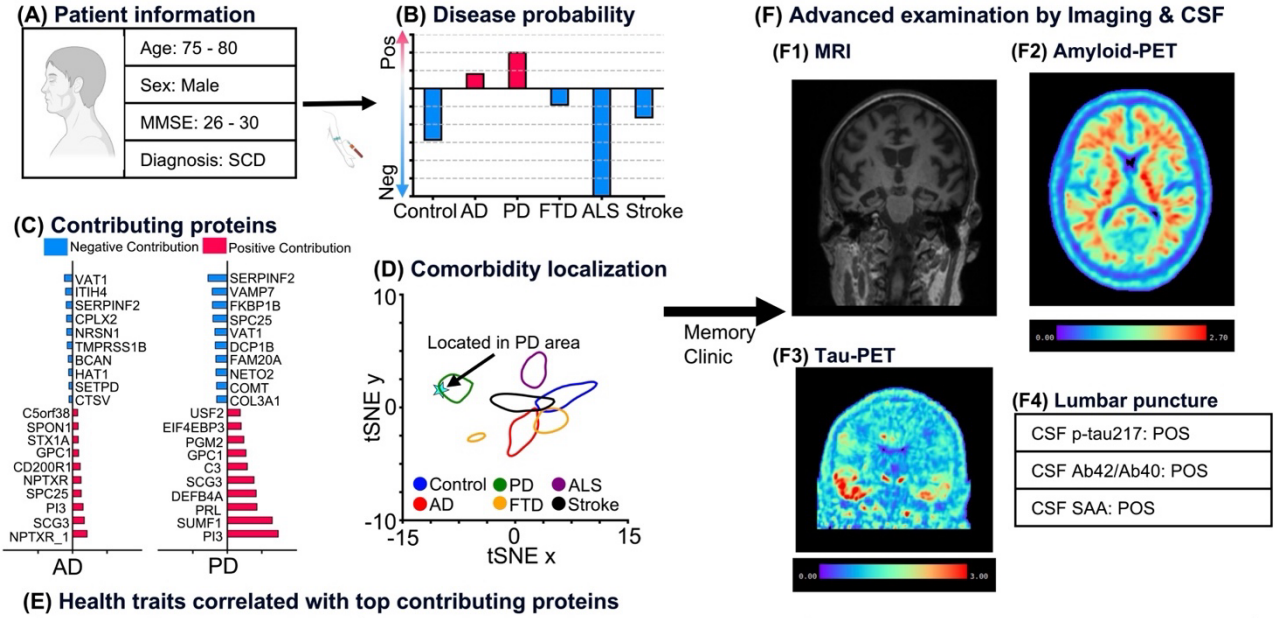
A possible limitation of ProtAIDe-Dx is that it is trained exclusively on clinical diagnosis, and most of these diagnoses were performed without biomarker support. Therefore, we sought to validate ProtAIDe-Dx in an external dataset with many disease-specific biomarkers (BioFINDER-2 dataset, N=1,786). Out-of-sample generalization is necessary for AI models such as ProtAIDe-Dx to be translated to clinical settings, but is challenging given that within-sample performance tends to be optimistic. We therefore used a leave-one-site-out cross-validation approach to evaluate the performance of ProtAIDe-Dx in out-of-sample generalization tasks. ProtAIDe-Dx continued to significantly outperform the XGBoost baseline model, but took a substantial dip in performance in both balanced accuracy score (Fig. 4A) and AUC (Fig. S4). The ensemble model did not aid performance. However, model performance was partially recovered by using finetuning (see Methods), a process involving calibration of a model to a new dataset using a select number of exemplary cases. This approach was particularly effective for improving discrimination of cognitively unimpaired controls from patients.

When applying ProtAIDe-Dx to the BioFINDER-2 dataset, diagnostic performance was consistent with that of the leave-one-site-out analysis (Fig. 4B). Looking at predicted probabilities across diagnosis revealed some expected trends (Fig. S5). For example, PD probabilities were elevated in PD patients, but also in dementia with Lewy bodies (DLB) cases, and stroke/TIA probability was elevated in patients with vascular dementia. Subsequently, we were interested whether these probabilities correlated with disease-specific biomarkers within disease groups (Fig. 4D, S6). Among cognitively unimpaired individuals, model probabilities of being cognitively unimpaired were lower for participants expressing AD, Lewy body or neurovascular pathology (Fig. 4D). This nominates cognitively unimpaired probabilities as a potential signal of general neurodegenerative malaise, and indicates that some “false positive” results from ProtAIDe-Dx are actually correctly identifying underlying preclinical neuropathology (Table S4). Similarly, AD probabilities were higher in non-AD cases with comorbid Aβ pathology (Fig. 4D, S7), and higher stroke/TIA probabilities were associated with greater WMH burden in both impaired and unimpaired individuals (Fig. 4D). PD probabilities did not show a significant relationship with presence of Lewy body pathology (as measured using CSF α-synuclein seed amplification assays; SAA), but were correlated with symptom progression (UPDRS) in PD cases (R=0.47, Fig. 4D). This latter finding is important given a lack of continuous biological signatures of PD progression.

Proteomics provide additive information to diagnosis in a clinically-relevant sample.

For models like ProtAIDe-Dx to be translated to real-life clinical applications, it is important to show evidence that they provide additive value in these settings. Therefore, using the same external BioFINDER-2 dataset (not involved in model training), we generalized ProtAIDe-Dx to a differential diagnostic task in a memory clinic sample (i.e., patients with subjective or objective cognitive impairment seeking memory clinic care). Due to advanced biomarkers, we have greater insight into the probable primary etiological diagnosis of AD, PD, FTD or Stroke/Vascular dementia in these cases, and we fit a series of models seeking to identify this primary diagnosis using a baseline model of only age and sex (Model 0), using just ProtAIDe-Dx and demographics (Model 1), using accessible clinical markers (Model 2: demographics, MMSE, mean cortical thickness of AD-signature meta-ROI (ADSignCT, [68]), plasma p-tau217, and plasma NEFL), and using all of these markers together (Model 3). The final model incorporating ProtAIDe-Dx with common clinical biomarkers (Model 3) achieved significantly better balanced classification accuracy than the model using only common clinical markers (Model 2), especially adding value in diagnosis of PD and neurovascular disease (Fig. 4E,

394 Table S5, S6). This shows ProtAIDe-Dx embeddings provide information complementary to existing
395 clinically accessible neurodegenerative disease biomarkers.
396 The above analysis shows the potential of ProtAIDe-Dx to provide significant additive
397 information to clinical diagnostic work-up. We therefore built a proof-of-concept for a diagnostic
398 report using ProtAIDe-Dx (Fig. 5, S8, S9). The report indicates diagnostic probability across all
399 conditions (Fig. 5B), and includes a localization of the patient to the GNPC disease probability space
400 (Fig. 2A,F; Fig. 5D). Importantly, the report uses model explanation technology to report which
401 proteins contributed to this specific individual's prediction (Fig. 5C), allowing a clear biological
402 explanation of the proteomic diagnosis. Physical traits linked to these patient-specific diagnostic
403 proteins are also provided by programmatically accessing a library of protein-trait associations (Fig.
404 5E), providing possibility for lifestyle interventions or further explanation.
405 We use three patient case examples to showcase the report. Case A (Fig. 5) was a 75-80 year-
406 old male entering the memory clinic with subjective cognitive complaints but objectively intact
407 cognition. ProtAIDe-Dx predicted underlying comorbid AD and Lewy body pathologies. This was
408 confirmed by PET scans showing cortical Aβ burden and temporal tau pathology (AD), as well as
409 positive CSF SAA indicating Lewy body pathology (Fig. 5F). Case B (Fig. S8) presented with MCI in
410 their early-60s, and was given a probable diagnosis of AD by ProtAIDe-Dx, likely due to proteins
411 indicating AD-related synaptic loss (NPTXR), *APOE* ε4 carriage (SPC25), and acetylcholinesterase
412 prescription (ACHE). CSF and PET biomarkers confirmed the presence of AD Aβ and tau pathology
413 in this patient. Finally, Case C was an early 80s male without any subjective or objective cognitive
414 impairment, recruited from the Malmö population. ProtAIDe-Dx predicted a diagnosis of cognitively
415 unimpaired with underlying neurovascular disease (Fig. S9). CSF and PET biomarkers were negative
416 for AD pathology, but the MRI showed enlarged lateral ventricles, periventricular WMH burden and
417 widespread cortical atrophy indicative of neurovascular disease. While not all ProtAIDe-Dx
418 probabilities were accurate and confirmed by biomarkers (Fig. 4B,E), these examples showcase the
419 potential of ProtAIDe-Dx in adding explainable proteomic data informative toward neurodegenerative
420 etiology to clinical workup.



(E) Health traits correlated with top contributing proteins

Protein	Interpretation	Positively Associated Traits	Negatively Associated Traits
NPTXR_1	Higher AD Probability	Cystatin C, Creatinine	BMI, Leucine
VAT1	Lower AD Probability	Cholesteryl Esters to Total Lipids in Small HDL percentage, Cholesterol to Total Lipids in Small HDL percentage	Phospholipids to Total Lipids in Medium HDL percentage, Phospholipids to Total Lipids in Small HDL percentage
PI3	Higher PD Probability	Creatinine, Cystatin C	Degree of Unsaturation, Lymphocyte percentage
SERPINF2	Lower PD Probability	Glycoprotein Acetyls, Concentration of Medium VLDL Particles	SHBG, Average Diameter for HDL Particles

Figure 5. Individual neurodegeneration risk report (Case A). (A) Demographics and cognition score of one older participant, who was diagnosed with SCD. (B) ProtAIDe-Dx predicted this participant with higher probabilities of AD and PD. The probabilities were normalized to be centered at zero. (C) Contributing proteins for making the decision were computed based on SHAP values. The top 10 positive and top 10 negative proteins were visualized. (D) The location of this participant on the diagnostic probability map indicates he was close to typical PD patients. (E) Health traits correlated with top contributing proteins were listed, informing clinicians to pay attention to these traits. (F) Advanced imaging and CSF biomarker examination at the memory clinic to confirm co-pathologies of AD and PD.

Discussion

The rapid growth of dementia populations worldwide urges scalable biomarkers for neurodegenerative diseases. Scalable biomarkers will undoubtedly revolutionize various fields of biomedicine, including pharmaceutical (e.g., trial recruitment, [69]), clinical (e.g., early-stage intervention, [70]), and research (e.g., pathology, [71]) pursuits. Following the recent success of AD biomarkers, biomarkers for many other neurodegenerative diseases are in development, such as extracellular vesicle (EV) biomarkers for FTD pathologies [72] and CSF seed amplification assays of α-synuclein for Lewy Body diseases [73]. However, while scalable biomarkers for certain diseases are on the horizon [74], many biomarkers under development are either invasive (CSF) or not yet scalable, and are usually singular, requiring multiple tests to assess different pathologies. Therefore, a one-shot, multi-disease biomarker panel that is minimally invasive, economical, and widely accessible is highly anticipated. This study presents an early attempt to apply deep neural networks on a massive (N=17,170) neurodegenerative disease sample and high-rank (7,595 proteins) plasma proteomics data, toward this goal. Our proposed ProtAIDe-Dx model synthesized novel biological insight and exhibited clinical utility by mining neurodegeneration-related signals from high-rank proteomics data at multiple levels. These levels ranged from discovering key proteins for robust prediction of certain diseases, to extraction of neurologically relevant protein networks, to simultaneous generation of probabilistic diagnoses across several neurological conditions. Moreover, when generalized to an out-of-sample memory clinic dataset, ProtAIDe-Dx improved automated differential diagnosis accuracy beyond the capability of currently accessible clinical biomarkers. This study distinguishes itself from other neurodegenerative disease proteomics studies [75–78] due to the focus here on making individual-level predictions on new cases, and by training a model to recognize multiple co-pathologies rather than on differential diagnosis. However, despite providing complementary information capable of enhancing current clinical workup, our data suggests that contemporary high-throughput plasma proteomics assays alone cannot replace currently available clinical markers. Altogether, this study sets a benchmark for future proteomics studies by providing a baseline for predictive performance, a resource sorely needed in the biomedicine AI field [79]. Below, we highlight promising findings, and we outline key challenges to be addressed in future work.

