Absorption and Co-expression Modules Show Where Polygenic and Proteomic Risk Scores Diverge in Neurodegenerative Diseases

Charles Zheng*

Department of Bioengineering,
University of Pennsylvania, Philadelphia, PA, USA

Manu Shivakumar*, Li Shen and Dokyoon Kim†

Department of Biostatistics, Epidemiology and Informatics,
University of Pennsylvania, Philadelphia, PA, USA
E-mail: dokyoon.kim@pennmedicine.upenn.edu†

Polygenic and proteomic risk scores are both proposed for pre-symptomatic stratification, yet the extent to which they provide overlapping or complementary information has not been measured across neurodegenerative disease. To quantify this overlap, we define absorption as the fraction of a polygenic score's predictive contribution accounted for by an out-of-fold proteomic score and estimate it among 9,434 to 9,820 UK Biobank participants. Absorption did not track h^2, as Alzheimer's with *APOE*, Alzheimer's without *APOE* and Parkinson's carried matched SNP h^2 of 0.068, 0.061 and 0.069 yet absorbed 0.73, 0.39 and 0.19, with amyotrophic lateral sclerosis at 0.22. Residual genetic signal remained in all four, indicating that proteomic risk scores did not fully capture the predictive information contained in polygenic risk scores. The proteins associated with a polygenic score and the proteins a proteomic score selects overlap no more often than chance, converging only where *APOE* dominates. Across 33 plasma co-expression modules built in 41,358 disease-free participants, germline signal concentrates in modules rather than spreading, with the summary component of seven modules associated with the score for Alzheimer's with *APOE* and none for Parkinson's. The lysosomal module M32 has 89% of its members associated with the score for Alzheimer's with *APOE*, an association absent without *APOE*, yet none was used by the proteomic score. Rebuilding the co-expression modules in *All of Us* gave an adjusted Rand index of 0.582 against the 0.686 attainable within that cohort, and 29 of 33 modules stayed together above a permutation null. We show that transfer is stable and that the disjointness of the proteomic and genomic protein sets is not an artifact of how the sets were defined.

Keywords: Polygenic risk score; proteomic risk score; multi-omics integration; protein co-expression network; module preservation; Alzheimer's disease; Parkinson's disease; APOE.

*These authors contributed equally to this work.
†This author is the corresponding author.

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

1. Introduction

Two stratification methods have been proposed and implemented for estimating patient risk before disease onset. The first uses the germline, static genome, summarized as a polygenic risk score (PRS). The second uses proteomic readings in plasma and sometimes cerebrospinal fluid (CSF), which represents the patient's state of health and physiology more dynamically and is now assayable at scale by affinity platforms.[1,2] With multi-modality data becoming increasingly available, clarifying how much one modality substitutes for the other becomes essential for determining whether PRS and ProRS provide redundant or complementary predictive information, and whether both are needed for risk stratification.

Recent work by Woerner *et al.* showed that proteomic risk scores (ProRS) trained on prevalent samples often surpassed PRS in predictive performance on incident cases while retaining complementary information, but also observed that when used in conjunction, the fraction of PRS's contribution after conditioning on ProRS scaled with heritability of the predicted trait.[3] However, Woerner *et al.* did not specifically analyze, on a protein or pathway resolution, how the two scores overlap or substitute for each other, nor did they individually categorize neurodegenerative diseases such as Alzheimer's Disease (AD), Parkinson's disease (PD), amyotrophic lateral sclerosis (ALS), and frontotemporal dementia (FTD).

Previous studies have shown that *APOE* $\varepsilon 4$ carriage leaves a reproducible imprint on the circulating proteome.[4–6] This presents evidence that plasma protein levels, from which the ProRS is built, are associated with genomic variants, so individual proteins may also associate with a PRS, which aggregates the effects of many loci. Therefore, we ask the question for each of AD, PD, ALS and FTD: to what extent can ProRS can account for proportion of genetic risk explained by the PRS signal, and does this correspond to the same individual proteins being prioritized by each score?

Proteomic and polygenic stratification may converge on the same biological processes without involving the same individual proteins. For example, genetically associated and disease-predictive proteins may occupy different roles within one pathway, and many individually weak associations may become detectable only when treated as a coordinated system. Co-expression modules provide such a higher-order representation by capturing biologically coherent processes, and protein modules recover reproducible biology across cohorts and tissues.[7–10] Prior serum work further showed that modules associated with incident disease and mortality are under cis- and trans-genetic regulation.[11] We therefore ask whether the predictive redundancy observed between PRS and ProRS reflects sparse associations across the plasma proteome or concentration within a limited number of co-expression modules, and whether the two scores converge at the module-level despite limited overlap among individual proteins.