Due to its diverse site composition and unstandardized data collection, the GNPC creates an excellent dataset for building and testing tools for translation to real-world clinical settings, where data curation and standardization rarely match that of most research datasets. Within the GNPC, ProtAIDe-Dx achieved impressive cross-validated diagnostic accuracy across diseases, especially for ALS and PD. However, this performance decreased in a leave-one-site-out validation setting that more closely simulates a true clinical translation. This indicates that, despite the large and diverse sample, site effects and overfitting remain a challenge for training generalizable models [80,81]. This cannot necessarily be explained by over-complexity of our deep architecture, as ProtAIDe-Dx consistently outperformed the less complex XGBoost model. It is likely that advances in generalizable data harmonization [82,83] or standardization [84] will be necessary to overcome these limitations. Model finetuning [85,86], a realistic application for a real out-of-site translation scenario, did help to recover performance, particularly for PD and control groups.

In general, several factors must be considered when evaluating the diagnostic performance of ProtAIDe-Dx. One of these factors is the difficulty of the task. For instance, FTD is a highly heterogeneous disorder with multiple underlying pathologies and many different clinical presentations [87,88]. Our model was trained on only 206 FTD participants but, despite these limitations, still managed greater-than-chance performance in out-of-site data, and outperformed XGBoost. This suggests that there are biological features consistent among FTD patients but unique to FTD disorders, which could

likely be exploited with larger samples and more in-depth annotation of patients. Similarly, we used TIA/stroke as a proxy for confirmed neurovascular disease. Prediction of TIA/stroke patients is a challenging task, given that it is a stochastic event that individuals with neurovascular disease are at risk for, rather than a disease itself. Furthermore, stroke is only one among many manifestations of neurovascular disease [35,89], and itself may often go undiagnosed. Despite these challenges, ProtAIDe-Dx learned a plasma proteomic pattern that not only provided greater-than-chance prediction of stroke patients, but also generalized to other indicators of neurovascular disease in a separate cohort.

However, one of the most important considerations in evaluating the performance of prediction models in neurological diseases relates to the accuracy of the original clinical diagnosis, which both forms the basis of training and serves as the standard for evaluation. Neurological and neurodegenerative diseases are notoriously difficult to diagnose [90,91], and in our particular case, models were trained and evaluated based on clinical diagnoses that, for the most part, lacked biomarker confirmation. In addition, the clinical diagnosis criteria varied across sites, given that GNPC is a retrospective collection of multiple cohorts. Given that there are multiple criteria for the diagnosis of AD [92–94] and multiple interpretations of those criteria, this issue will be present in any real-world application. In light of these limitations, many "false" predictions from ProtAIDe-Dx may not represent incorrect diagnoses, considering the prevalence of asymptomatic disease, co-pathology, and misdiagnosis. These concerns were validated on an external dataset with biomarker information. For example, we found that the model tends to predict cognitively unimpaired participants with older age, higher rates of abnormal CSF Aβ42/Aβ40 ratio, and higher CSF p-tau217 as AD dementia.

Along similar lines, in the main GNPC dataset, we found clusters of patients with one diagnosis that showed proteomic signatures in their blood plasma indicative of different diagnosis or etiology. For example, we found a minor cluster of AD patients with proteomic patterns similar to stroke/TIA patients. This might indicate misdiagnosis, could be indicative of co-pathology [95], or may even indicate a vascular etiology for a subgroup of AD patients [96,97]. Supporting the latter hypothesis, Tijms et al. described a proteomic AD subtype associated with blood-brain barrier dysfunction [77] that exhibited similar molecular features to our vascular-AD cluster, including increased levels of immunity-related proteins and reduced synaptic-related processes. By contrast, the "dominant" AD cluster in our study was characterized by a higher abundance of proteins involved in metabolism, cell death, and stress response, alongside lower levels of immune and defense-related proteins. These molecular patterns are corroborated by previous studies. Johnson et al. identified several modules related to AD, including one enriched for MAPK signaling and metabolic pathways, where MAPK is known to regulate cell survival, cell death, and stress responses, and another specifically associated with sugar metabolism [98]. Dysfunction in energy metabolism and brain immunity has also been recognized as a key alteration in AD pathophysiology in another study, with increased metabolic and glycolytic activity and decreased immune-related proteins emerging later in the disease continuum [34].

As another example of proteomic signatures perhaps signaling differing underlying etiologies, we also observed a sub-cluster of ALS patients with a higher rate of C9orf72 mutation carriers, that exhibited a proteomic signature more similar to FTD patients. This highlights the similar underlying molecular pathways known to be shared by some ALS and FTD patients [99]. Interestingly, differential abundance analysis revealed an upregulation of proteins related to cell death in this C9orf72 ALS cluster. Activation of programmed cell death, including apoptosis and other regulated death pathways, is a well-established feature of ALS pathology, observed in both sporadic and familial cases as well as in ALS mouse models [100]. Interestingly, C9orf72 has been implicated in the activation of p53, a major upstream regulator of apoptosis, with evidence showing that mutant C9orf72 promotes neurodegeneration through p53-driven cell death [101]. This raises the possibility that p53-related

522 mechanisms may represent a cell death pathway more specifically engaged in C9orf72-associated
523 ALS.
524 Findings of clinical syndromes demonstrating distinct proteomic profiles are in line with a
525 growing number of studies finding multiple proteomic or transcriptomic subtypes within the AD
526 population [77,102–104]. Together, these observations demonstrate that there is a set of neurodegenerative
527 signatures detectable in blood plasma that may be associated with differing clinical profiles, or that
528 may represent stratified responses to general neurological insult. In either case, a patient's underlying
529 biological response may be just as relevant to treatment and prognosis as their clinical presentation.
530 Future work should leverage longitudinal data to further explore the clinical meaningfulness of these
531 proteomic disease subgroups.
532 We saw many cases where ProtAIDe-Dx seemed to be learning a signature associated with a
533 specific neuropathology, rather than with a clinical syndrome per se. However, we also observed
534 scenarios where ProtAIDe-Dx appeared to be more attuned to symptoms than to underlying biology.
535 When generalized to a new sample, our model's PD probabilities were not associated with CSF α-
536 synuclein status, but were instead correlated with symptom progression in PD cases. While somewhat
537 disappointing with respect to differential diagnosis of motor disorders, there is still clinical potential
538 here, given that no continuous biological measures for tracking PD progression are currently available.
539 On a similar note, we found that lower control probabilities were associated with the presence of
540 multiple pathologies in the cognitively unimpaired group, underlying the potential of this signature as
541 a measure of general brain health.
542 Beyond producing diagnostic probabilities for multiple diseases, ProtAIDe-Dx also identified a
543 compact set of top predictive proteins that significantly influenced model performance on out-of-
544 sample data. Many previous studies have described lists of proteins associated with different diseases
545 [30,34,35,45,98,105,106], and it is important to distinguish between predictive and associational findings. With
546 larger sample sizes and a richer set of variables, association analyses are more prone to detecting
547 "significant" spurious and indirect associations between variables and group status. In contrast,
548 predictive modeling provides a more reliable assessment of a variable's discriminative roles, given
549 that it is validated on out-of-sample data [107]. As a result, predictive findings tend to be more
550 conservative and potentially more clinically useful. For example, a recent study identified 60,942
551 protein-disease associations between 2,920 proteins and 406 prevalent diseases, but only 30 top
552 proteins were needed to predict 124 prevalent diseases with AUC 0.8+ [45]. In our study, among the top
553 proteins discriminating different conditions, those proteins that discriminated controls from all
554 neurodegenerative conditions were perhaps the most interesting, including OMG, H1-2, and multiple
555 NPTXR aptamers. OMG was among the most robustly discriminative proteins in our study, and may
556 be involved in the formation and maintenance of myelin sheaths, which are essential for
557 neurodevelopment and brain regeneration following injury [108]. OMG may play a protective role in
558 maintaining brain health, a hypothesis that needs further investigation in future studies. H1-2 is
559 another novel protein in discriminating controls, which is involved in DNA damage response, gene
560 transcription regulation, and cell cycle [109]. Meanwhile, lower expression levels of NPTXR in
561 neurodegenerative groups have been observed in multiple studies [60,110,111], highlighting synapse loss as
562 a common feature of neurodegeneration [59]. These proteins could, in combination, form the basis of a
563 low-cost blood test for multi-cause neurological malaise, which might be useful in triaging or early
564 diagnosis and monitoring. Further studies should continue to investigate how these proteins change
565 over time in response to other outcomes of brain health.
566 Some key proteins for neurodegenerative diseases were identified by ProtAIDe-Dx as well.
567 NEFL, involved in neuronal damage, has been commonly found to have increased concentrations and

in multiple neurodegenerative diseases [112–114], including FTD [36,115], making it unsurprising as a protein predictive of FTD in the present study. Similarly, PLTP has previously been found to be differentially abundant in the CSF of FTD patients and was correlated with the severity of FTD [64] PLTP may affect TDP-43 or tau clearance via altered lipid dysregulation, given its role in lipid transport [116]. Ubiquitin (UBB) was another interesting protein found to be discriminative of FTD, given that ubiquitinated inclusions are present in some variants of FTD [65,117,118].

ProtAIDe-Dx also identified several proteins predictive of TIA/stroke, perhaps the most interesting being CD93. CD93, mainly expressed in vascular endothelial cells [119]plays a key role in maintaining vascular integrity and regulating inflammation [120,121], making it relevant to both stroke [119] and cerebral small vessel disease [122]. Nonsynonymous SNPs in CD93 have been linked to increased stroke risk, poor outcomes, and hypertension [123], while CD93 deficiency exacerbates blood–brain barrier breakdown and white matter damage [124]. In neurodegenerative diseases like vascular dementia and AD, CD93 may offer protection by limiting microglia inflammation [125] and supporting microvascular health. Other proteins of interest predicting TIA/stroke included RAN and ARRDC3. RAN has emerged as a potential factor in stroke pathology primarily via its role in microRNA pathways [126,127]. An ischemic stroke study showed that an RNA mutation at rs1435 showed a significant association with post-stroke mortality in the large-artery disease subtype [128]. ARRDC3, meanwhile, has been consistently associated with stroke risk factors such as obesity, diabetes, and hyperlipidemia [129–132]. A recent study found that ARRDC3 was significantly overexpressed in lung vessels of chronic thromboembolic pulmonary hypertension (CTEPH) patients compared to controls [133].

Finally, ProtAIDe-Dx identified proteins predictive of PD that warrant further exploration. CSF NPTXR has emerged as a potential biomarker for synaptic dysfunction [110,134,135] and Lewy body disorders [136,137], and plasma levels were predictive of PD in our dataset. NPTXR regulates synaptic plasticity and excitatory synapse formation, and recent studies have shown it to have a neuroprotective role against PD pathology [138,139]. SUMF1 has also been previously reported to be associated with PD[62]. The loss of SUMF1, which encodes an essential enzyme for activating cellular sulfatases [140], leads to lysosomal dysfunction and blocked autophagy [141]. Therefore, SUMF1 may play a role in impaired alpha-synuclein clearance, given that lysosomal dysfunction contributes to the spread of alpha-synuclein [142]. Finally, PRL demonstrates neuroprotective effects through activation of JAK/STAT and PI3K/Akt pathways and reduction of excitotoxicity in preclinical studies [143].

The joint learning framework of ProtAIDe-Dx captured a mixture of disease-specific and more general pathological/neuroprotective signals. For instance, it extracted ALS-discriminative signals using skeletomuscular enriched proteins, which reflect ALS-specific neuromuscular pathology [144]. Another interesting finding is that ProtAIDe-Dx captured sex-related patterns shared with PD-discriminative signals, supporting the observation that biological sex may modulate vulnerability to neurodegenerative diseases [145,146]. In addition to uni-disease signatures, ProtAIDe-Dx also captured some signals shared across multiple neurodegenerative diseases. For example, several embeddings were involved in neuronal cell death, inflammatory, or glial responses, as well as contributing to the discrimination of multiple neurodegenerative diseases. These signals are more likely to imply global neurodegenerative patterns rather than etiology-specific signatures [147]. Interestingly, several embeddings were shared across PD and stroke/TIA. These signatures are involved in neuroinflammation [148,149] and vascular degeneration [150,151] but do not correlate with amyloid pathology. These signals also captured age and cognition effects. Whether these signals may reflect a co-pathology pattern [152,153] between PD and vascular diseases or simply a general 'non-AD' neurodegeneration axis distinct from AD pathology [154] remains to be explored. In addition to disease

signatures, ProtAIDe-Dx extracted some potential brain health or resilience signals. These signals were enriched in multiple neuroprotective mechanisms, including neural integrity [155], immune/inflammatory response [156,157], metabolic maintenance [158], and synaptic plasticity [159]. Consistent with these neuroprotective pathways, biomarker evidence also revealed that these signatures exhibit correlations with younger age, improved inflammation/atrophy markers, and better cognition. An interesting biomarker finding is that some signals were associated with more abnormal levels across almost all biomarkers, but were not associated with cognition, suggesting a potential role of resilience [160]. Perhaps the most encouraging aspect of the proteomics embeddings is their generalizability and compactness. ProtAIDe-Dx compressed high-dimensional proteomics data into low-dimensional embeddings rich in neurodegeneration-related signals by classifying multiple neurodegenerative diseases. This suggests that the embeddings could be extended to new neurodegeneration-related tasks [161,162] and potentially to previously under-studied neurodegenerative diseases [163].

In all, ProtAIDe-Dx showed great promise in disease diagnosis, probing disease heterogeneity, identifying novel proteins of interest, and generating health- and disease-related signatures. However, there are still challenges that we wish to highlight to facilitate future studies exceeding the performance of ProtAIDe-Dx in real-world samples. A major challenge of ProtAIDe-Dx is that its diagnostic accuracy has not yet reached the level required for standalone clinical use. This limitation may be due to several factors. The performance of ProtAIDe-Dx was comparable to that of other studies using UK Biobank data [43,44,76], suggesting that there may be a performance ceiling of plasma proteomics as biomarkers. Another factor might be a limited set of specific aptamers that target arbitrary protein conformations, which are also secreted or surface-based and are detectable in blood. This is a particular challenge for brain diseases, since many disease-relevant proteins are likely brain-expressed and many of these do not cross the blood-brain barrier. Additionally, plasma p-tau biomarkers have achieved much better performance discriminating AD than we found here [7,164] due to the identification of specific peptides and post-translational modifications directly related to disease pathology. To achieve similar improvements, mass spectrometry and/or other approaches may be needed for more comprehensive screening of peptides and protein fragments [165]. Another challenge of ProtAIDe-Dx is its relatively poor site generalization performance, which hinders its applicability to new sites. This limitation may be due to several factors. Strong site effects in plasma proteomics [166] cannot be ignored and may be mitigated by statistical harmonization techniques [82,83,167]. As discussed, the training labels may not be sufficiently reliable due to the lack of biomarker support and imperfect harmonization of diagnostic criteria across cohorts. In addition, demographic, cognitive, and disease distributions varied across cohorts, potentially introducing distribution bias that hampers generalization. This issue could be addressed using adversarial approaches to unlearn distributional bias [168,169]. Hopefully, many of these challenges can be addressed or improved in future releases of GNPC. The third challenge is the potential confounding effects of various factors on protein levels. For example, medication use can significantly affect circulating protein levels, sometimes exceeding normal physiological ranges [61], and may therefore dominate model predictions. Our finding of ACHE predicting AD is likely driven by acetylcholinesterase inhibitors, a common treatment for MCI and AD, elevating plasma ACHE levels specifically in these cases. Another major confounding factor is aging, which is closely linked to the onset of neurodegenerative diseases and which strongly affects proteomic expression levels [170]. These concerns could be mitigated by incorporating covariates or using conditional prediction strategies in the model design, which is worthwhile to explore in future work.

In summary, ProtAIDe-Dx represents a pioneering attempt toward the development of scalable, minimally invasive, and multi-disease diagnostic tools for neurodegenerative diseases. Leveraging deep neural networks trained on the largest neurodegeneration-focused plasma proteomics dataset to date (N = 17,170; 7,595 proteins), the model achieved modest out-of-site performance across six different conditions. Beyond classification accuracy, ProtAIDe-Dx revealed shared molecular signatures across clinical diagnoses, identified a compact set of proteins predictive of multiple diseases, produced low-dimensional proteomics capturing disease- and resilience-related signals, and generated diagnostic probabilities that were correlated with PD symptom progression, CSF biomarkers, and imaging data in a separate memory clinic sample. Furthermore, the learned proteomic embeddings improved diagnostic accuracy beyond what was achievable using accessible clinical biomarkers alone. Despite its promise, predictive proteomics as a field faces several challenges, including insufficient diagnostic accuracy for clinical deployment, limited generalizability across sites, and potential confounding effects. Future work should address these limitations by incorporating more reliable diagnostic labels to improve accuracy, refining model architecture to enhance generalization, and implementing strategies to mitigate confounding biases. Overall, ProtAIDe-Dx establishes a robust benchmark for AI-driven proteomics tools, paving the way for precision medicine in neurodegenerative diseases.

Acknowledgements

This work was supported by the SciLifeLab & Wallenberg Data Driven Life Science Program (grant: KAW 2020.0239), the Crafoord Foundation (20230790), the Swedish Research Council (2024-03642) and the US National Institute of Health (U01 AG079847-02). The BioFINDER study was supported by the National Institute of Aging (R01AG083740), European Research Council (ADG-101096455), Alzheimer's Association (ZEN24-1069572, SG-23-1061717), GHR Foundation, Swedish Research Council (2022-00775, 2021-02219), ERA PerMed (ERAPERMED2021-184), Knut and Alice Wallenberg foundation (2022-0231), Strategic Research Area MultiPark (Multidisciplinary Research in Parkinson's disease) at Lund University, Swedish Alzheimer Foundation (AF-980907, AF-994229), Swedish Brain Foundation (FO2021-0293, FO2023-0163), WASP and DDLS Joint call for research projects (WASP/DDLS22-066), Parkinson foundation of Sweden (1412/22), Cure Alzheimer's fund, Rönström Family Foundation, Konung Gustaf V:s och Drottning Victorias Frimurarestiftelse, Skåne University Hospital Foundation (2020-O000028), Regionalt Forskningsstöd (2022-1259) and Swedish federal government under the ALF agreement (2022-Projekt0080, 2022-Projekt0107). The precursor of 18F-flutemetamol was sponsored by GE Healthcare. The precursor of 18F-RO948 was provided by Roche. The funding sources had no role in the design and conduct of the study; in the collection, analysis, interpretation of the data; or in the preparation, review, or approval of the manuscript.

Disclosures

OH is an employee of Lund University and Eli Lilly. NMC has received consultancy/speaker fees from Biogen, Eli Lilly, Owkin and Merck. JWV has received consultancy/speaker fees from Manifest Technologies.

Method

Datasets

Global Neurodegeneration Proteomics Consortium (GNPC)

Global Neurodegeneration Proteomics Consortium (GNPC, https://www.neuroproteome.org/), launched in 2023, is a mega consortium combining multiple dementia and population cohorts, including health aging, Alzheimer's Disease (AD), Parkinson's Disease (PD), frontotemporal dementia (FTD), amyotrophic lateral sclerosis (ALS), and stroke/TIA (GNPC main paper). The latest GNPC v1.3 release collected 20,426 participants from 22 contributors (sites), 3,942 of whom had longitudinal visits. We used baseline visit proteomics data for model development.

In this study, we selected 17,170 participants based on SomaLogic 7k proteomics availability. One site's proteomics data were extracted from serum, but we kept this site in the model to maximize the sample size and learn modality-agnostic signatures. Table S1 shows the demographics and cognition (Mini-Mental State Examination, MMSE) distribution of participants by site, and the corresponding distribution of clinical diagnoses can be found in Table S2. It is noted that diagnoses of some sites were confirmed by biomarkers, but not for most sites. For the 10,676 participants were neither diagnosed as Control, nor with any of the five above mentioned diseases, we utilized them in model development as "negative group". For example, a participant without diagnosing as Control or five diseases, could still be useful as negative signal to guide training process. In the following analysis, the patients with either Stroke or TIA were labeled with Stroke in figures for ease of visualizations but are referred to stroke/TIA in the Main Text.

All GNPC blood samples were shipped separately by each contributor and were analyzed by SomaLogic, coordinated by Gates Ventures. We removed the outlier values for each protein that exceeded six standard deviations.

BioFINDER-2 cohort

The Swedish BioFINDER-2 dataset (https://biofinder.se/two/, NCT03174938) is a prospective cohort in the south of Sweden spanning the full continuum of AD as well as including patients with non-AD neurodegenerative diseases. Participants are deeply phenotyped, including clinical assessment, CSF/blood sampling, PET, and MRI imaging data. BioFINDER-2 is a participating site in GNPC, but in a sub-analysis aiming to evaluate the disease probability predicted by the different models in relation to key markers of AD and neurodegenerative diseases, we focused specifically on this cohort. For the BioFINDER-2 sub-analysis, we selected 1,786 participants with plasma SomaLogic 7k proteomics data. The demographics and cognition distribution of 1,786 participants can be found in Table S7.