We address these questions in UK Biobank, using 2,920 Olink analytes, polygenic scores built exclusively from non-UK-Biobank discovery GWAS, and up to 41,358 participants without a recorded diagnosis for the proteome-wide work. Our module-level analysis replicated modestly in 9,211 disease-free participants of *All of Us*, assayed on a later Olink panel. Our contributions are fourfold. We quantify the extent to which proteomic risk captures the predictive contribution of polygenic risk as an effect size with a bootstrap interval, and find it substantial for AD, partial in every disease examined. We then show that the proteins ProRS

selects and the ones associated with polygenic risk are largely different populations, overlapping no more often than chance except for AD with APOE. We then analyzed 33 data-derived co-expression modules and observed the two sets' modular separation as well as patterns of their respective enrichment. Finally, we ask whether those module definitions are stable across cohort transfer by rebuilding them in All of Us and measuring their agreement with UK Biobank, which identifies which modules are reproducible across cohorts, for use in later studies relating germline or proteomic risk to the plasma proteome.

2. Methods

2.1. *Cohorts, phenotyping and quality control*

Discovery analyses used UK Biobank,[12,13] restricted to participants of European genetic ancestry with Olink Explore 3072 proteomics at the baseline instance.[1] Cases were defined from linked hospital and death records by ICD-10 code (AD G30, PD G20, ALS G12.2, FTD G31.0) and classified as prevalent or incident relative to venepuncture. Three distinct analysis populations are used. The nested logistic models use the complete-case intersection of proteome, genotype and covariates with a downsampled control pool, giving 9,434–9,820 participants per disease. The association analyses use every participant with genotype data but without a recorded diagnosis at venepuncture, including the incident diagnoses, giving 38,602–39,807 per disease. Modules are built once in the 41,358 participants with proteomic data and no recorded AD diagnosis. These participants were not screened for PD, ALS or FTD, so about 700 of them may carry one of those diagnoses. Covariates throughout were age, age squared, sex and the first ten genotype principal components. Olink NPX values were used as released by UK Biobank, with no additional plate or processing-batch adjustment, because plate and batch identifiers were not available in our data extract.

Modules and absorption were assessed in an external cohort. *All of Us* contributed 9,969 participants with Olink Explore HT plasma proteomics.[14,15] Of the 2,920 UK Biobank module proteins, 2,819 were present on Explore HT by gene symbol and survived QC. Proteins missing in more than 20% of participants and participants missing more than 20% of proteins were dropped, and the remainder were standardized and mean-imputed, whereas the UK Biobank matrix was k-nearest-neighbour imputed.

2.2. *Polygenic risk scores*

To preclude sample-overlap inflation, discovery summary statistics were taken exclusively from non-UK-Biobank GWAS (Table 1).[16–19] Posterior effect sizes were obtained with PRS-CS under the continuous-shrinkage prior against the 1000 Genomes European LD reference,[20] and scores computed with PLINK.[21] Two AD scores were constructed, one genome-wide and one excluding chr19:44.4–46.5 Mb (GRCh37),[22] with the exclusion applied to the summary statistics before shrinkage so that the resulting posterior is *APOE*-free by construction. SNP heritability on the liability scale was estimated by LD score regression.[23]

2.3. *Proteomic risk scores*

ProRS[3] were fitted by LASSO logistic regression with `glmnet`,[24] with proteins penalized and covariates forced in unpenalized, and λ selected by inner cross-validation. A repeated stratified five-fold design was used in which all prevalent cases enter every training fold while incident cases and a downsized control sample are scored strictly out of fold; the entire procedure was repeated three times with distinct seeds, giving fifteen selection opportunities per protein. In each repetition, controls were randomly downsampled to 10,000 of the roughly 41,000 to 43,000 available, and these controls together with all incident cases were divided into five case-control-stratified folds, with prevalent cases added to every training set. This gave roughly 13 to 38 controls per case in training depending on disease, and the ratio was not varied. Because the ProRS is never exposed to the genetic score, one cached ProRS per disease is compared against every genetic score. A protein counts as ProRS-selected when it is selected in at least 7 of the 15 folds. As a sensitivity analysis on this threshold itself, the cut was swept from 1 to 15 folds; the observed-to-expected overlap between the PRS-associated and ProRS-selected sets for AD with *APOE* rose across the entire range, from parity at $t = 1$ to 5.7-fold at $t = 15$, crossing nominal significance ($p < 0.05$) at $t = 3$ and remaining significant through $t = 15$, while none of the other three scores reached nominal significance at any threshold (Supplementary Fig. S1). The reported separation is therefore not because of the selected 7-of-15 cut. As a sensitivity analysis, every ProRS was refitted by elastic net at $\alpha = 0.5$ with the penalty being the only difference.