We grouped participants into 6 different groups: Cognitively unimpaired (CU), SCD, MCI, AD, Parkinsonism and Other diseases. The BioFINDER-2 dataset inclusion and exclusion criteria have been described in detail previously [171]. Briefly, CU participants needed to have a MMSE of at least 27 (if < 66 years old) or 26 (if >= 66 years old) and no signs of cognitive symptoms as assessed by physicians specialized in cognitive disorders. Participants with SCD or MCI were all referred to a memory clinic due to cognitive symptoms, had an MMSE between 24 and 30 and did not fulfill criteria for any dementia according to the DSM-5. Participants were classified as MCI if they performed at least 1.5 standard deviations below the normative score on at least one cognitive domain from an extensive neuropsychological test battery [12], while SCD participants performed better than -1.5 standard deviations. Patients with dementia fulfilled the DSM-5 criteria for dementia and all patients with AD dementia were Aβ-positive (based on CSF Aβ42/Aβ40). Patients with non-AD

743 neurodegenerative diseases were also included in the cohort. Patients with PD, dementia with Lewy
744 bodies, atypical parkinsonism disorders formed the group "Parkinsonism". Patients with other diseases
745 were grouped together (labeled "Other") and included patients with FTD spectrum disorders, vascular
746 dementia, and one patient with etiology not otherwise specified. Clinical diagnosis of AD dementia or
747 other neurodegenerative diseases was determined by experienced clinicians.
748 To benchmark generalization performances on BioFINDER-2 (Fig. 4B) against leave-one-site-
749 out on GNPC, we performed additional diagnostic grouping to match GNPC based on clinical
750 diagnosis and biomarkers of BioFINDER-2. Participants with clinical diagnosis as Normal were
751 labeled with Control; patients with abnormal CSF Aβ42/Aβ40 ratio and clinical diagnosis as AD are
752 labeled as AD; patients with a clinical diagnosis with PD, DLB, or Parkinsonism (not otherwise
753 specified) were labeled with PD; patients with a clinical diagnosis behavioral variant FTD, semantic
754 variant of primary progressive aphasia, FTD (not otherwise specified) or a SCD with a MAPT
755 mutation were labeled with FTD; patients with infarcts were labeled with Stroke. It is noted that stroke
756 is not exclusive to other diagnosis categories. For example, an AD patient might also be diagnosed
757 with Stroke. In summary, we got 609 control participants, 261 AD patients, 135 PD patients, 43 FTD
758 patients and 117 stroke patients.
759
760 **Biomarkers of interest in BioFINDER-2**
761 Multiple biomarkers and the MMSE as a global measure of cognition were investigated in
762 BioFINDER-2. The AD biomarkers were Aβ status (based on CSF Aβ42/Aβ40) and tau-PET
763 standardized uptake value ratio (SUVR) in a temporal meta-ROI (tracer ^{18}F-RO948) [172,173]. The
764 positivity of CSF p-tau217 was determined at a cutoff of 11.42 pg/ml [12]. Structural MRI markers of
765 interest were the cortical thickness in an AD signature comprised of temporal lobe regions [68], whole-
766 brain cortical thickness, and ventricular volume (average of lateral ventricle volume in both
767 hemispheres divided by total intracranial volume). T1-weighted MRI were processed with FreeSurfer
768 v6.0. White matter hyperintensity burden (divided by total intracranial volume) was measured based
769 on FLAIR and T1 images processed with Sequence Adaptive Multimodal SEGmentation (SAMSEG)
770 tool from FreeSurfer v7.1. Alpha-synuclein status was available from seed amplification assay done in
771 CSF, as described previously [12]. For a subset of participants in the Parkinsonism group (n=100), the
772 Unified Parkinson's Disease Rating Scale (UPDRS) was also available.
773
774 **Model cross-validation and generalization**
775 We conducted a 10-fold cross-validation procedure to evaluate the performances of ProtAIDe-
776 Dx models on GNPC. Participants from each site were evenly split into 10 folds, so each fold
777 contained data from all sites. A 8-1-1 train-validation-test split was employed, with 8 of 10 folds as a
778 training split, one fold as a validation split, and the remaining one as a test split. This process was
779 repeated 10 times, with each fold serving as the test split once.
780 To test the site generalization performance of the proposed ProtAIDe-Dx models, a leave-one-
781 contributor-out scheme was employed separately. Data from one site was reserved as the test split,
782 while the data from the remaining contributors served as a training-validation split. The sites used as
783 test sets here were selected based on the following criteria: for any of the six conditions, 1) this site
784 had 200 or more participants with non-missing diagnoses; 2) this site had at least 5 participants with
785 minor diagnosis categories for this condition. In total, 13 testing sites fit this criteria. Detailed
786 information can be found in Table S8. Notably, the BioFINDER-2 cohort was part of GNPC,
787 therefore, we excluded BioFINDER-2 participants when training and tuning the ProtAIDe-Dx model
788 for testing on BioFINDER-2. Note that the cross-validation procedure above, and the leave-one-site-

out procedure described here, were separate evaluations and altogether separate models with no interaction.

Machine learning model development

Feature selection

A feature selection procedure was employed to reduce the number of input proteins. On each training and validation split, we conducted both GLM association analysis and XGBoost [47] predictive analysis to select informative proteins as input features. Since GLM does not require any hyperparameters to set, we merged training and validation participants together to run the GLM association analysis for each of the six conditions and each of the 7,595 SomaLogic 7k proteins following:

$$Protein \sim Condition + Age + Sex + AverageProteinLevel \quad (1)$$

AverageProteinLevel is the average protein level across all 7k proteins to control for individual expression differences [174].

After running 7,595 GLMs for each target, we first selected proteins with fold change of beta *Condition* larger > 2 or smaller than 0.5 and then picked the top 5 proteins with the smallest corrected p values of beta *Condition*.

We also ran XGBoost predictive models on each training and validation split to select predictive proteins. We merged training and validation participants together and re-split these participants into 10 subfolds, 9 used for training models and 1 as validation split to tune model hyperparameters. This procedure was repeated 10 times to predict each condition, with each subfold serving as the validation split once. After all 10 models for each condition were trained, proteins internally selected for building models by all 10 XGBoost models were kept.

In this way, the number of input proteins was reduced from 7,595 to 200 - 300 proteins as input features, which vary across each train-val-test split. In total, 738 proteins were kept across 10 train-validation-test splits. This feature selection procedure was performed for both cross-validation and leave-one-contributor-out experiments.

Classification metrics

Due to imbalance distribution of six conditions, we choose balanced classification accuracy (BCA) as classification metric, which was commonly used in AD classification tasks [175,176]. We also reported Area Under Curve (AUC) scores for reference.

XGBoost as baseline approach

XGBoost models were employed as baseline machine approaches to classify binary targets separately. We therefore trained six XGBoost models for classifying six binary conditions. The input proteins to XGBoost models were z-normalized by mean and standard deviation computed from the training split. To get optimal validation prediction accuracies, a grid search on hyperparameters, including maximum tree depth, subsampling of training data, and probability threshold, was performed on the validation split. The XGBoost model trained with optimal hyperparameters was then applied to the test split. This procedure was repeated for each training-validation-test split, including cross-validation and leave-one-site-out.

ProtAIDe-Dx

The ProtAIDe-Dx models were implemented as multilayer neural networks. Input proteins were fed into multiple multilayer neural networks to classify six binary conditions jointly. The loss function of the proposed ProtAIDe-Dx model is a weighted combination of binary cross-entropy loss L_{BCE} and multi-class rank loss [177] L_{RL}.

$$L_{BCE} = -\frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{6}\left[\left(y_{k,i}(1-\alpha)+\frac{\alpha}{2}\right)\log\hat{y}_{k,i} + \left((1-y_{k,i})(1-\alpha)+\frac{\alpha}{2}\right)\log(1-\hat{y}_{k,i})\right] \quad (2)$$

Where N is the number of participants, k is the participant index, i is the index of conditions, $y_{k,i}$ is the true label for the participant k and condition i, α is a label smoothing factor, $\hat{y}_{k,i}$ is the model predicted probability for the participant k and condition i.

The role of rank loss L_{RL} is to constrain the rank of predicted probabilities across conditions, enabling better jointly learning of information across targets. For any two conditions i and j, the rank loss ${L_{RL}}^{i,j}$ is

$$L_{RL}^{i,j} = \frac{1}{N}\sum_{k=1}^{N}\max\left[0, \left(\hat{y}_{k,i} - \hat{y}_{k,j}\right)\left(y_{k,j} - y_{k,i}\right) + \varepsilon\right] \quad (3)$$

Where N is the number of participants, k is the participant index, i and j are two indexes of conditions. $\hat{y}_{k,i}$ and $\hat{y}_{k,j}$ are predicted participant k's probabilities for condition i and j, $y_{k,i}$ and $y_{k,j}$ are k's true labels for condition i and j, ε is a hyperparameter to be set as 0.25.

The rank loss across all conditions is

$$L_{RL} = \frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{6}\sum_{j=i+1}^{6} L_{RL}^{i,j} \quad (4)$$

Therefore, the overall loss function for the proposed ProtAIDe-Dx model is

$$L = L_{BCE} + \lambda L_{RL} \quad (5)$$

Where λ is a hyperparameter to control the weight of L_{RL}.

The probability threshold for each target was determined for the highest validation F1 score. We performed a hyperparameter search on the validation split to get the optimal validation balanced accuracy (BCA) using Optuna [178] with 50 trials. The search range of hyperparameters can be found in the following Table S9.

We performed several additional postprocessing steps before proteomics data was input into ProtAIDe-Dx models. First, each participant's proteomics values were normalized by their averaged protein levels [174]. Second, we fitted a 10-nearest neighbor data imputer onto the training split to impute missing protein entries. Third, a Gaussian rank normalizer [179] was fitted on training split to ensure normalized proteomics are following a normal distribution.

Proposed Ensemble of XGBoost and ProtAIDe-Dx

We conducted an ensemble approach for the developed XGBoost and ProtAIDe-Dx models. The ensemble approach's output probability for each condition was a weighted sum between XGBoost and ProtAIDe-Dx: $w * p_{ProtAIDe-Dx} + (1 - w) * p_{XGBoost}$. On the validation split, we searched for the optimal weight w, ranging from 0 to 1 with the step as 0.01, to get highest validation balanced accuracy. This procedure was repeated for each training-validation-test split, including cross-validation and leave-one-site-out.

870 **Model generalization to new sites by k-shot transfer learning**

871 We developed a k-shot transfer learning framework [86,180,181] to help the developed model
872 generalize to new tasks or new sites. Taking generalization to new sites as an example, only K
873 participants were needed to transfer the ProtAIDe-Dx model to this new site. In this study, we picked
874 K=100 to balance available sample size for each condition and generalization performances. First, the
875 ProtAIDe-Dx model was applied to new sites to get proteomic embeddings. Second, a simple logistic
876 regression model was trained on K participants' proteomics embedding and tested on the remaining
877 participants. The performances were evaluated on remaining participants. Similarly, the developed
878 ProtAIDe-Dx model could transfer to new neurodegeneration-related tasks by training a simple
879 machine learning model on K participants.

880 The underlying assumption is that ProtAIDe-Dx's embeddings have captured enough signals to
881 represent neurodegenerative diseases broadly. Therefore, training a simple machine learning model on
882 low-dimensional proteomics embedding with limited subjects will avoid overfitting issues compared
883 to directly training new models on raw proteins.

884

885 **Estimating predictive importance**

886 To estimate the predictive importance of individual proteins, the PermFIT approach [49] was
887 adopted on each test split. For each input protein, the PermFIT approach randomly permuted this
888 protein across participants and computed the cross-entropy differences compared to un-permutated
889 data. The mean and standard deviation of cross-entropy differences for this protein were obtained from
890 100 permutations, and p-values were calculated based on the mean and standard deviation, assuming a
891 normal distribution.

892 In this study, we measured the predictive importance of proteins in two ways, including Fold
893 Count and Z statistics. Once PermFIT was done on all 10 splits, the Fold Count approach counted the
894 frequency of important proteins across those splits using FDR-corrected p values. The Z statistics
895 approach computed the z scores (mean divided by standard deviations) on each test split and then
896 averaged z scores across 10 splits to get the predictive importance of proteins.

897 To estimate the importance of individual model embeddings in predicting specific diagnoses,
898 we used the Haufe approach. Haufe and colleagues [182] proved that covariance is more reliable than
899 weights (betas) to interpret feature importances in linear models. Given the embeddings were linearly
900 combined to make predictions, we computed the covariance between embeddings and predicted
901 probabilities for each condition to estimate the embeddings' feature importance.

902

903 **Model evaluation on GNPC**

904 In addition to reporting classification accuracy, we performed multiple experiments to validate
905 ProtAIDe-Dx models on GNPC data under cross-validation procedure. First, we correlated predicted
906 AD probability with MMSE scores. Given that each train-validation-test split had different probability
907 thresholds, we first divided AD probability by corresponding AD probability threshold for
908 normalization. Then, we took the logarithm of the mean AD probabilities of participants with the same
909 MMSE score on each testing split. Finally, we computed the Pearson correlation between MMSE
910 scores and the logarithms of the mean AD probabilities. Similarly, we computed the logarithms of
911 mean AD probabilities by *APOE* ε2 or ε4 carrier group on each test split, then averaged them across
912 10 splits.

913

914

915

916 **Model generalization to new task: prediction of longitudinal clinical progression**

917 We validated the generalization ability of nonlinear proteomics embeddings to new tasks.
918 More specifically, we took the embeddings from baseline visits and used these to predict whether
919 participants with CDR 0 at baseline visits would progress to CDR 0.5 or 1 in their following visits or
920 not. For 3,942 participants with longitudinal visits, we selected participants with CDR 0 at baseline
921 visits and non-decreasing CDR during the following visits. For example, participants with CDR 0, 1, 0
922 in their first, second, and third visits would be excluded. In total, we selected 1,945 participants with
923 stable CDR 0 in their following visits and 107 participants with progressive CDR in their following
924 visits. It is noted that the baseline visits of the 1,945+107 = 2,052 participants have been used in the
925 10-fold cross-validation procedure for ProtAIDe-Dx development. We trained Logistic Regression
926 (LR) models to predict whether these 2,052 participants would progress to CDR 0.5 or 1 in their
927 following visits or not, following the former 10-fold cross-validation procedure. We adopted default
928 hyper-parameters for logistic regression models from the sklearn package [183].
929

930 **Dimensionality reduction of predicted probabilities with tSNE**

931 We selected 6,124 participants who were either diagnosed as Recruited Control (N=1,095), AD
932 (N=1,889), PD (N=2,284), FTD (N=176), ALS (N=435), or stroke/TIA (N=335) to project onto a 2-
933 dimensional probability map. We excluded subjects with all negatives or multiple positives for the six
934 targets. It is noted that the predicted probabilities of the 6,124 participants were all from test splits,
935 following the former 10-fold cross-validation procedure.
936 The dimensionality reduction algorithm was chosen as tSNE from the OpenTSNE library [184].
937 We used OpenTSNE over sklearn because OpenTSNE supports inference on new data with fitted
938 tSNE models, allowing us to investigate new participants such as MCI-SCI patients on the same tSNE
939 map. The parameters of tSNE were default values except for setting perplexity to 1000 to better
940 capture global structures.
941

942 **Assessing disease heterogeneity**

943 For each target, KMeans clustering was performed on the 2-dimensional tSNE probability
944 maps of participants within each diagnostic category. To pick the optimal number of clusters, we
945 looped it over from 2 to 10 to pick the one with the highest Silhouette score. In summary, we got 2
946 Control clusters, 3 AD clusters, 4 PD clusters, 2 FTD clusters, 2 ALS clusters, and 5 stroke/TIA
947 clusters.
948 To figure out the differentially expressed proteins across clusters, we ran GLMs across 7k
949 proteins on positive participants for each diagnostic category. Therefore, for each diagnostic category,
950 we have GLMs following the formula:

951 $$Protein \sim C_1 + \cdots + C_K + Age + Sex + Site + AverageProeinLevel \quad (6)$$

952 $C_1 \ldots C_K$ are the binary variables indicating which cluster the participants were from. $Site$ is the
953 categorical variable that indicates which site the participants were from.
954

955 **Enrichment analysis**

956 To evaluate the cell-type expression of genes of interest, we utilized two independent single-
957 cell RNA sequencing (scRNA-seq) datasets. The first dataset comprised single-cell transcriptomes
958 from 2.3 million cells obtained from the aged human prefrontal cortex of 427 participants in the
959 Religious Order Study (ROS) and the Rush Memory and Aging Project (MAP) [185]. This dataset
960 includes neuronal, glial, and vascular cell types. The second dataset, the Human Brain Vascular Atlas,

profiled mainly vascular and perivascular cell types, using 143,793 single-cell transcriptomes from the hippocampus and cortex of eight post-mortem samples [186].

For the two datasets, we downloaded the Seurat objects and used the R package Seurat v4.3.0 [187,188], applying the AverageExpression function to compute cell-type expression levels. We then calculated the proportion of gene expression across all major neuronal and non-neuronal populations.

For organ enrichment analysis, we used organ-enriched genes as defined in previous studies [189] based on the Gene Tissue Expression Atlas (GTEx) bulk RNA-seq database. A gene was classified as organ-enriched if its expression was at least four-fold higher in a specific organ or tissue compared to any other, following the Human Protein Atlas definition [190]. We considered organ categories including adipose tissue, artery, brain, esophagus, heart, immune, intestine, kidney, liver, lung, muscle, pancreas, skin, stomach, and whole blood. Cell- and organ-enriched genes were then mapped to proteins quantified in the SomaScan assay.

To determine whether a specific set of proteins exhibited preferential expression in certain cell types, we calculated the average expression level of the significant protein list across each cell type. To determine whether this enrichment exceeded random expectation, we generated 10,000 randomly sampled gene lists from the SomaLogic background set, each maintaining the same protein count as the significant list. A probability distribution was then constructed based on the average expression levels of these random lists across cell types. This allowed us to quantify the likelihood that the observed enrichment was greater than expected by chance, while accounting for background expression variability. Cell types with a false discovery rate (pFDR)<0.05 (Benjamini–Hochberg correction) were considered significantly enriched.

We performed overrepresentation analyses of KEGG pathways and Gene Ontology (GO) Biological Process (BP) terms using enrichKEGG() and enrichGO() functions in clusterProfiler (v4.14.4)[191] and org.Hs.eg.db (v3.20.0)[192] in R (v4.4.2). The input gene lists consisted of proteins that were significantly up-, down-regulated or both in disease clusters defined by logFoldChange and false discovery rate (pFDR)<0.05 (Benjamini–Hochberg correction). We used the 6,409 human proteins measured by SomaLogic (7k panel) as the background set for overrepresentation analysis. We reported only those KEGG pathways and GO BP terms that met a false discovery rate (FDR) threshold of <0.05. To reduce redundancy among significantly enriched BP terms, we applied the clusterProfiler simplify() function using Wang's semantic similarity measure with a cutoff of 0.7.

Enrichment analyses were performed on proteins differentially expressed across tSNE clusters. Enrichment analysis was additionally performed on embedding-specific proteins. To annotate the biological process of each nonlinear proteomics embedding, we found "embedding specific" proteins using regression. For each embedding, we took the 738 proteins from the cross-validation feature selection to train an XGBoost regression model, with grid search on validation split for the lowest mean squared errors. It is noted that the model was trained to generalize to BioFINDER-2 has 32 embedding dimensions. Once these models were trained, we thresholded proteins with `total_gain` over 50 as "embedding specific" proteins for enrichment analysis.

Model evaluation on BioFINDER-2

In the BioFINDER-2 dataset, where we correlated each biomarker with the different predicted probabilities separately in each of the six diagnostic groups (CU, SCD, MCI, AD, Parkinsonism, Other). FDR correction was applied across all comparisons performed. We also calculated the Pearson correlation between embeddings and biomarkers for BioFINDER-2 participants who had corresponding biomarker available. FDR correction was applied across all comparisons performed.

We further investigated whether proteomics embeddings would provide unique or additive value in distinguishing multiple neurodegenerative diseases in the BioFINDER-2 dataset. We selected 231 AD patients (with abnormal CSF Aβ42/Aβ40 ratio), 111 PD patients (with positive CSF alpha-synuclein status), 39 FTD patients, and 20 Stroke patients (with infarcts), excluding patients with multiple neurodegenerative diseases. We performed a five-fold cross-validation, ensuring the diagnosis distribution was balanced across folds. Three out of five folds were used to train an SVM model, one fold was used to tune hyperparameter C from 0.001 to 100 to get an optimal balanced accuracy score, and the remaining one was used for the test. This procedure was repeated five times to ensure each fold was used as a test fold.

Four sets of features were selected to validate the effectiveness of our proteomics embeddings. Model 0 only used age and sex; Model 1 used age, sex, and the top 5 principal components of proteomic embeddings; Model 2 used age and sex plus clinical biomarkers, including MMSE, plasma p-tau217, plasma NEFL, and mean cortical thickness of AD-signature meta-ROI; and Model 3 used all features in Model 0, Model 1 and Model 2. It is noted that age, MMSE, mean cortical thickness of AD-signature meta-ROI, plasma p-tau217 and plasma NEFL were z-normalized. Both PCA and z normalization were fitted on training splits. Once models were trained, we concatenated participants from five test folds to perform bootstraps 1000 times to perform statistical tests.

Individual disease risk report

To assess contributing proteins for individual prediction, SHapley Additive exPlanations (SHAP) values [193] were computed for each of six conditions, with training participants as the background set. To associate the top contributing proteins with health-related traits, we applied a proteome-phenome atlas (Deng et al., 2025) to find the top correlated traits for each protein. The proteome-phenome atlas was computed based on Olink proteomic data. Therefore, there might be no matched results for some of the contributing proteins; we only included available protein-trait associations.

Statistical analysis

We applied a corrected resample t-test to compare model performances in cross-validation and leave-one-site-out experiments. To examine whether the distribution of variables varies across tSNE clusters or not, we conducted two-proportion z tests for binary variables with two clusters, chi-square tests of independence for both binary variables with three or more groups and multi-category variables, t-test for continuous variables with two clusters, and one-way ANOVA for continuous variables with three or more groups. To compare model performances in comorbidity detection, we applied t-tests on bootstrapped accuracies. It is noted that all p values were FDR corrected.

Data Availability

GNPC (https://www.neuroproteome.org/) will be open-acess upon the publication of this study. Pseudonymized BioFINDER-2 data will be shared by request from a qualified academic investigator for the sole purpose of replicating procedures and results presented in the article and as long as data transfer is in agreement with EU legislation on the general data protection regulation and decisions by the Swedish Ethical Review Authority and Region Skåne, which should be regulated in a material transfer agreement.

Code Availability

Code for the ProtAIDe-Dx model development/evaluation/biological annotations can be found at [GitHub link]. Co-author (YX) reviewed the code before merging it into the GitHub repository to reduce the chance of coding errors.

Reference

1. van Dyck, C. H. *et al.* Lecanemab in early Alzheimer's disease. *N. Engl. J. Med.* **388**, 9–21 (2023).
2. Sims, J. R. *et al.* Donanemab in Early Symptomatic Alzheimer Disease: The TRAILBLAZER-ALZ 2 Randomized Clinical Trial. *JAMA* **330**, 512–527 (2023).
3. Mummery, C. J. *et al.* Tau-targeting antisense oligonucleotide MAPTRx in mild Alzheimer's disease: a phase 1b, randomized, placebo-controlled trial. *Nat. Med.* **29**, 1437–1447 (2023).
4. Kirkeby, A. *et al.* Preclinical quality, safety, and efficacy of a human embryonic stem cell-derived product for the treatment of Parkinson's disease, STEM-PD. *Cell Stem Cell* **30**, 1299-1314.e9 (2023).
5. Boros, B. D., Schoch, K. M., Kreple, C. J. & Miller, T. M. Antisense oligonucleotides for the study and treatment of ALS. *Neurotherapeutics* **19**, 1145–1158 (2022).
6. Bernstein, A. *et al.* Dementia assessment and management in primary care settings: a survey of current provider practices in the United States. *BMC Health Serv. Res.* **19**, 919 (2019).
7. Palmqvist, S. *et al.* Blood biomarkers to detect Alzheimer disease in primary care and secondary care. *JAMA: the journal of the American Medical Association* vol. 332 1245–1257 (2024).
8. Rabinovici, G. D. *et al.* Association of amyloid positron emission tomography with subsequent change in clinical management among Medicare beneficiaries with mild cognitive impairment or dementia. *JAMA* **321**, 1286–1294 (2019).
9. Robinson, J. L. *et al.* Pathological combinations in neurodegenerative disease are heterogeneous and disease-associated. *Brain* **146**, 2557–2569 (2023).
10. Mehta, D., Jackson, R., Paul, G., Shi, J. & Sabbagh, M. Why do trials for Alzheimer's disease drugs keep failing? A discontinued drug perspective for 2010-2015. *Expert Opin. Investig. Drugs* **26**, 735–739 (2017).
11. Landau, S. M. *et al.* Cohort heterogeneity and AD biomarkers in older adults without significant cognitive impairment: A comparison of U.S. POINTER and ADNI. *Alzheimers. Dement.* **19**, (2023).
12. Palmqvist, S. *et al.* Cognitive effects of Lewy body pathology in clinically unimpaired individuals. *Nat. Med.* **29**, 1971–1978 (2023).
13. Gaugler, J. E. *et al.* Characteristics of patients misdiagnosed with Alzheimer's disease and their medication use: an analysis of the NACC-UDS database. *BMC Geriatr.* **13**, 137 (2013).
14. Olsson, B. *et al.* CSF and blood biomarkers for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis. *Lancet Neurol.* **15**, 673–684 (2016).
15. Hansson, O. *et al.* Association between CSF biomarkers and incipient Alzheimer's disease in patients with mild cognitive impairment: a follow-up study. *Lancet Neurol.* **5**, 228–234 (2006).
16. Mattsson, N. *et al.* CSF biomarkers and incipient Alzheimer disease in patients with mild cognitive impairment. *JAMA* **302**, 385–393 (2009).
17. VandeVrede, L. *et al.* Detection of Alzheimer neuropathology in Alzheimer and non-Alzheimer clinical syndromes with blood-based biomarkers. *JAMA Neurol.* (2025) doi:10.1001/jamaneurol.2024.5017.
18. Ossenkoppele, R., van der Kant, R. & Hansson, O. Tau biomarkers in Alzheimer's disease: towards implementation in clinical practice and trials. *Lancet Neurol.* **21**, 726–734 (2022).
19. Schöll, M. *et al.* PET imaging of tau deposition in the aging human brain. *Neuron* **89**, 971–982 (2016).
20. Klunk, W. E. *et al.* Imaging brain amyloid in Alzheimer's disease with Pittsburgh Compound-B: Imaging Amyloid in AD with PIB. *Ann. Neurol.* **55**, 306–319 (2004).
21. Thijssen, E. H. *et al.* Diagnostic value of plasma phosphorylated tau181 in Alzheimer's disease and frontotemporal lobar degeneration. *Nat. Med.* **26**, 387–397 (2020).
22. Janelidze, S. *et al.* Plasma P-tau181 in Alzheimer's disease: relationship to other biomarkers, differential diagnosis, neuropathology and longitudinal progression to Alzheimer's dementia. *Nat. Med.* **26**, 379–386 (2020).
23. Mooldijk, S. S., Licher, S. & Wolters, F. Characterizing demographic, racial, and geographic diversity in dementia research: A systematic review. *JAMA Neurol.* (2021) doi:10.1001/jamaneurol.2021.2943.
24. Chin, A. L., Negash, S. & Hamilton, R. Diversity and disparity in dementia: the impact of ethnoracial differences in Alzheimer disease. *Alzheimer Dis. Assoc. Disord.* **25**, 187–195 (2011).
25. McAleese, K. E. *et al.* Concomitant neurodegenerative pathologies contribute to the transition from mild cognitive impairment to dementia. *Alzheimers. Dement.* **17**, 1121–1133 (2021).
26. Quadalti, C. *et al.* Clinical effects of Lewy body pathology in cognitively impaired individuals. *Nat. Med.* **29**, 1964–1970 (2023).
27. Sattlecker, M. *et al.* Alzheimer's disease biomarker discovery using SOMAscan multiplexed protein technology. *Alzheimers. Dement.* **10**, 724–734 (2014).
28. Hye, A. *et al.* Proteome-based plasma biomarkers for Alzheimer's disease. *Brain* **129**, 3042–3050 (2006).
29. O'Bryant, S. E. *et al.* A serum protein-based algorithm for the detection of Alzheimer disease. *Arch.*

Neurol. **67**, 1077–1081 (2010).
30. Sun, B. B. *et al.* Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).
31. Argentieri, M. A. *et al.* Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations. *Nat. Med.* **30**, 2450–2460 (2024).
32. Carrasco-Zanini, J. *et al.* Proteomic signatures improve risk prediction for common and rare diseases. *Nat. Med.* **30**, 2489–2498 (2024).
33. Oh, H. S.-H. *et al.* Organ aging signatures in the plasma proteome track health and disease. *Nature* **624**, 164–172 (2023).
34. Pichet Binette, A. *et al.* Proteomic changes in Alzheimer's disease associated with progressive Aβ plaque and tau tangle pathologies. *Nat. Neurosci.* **27**, 1880–1891 (2024).
35. Hristovska, I. *et al.* Identification of distinct and shared biomarker panels in different manifestations of cerebral small vessel disease through proteomic profiling. *medRxiv* (2024) doi:10.1101/2024.06.10.24308599.
36. Saloner, R. *et al.* Large-scale network analysis of the cerebrospinal fluid proteome identifies molecular signatures of frontotemporal lobar degeneration. *Res. Sq.* (2024) doi:10.21203/rs.3.rs-4103685/v1.
37. Saloner, R. *et al.* Interrogating the plasma proteome of repetitive head impact exposure and chronic traumatic encephalopathy. *bioRxiv* (2024) doi:10.1101/2024.10.01.616097.
38. Rooney, M. R. *et al.* Plasma proteomic comparisons change as coverage expands for SomaLogic and Olink. *medRxiv* (2024) doi:10.1101/2024.07.11.24310161.
39. Wang, Q. S. *et al.* Platform-dependent effects of genetic variants on plasma APOL1 and their implications for kidney disease. *bioRxivorg* (2025) doi:10.1101/2025.01.30.635763.
40. Candia, J. *et al.* Variability of 7K and 11K SomaScan plasma proteomics assays. *J. Proteome Res.* **23**, 5531–5539 (2024).
41. Shen, X. *et al.* Nonlinear dynamics of multi-omics profiles during human aging. *Nat. Aging* 1–16 (2024).
42. Topol, E. J. The revolution in high-throughput proteomics and AI. *Science* **385**, eads5749 (2024).
43. Guo, Y. *et al.* Plasma proteomic profiles predict future dementia in healthy adults. *Nat. Aging* **4**, 247–260 (2024).
44. You, J. *et al.* Plasma proteomic profiles predict individual future health risk. *Nat. Commun.* **14**, 7817 (2023).
45. Deng, Y.-T. *et al.* Atlas of the plasma proteome in health and disease in 53,026 adults. *Cell* **188**, 253-271.e7 (2025).
46. Gong, J. *et al.* Unraveling the role of plasma proteins in dementia: insights from two cohort studies in the UK, with causal evidence from Mendelian randomization. *medRxiv* 2024.06.04.24308415 (2024).
47. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, San Francisco California USA, 2016).
48. Grinsztajn, L., Oyallon, E. & Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? in *Proceedings of the 36th International Conference on Neural Information Processing Systems* 507–520 (Curran Associates Inc., Red Hook, NY, USA, 2022).
49. Mi, X., Zou, B., Zou, F. & Hu, J. Permutation-based identification of important biomarkers for complex diseases via machine learning models. *Nat. Commun.* **12**, 3008 (2021).
50. Rosengren, L. E., Karlsson, J. E., Karlsson, J. O., Persson, L. I. & Wikkelsø, C. Patients with amyotrophic lateral sclerosis and other neurodegenerative diseases have increased levels of neurofilament protein in CSF. *J. Neurochem.* **67**, 2013–2018 (1996).
51. Verde, F., Otto, M. & Silani, V. Neurofilament light chain as biomarker for amyotrophic lateral sclerosis and frontotemporal dementia. *Front. Neurosci.* **15**, 679199 (2021).
52. Bergström, S. *et al.* A panel of CSF proteins separates genetic frontotemporal dementia from presymptomatic mutation carriers: a GENFI study. *Mol. Neurodegener.* **16**, 79 (2021).
53. Lu, C.-H. *et al.* Neurofilament light chain: A prognostic biomarker in amyotrophic lateral sclerosis. *Neurology* **84**, 2247–2257 (2015).
54. Lambert, J.-C. *et al.* Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat. Genet.* **41**, 1094–1099 (2009).
55. Liu, X. *et al.* Clusterin transduces Alzheimer-risk signals to amyloidogenesis. *Signal Transduct. Target. Ther.* **7**, 325 (2022).
56. Balcomb, K. *et al.* SMOC1 colocalizes with Alzheimer's disease neuropathology and delays Aβ aggregation. *Acta Neuropathol.* **148**, 72 (2024).
57. Roberts, J. A. *et al.* Unbiased proteomics and multivariable regularized regression techniques identify

SMOC1, NOG, APCS, and NTN1 in an Alzheimer's disease brain proteomic signature. *NPJ Aging* **9**, (2023).
58. Morató, X. *et al.* Associations of plasma SMOC1 and soluble IL6RA levels with the progression from mild cognitive impairment to dementia. *Brain Behav. Immun. Health* **42**, 100899 (2024).
59. Zhou, J. *et al.* The neuronal pentraxin Nptx2 regulates complement activity and restrains microglia-mediated synapse loss in neurodegeneration. *Sci. Transl. Med.* **15**, eadf0141 (2023).
60. Nilsson, J. *et al.* Cerebrospinal fluid biomarker panel for synaptic dysfunction in a broad spectrum of neurodegenerative diseases. *Brain* **147**, 2414–2427 (2024).
61. Bolsewig, K. *et al.* Increased plasma DOPA decarboxylase levels in Lewy body disorders are driven by dopaminergic treatment. *Nat. Commun.* **16**, 1139 (2025).
62. Rutledge, J. *et al.* Comprehensive proteomics of CSF, plasma, and urine identify DDC and other biomarkers of early Parkinson's disease. *Acta Neuropathol.* **147**, 52 (2024).
63. Karayel, O. *et al.* Proteome profiling of cerebrospinal fluid reveals biomarker candidates for Parkinson's disease. *Cell Rep. Med.* **3**, 100661 (2022).
64. Hok-A-Hin, Y. S. *et al.* Large-scale CSF proteome profiling identifies biomarkers for accurate diagnosis of Frontotemporal Dementia. *medRxiv* 2024.08.19.24312100 (2024).
65. Fischer, D. F. *et al.* Disease-specific accumulation of mutant ubiquitin as a marker for proteasomal dysfunction in the brain. *FASEB J.* **17**, 2014–2024 (2003).
66. Le Guerroué, F. & Youle, R. J. Ubiquitin signaling in neurodegenerative diseases: an autophagy and proteasome perspective. *Cell Death Differ.* **28**, 439–454 (2021).
67. Kövari, E., Gold, G., Giannakopoulos, P. & Bouras, C. Cortical ubiquitin-positive inclusions in frontotemporal dementia without motor neuron disease: a quantitative immunocytochemical study. *Acta Neuropathol.* **108**, 207–212 (2004).
68. Jack, C. R., Jr *et al.* Defining imaging biomarker cut points for brain aging and Alzheimer's disease. *Alzheimers. Dement.* **13**, 205–216 (2017).
69. Pascoal, T. A. *et al.* Insights into the use of biomarkers in clinical trials in Alzheimer's disease. *EBioMedicine* **108**, 105322 (2024).
70. Chatterjee, P. *et al.* Diagnostic and prognostic plasma biomarkers for preclinical Alzheimer's disease. *Alzheimers. Dement.* **18**, 1141–1154 (2022).
71. Cousins, K. A. Q. *et al.* Distinguishing frontotemporal lobar degeneration tau from TDP-43 using plasma biomarkers. *JAMA Neurol.* **79**, 1155–1164 (2022).
72. Chatterjee, M. *et al.* Plasma extracellular vesicle tau and TDP-43 as diagnostic biomarkers in FTD and ALS. *Nat. Med.* **30**, 1771–1783 (2024).
73. Siderowf, A. *et al.* Assessment of heterogeneity among participants in the Parkinson's Progression Markers Initiative cohort using α-synuclein seed amplification: a cross-sectional study. *Lancet Neurol.* **22**, 407–417 (2023).
74. Kuang, Y. *et al.* A skin-specific α-Synuclein seeding amplification assay for diagnosing Parkinson's disease. *NPJ Parkinsons Dis.* **10**, 129 (2024).
75. Johnson, E. C. B. *et al.* Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. *Nat. Med.* **26**, 769–780 (2020).
76. Gan, Y.-H. *et al.* Large-scale proteomic analyses of incident Parkinson's disease reveal new pathophysiological insights and potential biomarkers. *Nat. Aging* 1–16 (2025).
77. Tijms, B. M. *et al.* Cerebrospinal fluid proteomics in patients with Alzheimer's disease reveals five molecular subtypes with distinct genetic risk profiles. *Nat. Aging* **4**, 33–47 (2024).
78. Walker, K. A. *et al.* Proteomics analysis of plasma from middle-aged adults identifies protein markers of dementia risk in later life. *Sci. Transl. Med.* **15**, eadf5681 (2023).
79. Mahmood, F. A benchmarking crisis in biomedical machine learning. *Nat. Med.* 1–1 (2025).
80. Goecks, J., Jalili, V., Heiser, L. M. & Gray, J. W. How machine learning will transform biomedicine. *Cell* **181**, 92–101 (2020).
81. Wainberg, M., Merico, D., Delong, A. & Frey, B. Deep learning in biomedicine. *Nat. Biotechnol.* **36**, 829–838 (2018).
82. An, L. *et al.* DeepResBat: Deep residual batch harmonization accounting for covariate distribution differences. *Med. Image Anal.* **99**, 103354 (2025).
83. Fortin, J.-P. *et al.* Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage* **167**, 104–120 (2018).
84. Bethlehem, R. a. I. *et al.* Brain charts for the human lifespan. *Nature* **604**, 525–533 (2022).
85. Tajbakhsh, N. *et al.* Convolutional neural networks for medical image analysis: Full training or fine

tuning? *IEEE Trans. Med. Imaging* **35**, 1299–1312 (2016).
86. He, T. *et al.* Meta-matching as a simple framework to translate phenotypic predictive models from big to small data. *Nat. Neurosci.* **25**, 795–804 (2022).
87. Seelaar, H., Rohrer, J. D., Pijnenburg, Y. A. L., Fox, N. C. & van Swieten, J. C. Clinical, genetic and pathological heterogeneity of frontotemporal dementia: a review. *J. Neurol. Neurosurg. Psychiatry* **82**, 476–486 (2011).
88. Josephs, K. A. *et al.* Neuropathological background of phenotypical variability in frontotemporal dementia. *Acta Neuropathol.* **122**, 137–153 (2011).
89. Lorenzini, L. *et al.* Association of vascular risk factors and cerebrovascular pathology with Alzheimer disease pathologic changes in individuals without dementia. *Neurology* **103**, e209801 (2024).
90. Bradford, A., Kunik, M. E., Schulz, P., Williams, S. P. & Singh, H. Missed and delayed diagnosis of dementia in primary care: prevalence and contributing factors. *Alzheimer Dis. Assoc. Disord.* **23**, 306–314 (2009).
91. Rizzo, G. *et al.* Accuracy of clinical diagnosis of dementia with Lewy bodies: a systematic review and meta-analysis. *J. Neurol. Neurosurg. Psychiatry* **89**, 358–366 (2018).
92. Jack, C. R., Jr *et al.* Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's Association Workgroup. *Alzheimers. Dement.* **20**, 5143–5169 (2024).
93. Ma, F. Diagnostic and statistical manual of mental disorders-5 (DSM-5). in *Encyclopedia of Gerontology and Population Aging* 1414–1425 (Springer International Publishing, Cham, 2021).
94. Dubois, B. *et al.* Clinical diagnosis of Alzheimer's disease: recommendations of the International Working Group. *Lancet Neurol.* **20**, 484–496 (2021).
95. Attems, J. & Jellinger, K. A. The overlap between vascular disease and Alzheimer's disease - lessons from pathology. *BMC Med.* **12**, (2014).
96. Royall, D. R. Alzheimer disease as a vascular disorder: nosological evidence. *Stroke; a journal of cerebral circulation* vol. 33 2147–8; author reply 2147-8 (2002).
97. Inoue, Y., Shue, F., Bu, G. & Kanekiyo, T. Pathophysiology and probable etiology of cerebral small vessel disease in vascular dementia and Alzheimer's disease. *Mol. Neurodegener.* **18**, 46 (2023).
98. Johnson, E. C. B. *et al.* Large-scale deep multi-layer analysis of Alzheimer's disease brain reveals strong proteomic disease-related changes not observed at the RNA level. *Nat. Neurosci.* **25**, 213–225 (2022).
99. Balendra, R. & Isaacs, A. M. C9orf72-mediated ALS and FTD: multiple pathways to disease. *Nat. Rev. Neurol.* **14**, 544–558 (2018).
100. Nógrádi, B. *et al.* Mechanism of motoneuronal and pyramidal cell death in amyotrophic lateral sclerosis and its potential therapeutic modulation. *Cell Death Discov.* **10**, 291 (2024).
101. Maor-Nof, M. *et al.* p53 is a central regulator driving neurodegeneration caused by C9orf72 poly(PR). *Cell* **184**, 689-708.e20 (2021).
102. Neff, R. A. *et al.* Molecular subtyping of Alzheimer's disease using RNA sequencing data reveals novel mechanisms and targets. *Sci. Adv.* **7**, eabb5398 (2021).
103. Higginbotham, L. *et al.* Unbiased classification of the elderly human brain proteome resolves distinct clinical and pathophysiological subtypes of cognitive impairment. *Neurobiol. Dis.* **186**, 106286 (2023).
104. Iturria-Medina, Y. *et al.* Unified epigenomic, transcriptomic, proteomic, and metabolomic taxonomy of Alzheimer's disease progression and heterogeneity. *Sci. Adv.* **8**, eabo6764 (2022).
105. Dhindsa, R. S. *et al.* Rare variant associations with plasma protein levels in the UK Biobank. *Nature* **622**, 339–347 (2023).
106. Ali, M. *et al.* Multi-cohort cerebrospinal fluid proteomics identifies robust molecular signatures across the Alzheimer disease continuum. *Neuron* (2025) doi:10.1016/j.neuron.2025.02.014.
107. Bzdok, D., Varoquaux, G. & Steyerberg, E. W. Prediction, not association, paves the road to precision medicine. *JAMA Psychiatry* **78**, 127–128 (2021).
108. Vourc'h, P. & Andres, C. Oligodendrocyte myelin glycoprotein (OMgp): evolution, structure and function. *Brain Res. Brain Res. Rev.* **45**, 115–124 (2004).
109. Lai, S., Jia, J., Cao, X., Zhou, P.-K. & Gao, S. Molecular and cellular functions of the linker histone H1.2. *Front. Cell Dev. Biol.* **9**, 773195 (2021).
110. 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.* (2025) doi:10.1038/s41591-025-03565-2.
111. Kirsebom, B.-E. *et al.* CSF synapse marker changes in predementia Alzheimer´s Disease: AMPA receptor modulators NPTX2 and NPTXR mediate cognition within A/T/N stages. *Alzheimers. Dement.* **19**, e079670 (2023).
112. Giannini, L. A. A. *et al.* Clinical value of longitudinal serum neurofilament light chain in prodromal genetic frontotemporal dementia. *Neurology* **101**, e1069–e1082 (2023).

113. Ashton, N. J. *et al.* A multicentre validation study of the diagnostic value of plasma neurofilament light. *Nat. Commun.* **12**, 3400 (2021).
114. Rohrer, J. D. *et al.* Serum neurofilament light chain protein is a measure of disease intensity in frontotemporal dementia. *Neurology* **87**, 1329–1336 (2016).
115. Sogorb-Esteve, A. *et al.* Proteomic analysis reveals distinct cerebrospinal fluid signatures across genetic frontotemporal dementia subtypes. *Sci. Transl. Med.* **17**, eadm9654 (2025).
116. Zhang, M. *et al.* Structural basis of the lipid transfer mechanism of phospholipid transfer protein (PLTP). *Biochim. Biophys. Acta Mol. Cell Biol. Lipids* **1863**, 1082–1094 (2018).
117. Marsan, E. *et al.* Astroglial toxicity promotes synaptic degeneration in the thalamocortical circuit in frontotemporal dementia with GRN mutations. *J. Clin. Invest.* **133**, (2023).
118. Miklossy, J. *et al.* Lrrk2 and chronic inflammation are linked to pallido-ponto-nigral degeneration caused by the N279K tau mutation. *Acta Neuropathol.* **114**, 243–254 (2007).
119. Piani, F. *et al.* Diagnostic and prognostic role of CD93 in cardiovascular disease: A systematic review. *Biomolecules* **13**, 910 (2023).
120. Vemuri, K. *et al.* CD93 maintains endothelial barrier function and limits metastatic dissemination. *JCI Insight* **9**, (2024).
121. Tossetta, G., Piani, F., Borghi, C. & Marzioni, D. Role of CD93 in health and disease. *Cells* **12**, 1778 (2023).
122. Cordon, J. *et al.* Identification of clinically relevant brain endothelial cell biomarkers in plasma. *Stroke* **54**, 2853–2863 (2023).
123. Mälarstig, A. *et al.* Plasma CD93 concentration is a potential novel biomarker for coronary artery disease: Plasma CD93 concentration in CAD. *J. Intern. Med.* **270**, 229–236 (2011).
124. Griffiths, M. R., Botto, M., Morgan, B. P., Neal, J. W. & Gasque, P. CD93 regulates central nervous system inflammation in two mouse models of autoimmune encephalomyelitis. *Immunology* **155**, 346–355 (2018).
125. Liu, C., Cui, Z., Wang, S. & Zhang, D. CD93 and GIPC expression and localization during central nervous system inflammation. *Neural Regen. Res.* **9**, 1995–2001 (2014).
126. Al-Jehani, H. M., Mousa, A. H., Alhamid, M. A. & Al-Mufti, F. Role of microRNA in the risk stratification of ischemic strokes. *Front. Neurol.* **16**, 1499493 (2025).
127. Rink, C. & Khanna, S. MicroRNA in ischemic stroke etiology and pathology. *Physiol. Genomics* **43**, 521–528 (2011).
128. Kim, J. O. *et al.* Association of MicroRNA biogenesis genes polymorphisms with ischemic stroke susceptibility and post-stroke mortality. *J. Stroke* **20**, 110–121 (2018).
129. Alloubani, A., Nimer, R. & Samara, R. Relationship between hyperlipidemia, cardiovascular disease and stroke: A systematic review. *Curr. Cardiol. Rev.* **17**, e051121189015 (2021).
130. Lu, S. *et al.* Prevalence of hypertension, diabetes, and dyslipidemia, and their additive effects on myocardial infarction and stroke: a cross-sectional study in Nanjing, China. *Ann. Transl. Med.* **7**, 436 (2019).
131. Lin, H. *et al.* Healthy diet is associated with gene expression in blood: the Framingham Heart Study. *Am. J. Clin. Nutr.* **110**, 742–749 (2019).
132. Batista, T. M. *et al.* Arrestin domain-containing 3 (Arrdc3) modulates insulin action and glucose metabolism in liver. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 6733–6740 (2020).
133. Lin, J., Lin, S., Zhang, Y. & Liu, W. Identification of Ferroptosis-related potential biomarkers and immunocyte characteristics in Chronic Thromboembolic Pulmonary Hypertension via bioinformatics analysis. *BMC Cardiovasc. Disord.* **23**, 504 (2023).
134. Nilsson, J. *et al.* Cerebrospinal fluid biomarkers of synaptic dysfunction are altered in Parkinson's disease and related disorders. *Mov. Disord.* **38**, 267–277 (2023).
135. Bartl, M. *et al.* Author Correction: Lysosomal and synaptic dysfunction markers in longitudinal cerebrospinal fluid of de novo Parkinson's disease. *NPJ Parkinsons Dis.* **10**, 121 (2024).
136. Frigerio, I. *et al.* Regional differences in synaptic degeneration are linked to alpha-synuclein burden and axonal damage in Parkinson's disease and dementia with Lewy bodies. *Acta Neuropathol. Commun.* **12**, 4 (2024).
137. Holmes, S. E. *et al.* Synaptic loss and its association with symptom severity in Parkinson's disease. *NPJ Parkinsons Dis.* **10**, 42 (2024).
138. Al-Kuraishy, H. M., Jabir, M. S., Al-Gareeb, A. I. & Albuhadily, A. K. The conceivable role of prolactin hormone in Parkinson disease: The same goal but with different ways. *Ageing Res. Rev.* **91**, 102075 (2023).
139. Nguyen, H. D., Jo, W. H., Hoang, N. H. M. & Kim, M.-S. In silico identification of the potential

molecular mechanisms involved in protective effects of prolactin on motor and memory deficits induced by 1,2-Diacetylbenzene in young and old rats. *Neurotoxicology* **93**, 45–59 (2022).
140. Buono, M. & Cosma, M. P. Sulfatase activities towards the regulation of cell metabolism and signaling in mammals. *Cell. Mol. Life Sci.* **67**, 769–780 (2010).
141. Frankel, L. B. *et al.* A non-conserved miRNA regulates lysosomal function and impacts on a human lysosomal storage disorder. *Nat. Commun.* **5**, 5840 (2014).
142. Udayar, V., Chen, Y., Sidransky, E. & Jagasia, R. Lysosomal dysfunction in neurodegeneration: emerging concepts and methods. *Trends Neurosci.* **45**, 184–199 (2022).
143. Ramos-Martinez, E., Ramos-Martínez, I., Molina-Salinas, G., Zepeda-Ruiz, W. A. & Cerbon, M. The role of prolactin in central nervous system inflammation. *Rev. Neurosci.* **32**, 323–340 (2021).
144. Loeffler, J.-P., Picchiarelli, G., Dupuis, L. & Gonzalez De Aguilar, J.-L. The role of skeletal muscle in amyotrophic lateral sclerosis: Skeletal muscle in ALS. *Brain Pathol.* **26**, 227–236 (2016).
145. Pinares-Garcia, P., Stratikopoulos, M., Zagato, A., Loke, H. & Lee, J. Sex: A significant risk factor for neurodevelopmental and neurodegenerative disorders. *Brain Sciences* **8**, (2018).
146. Reekes, T. H. *et al.* Sex specific cognitive differences in Parkinson disease. *NPJ Parkinsons Dis.* **6**, 7 (2020).
147. Wilson, D. M., 3rd *et al.* Hallmarks of neurodegenerative diseases. *Cell* **186**, 693–714 (2023).
148. Kouli, A. *et al.* Neuroinflammation is linked to dementia risk in Parkinson's disease. *Brain* **147**, 923–935 (2024).
149. Wang, Q., Zheng, J., Pettersson, S., Reynolds, R. & Tan, E.-K. The link between neuroinflammation and the neurovascular unit in synucleinopathies. *Sci. Adv.* **9**, eabq1141 (2023).
150. Guan, J. *et al.* Vascular degeneration in Parkinson's disease: Vascular degeneration of Parkinson's disease. *Brain Pathol.* **23**, 154–164 (2013).
151. Lau, K., Kotzur, R. & Richter, F. Blood-brain barrier alterations and their impact on Parkinson's disease pathogenesis and therapy. *Transl. Neurodegener.* **13**, 37 (2024).
152. Liu, Y., Xue, L., Zhang, Y. & Xie, A. Association between stroke and Parkinson's disease: A meta-analysis. *J. Mol. Neurosci.* **70**, 1169–1176 (2020).
153. Toledo, J. B. *et al.* Dementia with Lewy bodies: Impact of co-pathologies and implications for clinical trial design. *Alzheimers. Dement.* **19**, 318–332 (2023).
154. Vos, S. J. B., Delvenne, A., Jack, C. R., Jr, Thal, D. R. & Visser, P. J. The clinical importance of suspected non-Alzheimer disease pathophysiology. *Nat. Rev. Neurol.* **20**, 337–346 (2024).
155. Jiang, Q. *et al.* Antiageing strategy for neurodegenerative diseases: from mechanisms to clinical advances. *Signal Transduct. Target. Ther.* **10**, 76 (2025).
156. Schwartz, M. & Shechter, R. Systemic inflammatory cells fight off neurodegenerative disease. *Nat. Rev. Neurol.* **6**, 405–410 (2010).
157. Zipp, F. & Aktas, O. The brain as a target of inflammation: common pathways link inflammatory and neurodegenerative diseases. *Trends Neurosci.* **29**, 518–527 (2006).
158. Cai, H. *et al.* Metabolic dysfunction in Alzheimer's disease and related neurodegenerative disorders. *Curr. Alzheimer Res.* **9**, 5–17 (2012).
159. Pilato, F. *et al.* Synaptic plasticity in neurodegenerative diseases evaluated and modulated by in vivo neurophysiological techniques. *Mol. Neurobiol.* **46**, 563–571 (2012).
160. Masurkar, A. V., Marsh, K., Morgan, B., Leitner, D. & Wisniewski, T. Factors affecting resilience and prevention of Alzheimer's disease and related dementias. *Ann. Neurol.* **96**, 633–649 (2024).
161. Sisodia, P. S., Ameta, G. K., Kumar, Y. & Chaplot, N. A review of deep transfer learning approaches for class-wise prediction of Alzheimer's disease using MRI images. *Arch. Comput. Methods Eng.* **30**, 2409–2429 (2023).
162. Tanveer, M. *et al.* Classification of Alzheimer's disease using ensemble of deep neural networks trained through transfer learning. *IEEE J. Biomed. Health Inform.* **26**, 1453–1463 (2022).
163. Singh, K. & Malhotra, D. Meta-health: Learning-to-learn (meta-learning) as a next generation of deep learning exploring healthcare challenges and solutions for rare disorders: A systematic analysis. *Arch. Comput. Methods Eng.* **30**, 4081–4112 (2023).
164. Mielke, M. M. *et al.* Performance of plasma phosphorylated tau 181 and 217 in the community. *Nat. Med.* **28**, 1398–1405 (2022).
165. Barthélemy, N. R. *et al.* Highly accurate blood test for Alzheimer's disease is similar or superior to clinical cerebrospinal fluid tests. *Nat. Med.* **30**, 1085–1095 (2024).
166. Halvey, P. *et al.* Variable blood processing procedures contribute to plasma proteomic variability. *Clin. Proteomics* **18**, 5 (2021).
167. Chen, A. A. *et al.* Mitigating site effects in covariance for machine learning in neuroimaging data. *Hum.*

Brain Mapp. **43**, (2022).
168. Yang, J., Soltan, A., Eyre, D., Yang, Y. & Clifton, D. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ Digit. Med.* **6**, (2023).
169. Xu, H., Liu, X., Li, Y., Jain, A. K. & Tang, J. To be Robust or to be Fair: Towards Fairness in Adversarial Training. *arXiv [cs.LG]* (2020).
170. Moaddel, R. *et al.* Proteomics in aging research: A roadmap to clinical, translational research. *Aging Cell* **20**, e13325 (2021).
171. Palmqvist, S. *et al.* Discriminative accuracy of plasma phospho-tau217 for Alzheimer disease vs other neurodegenerative disorders. *JAMA* **324**, 772–781 (2020).
172. Pichet Binette, A. *et al.* Amyloid-associated increases in soluble tau relate to tau aggregation rates and cognitive decline in early Alzheimer's disease. *Nat. Commun.* **13**, 6635 (2022).
173. Leuzy, A. *et al.* Comparison of group-level and individualized brain regions for measuring change in longitudinal tau positron emission tomography in Alzheimer disease. *JAMA Neurol.* **80**, 614–623 (2023).
174. Karlsson, L. *et al.* Cerebrospinal fluid reference proteins increase accuracy and interpretability of biomarkers for brain diseases. *Nat. Commun.* **15**, 3676 (2024).
175. Marinescu, R. V. *et al.* The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge: Results after 1 Year Follow-up. *J. Mach. Learn. Biomed. Imaging* **1**, 1–60 (2021).
176. Nguyen, M. *et al.* Predicting Alzheimer's disease progression using deep recurrent neural networks. *Neuroimage* **222**, 117203 (2020).
177. Xue, C. *et al.* AI-based differential diagnosis of dementia etiologies on multimodal data. *Nat. Med.* **30**, 2977–2989 (2024).
178. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 2623–2631 (Association for Computing Machinery, New York, NY, USA, 2019).
179. Zając, Z. Preparing continuous features for neural networks with GaussRank. https://fastml.com/preparing-continuous-features-for-neural-networks-with-rankgauss/.
180. Chen, P. *et al.* Multilayer meta-matching: Translating phenotypic prediction models from multiple datasets to small data. *Imaging Neuroscience* **2**, 1–22 (2024).
181. Wulan, N. *et al.* Translating phenotypic prediction models from big to small anatomical MRI data using meta-matching. *Imaging Neuroscience* **2**, 1–21 (2024).
182. Haufe, S. *et al.* On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage* **87**, 96–110 (2014).
183. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **abs/1201.0490**, 2825–2830 (2011).
184. Policar, P. G., Strazar, M. & Zupan, B. OpenTSNE: A modular *python* library for t-SNE dimensionality reduction and embedding. *J. Stat. Softw.* **109**, 1–30 (2024).
185. Mathys, H. *et al.* Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to Alzheimer's disease pathology. *Cell* **186**, 4365-4385.e27 (2023).
186. Yang, A. C. *et al.* A human brain vascular atlas reveals diverse mediators of Alzheimer's risk. *Nature* **603**, 885–892 (2022).
187. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
188. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
189. Zhao, R. *et al.* Plasma proteomics-based organ-specific aging for all-cause mortality and cause-specific mortality: a prospective cohort study. *GeroScience* 1–13 (2024).
190. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
191. Xu, S. *et al.* Using clusterProfiler to characterize multiomics data. *Nat Protoc* **19**, 3292–3320 (2024).
192. Carlson, M. *Org.Hs.Eg.Db*. (Bioconductor, 2017). doi:10.18129/B9.BIOC.ORG.HS.EG.DB.
193. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems* **30**, (2017).