2.4. *Proteomic absorption of polygenic predictive signal*

For each disease we fitted four nested logistic models on an identical complete-case set, these being covariates alone, covariates plus PRS, covariates plus ProRS, and both.[3] Using ΔR^2 increment,[25] absorption is

$$A \;=\; 1 - \frac{\Delta R^2(\mathrm{PRS} \mid \mathrm{cov} + \mathrm{ProRS})}{\Delta R^2(\mathrm{PRS} \mid \mathrm{cov})},$$

the proportion of the polygenic score's own predictive contribution accounted by the proteome. Confidence intervals came from a case-control-stratified bootstrap with 2000 resamples, and between-disease comparisons use a paired bootstrap that resamples the shared control pool once and applies the same resample to both diseases.[26] The DeLong test on the area under the curve is also reported.[27]

2.5. *PRS-associated proteins*

A protein was called PRS-associated if its abundance tracked that score among participants without a recorded diagnosis at proteomic assessment, including controls and incident cases. Prevalent cases were excluded so that established disease, its treatment, or its physiological consequences could not alter protein abundance and be misattributed to a PRS association. Each of the 2,920 assayed proteins was regressed individually on the standardized PRS with age, age^2, sex and PC1–10 as covariates, and the resulting p-values were Benjamini–Hochberg corrected at $q < 0.05$ within each score, and every subsequent q value carries the same correction. Two genomic regions are reported separately from the rest. The major histocompatibility

complex and the 17q21.31 inversion[1,28,29] were associated with a very wide range of traits and contain strong *cis*-pQTLs, so every polygenic score carried weight there and every score therefore appeared to associate with the proteins those regions encode. Proteins encoded in the two regions are held out of the proteome-share denominator for that reason and counted separately (Fig. 1a).

2.6. *Plasma proteome co-expression modules*

In the 41,358 disease-free participants, each protein was residualized on the covariate design by ordinary least squares, retaining residuals only. Residuals were standardized and Pearson-correlated pairwise, and the signed distance $1 - r$ clustered by Ward's minimum-variance criterion following established practice.[7] The resulting dendrogram was cut at twelve candidate resolutions spanning $K = 8$ to 50 clusters. Each resolution was characterized on reproducibility, measured by partitioning participants at random into halves and scoring each half's partition against the full-data partition by the adjusted Rand index (ARI) over five independent splits, as well as the smallest module the cut produced (Fig. 2c). Reproducibility rises monotonically across the sweep and so cannot select a resolution on its own; thus, we selected for $K = 35$ as the finest resolution before the smallest module size degrades from a plateau held at 17 proteins from $K = 20$ through $K = 35$ to 8 at $K = 40$ and 6 at $K = 50$, relative to a floor of 15 proteins per module (dashed line in Fig. 2c). The two neighboring resolutions, $K = 30$ and $K = 40$, both have split-half reproducibility within 0.03 of the selected cutoff (0.742 and 0.771 against 0.766 at $K = 35$) and agree with the retained 33-module partition at ARI 0.962 and 0.986 respectively, so $K = 35$ is stable relative to its neighbors (Supplementary Table S1). Each module was summarized by its eigenprotein, the score on the first principal component of its members' residualized abundances, which assigned one value to every participant. Any pair of eigenproteins correlating above 0.85 was merged. The retained cut requested 35 clusters, and two such merges returned the 33 modules analyzed throughout (Fig. 2c).

Modules were annotated by hypergeometric over-representation of Gene Ontology terms with the assayed panel as background. Module-level association was assessed by regressing each module's standardized eigenprotein on the standardized polygenic score with the same covariates, in disease-free participants only so that disease-consequence cannot confound any discovery. Each module was additionally tested for over-representation of the PRS-associated set and of the ProRS-selected set by hypergeometric test against the assayed panel, giving three per-module axes in total.

2.7. *Cross-cohort comparison*

Module transfer was assessed in *All of Us* using proteins shared with UK Biobank. The two cohorts were residualized and partitioned independently through the same clustering procedure, and agreement was quantified by the adjusted Rand index (ARI). As an internal reference ceiling, *All of Us* was split in half ten times, each pair of half-sample partitions was compared, and the median ARI was retained. Module-specific preservation was assessed by the fraction of each UK Biobank module's protein pairs assigned to the same *All of Us* module. This co-location fraction was compared with 2,000 size-matched random protein

sets, converted to a z-score, and tested using one-sided p-values with Benjamini–Hochberg correction across 33 modules. In a complementary analysis, UK Biobank module membership was held fixed without reclustering *All of Us*. Module coherence, defined as the variance explained by the first principal component, was recomputed in *All of Us* and compared with the corresponding UK Biobank values.[30] UK Biobank modules were derived in European-similar disease-free participants, whereas the *All of Us* analysis included all ancestries because restricting to the 2,681 European-similar participants would reduce the available sample from 9,211 by approximately fourfold.

Absorption was additionally evaluated in *All of Us* by applying the frozen UK Biobank ProRS weights without refitting. Refitting was not attempted because only 20 to 56 disease cases were available (Supplementary Table S2), so the transferred score tests the UK Biobank weights rather than the *All of Us* proteome's best achievable fit.

3. Results

3.1. *Proteomic substitution for polygenic risk is partial in every disease and its magnitude tracks one locus*

Adding ProRS to the covariates model improved discrimination in every disease and by a wider margin than what PRS achieved, with AD rising from an area under the curve of 0.819 for covariates alone to 0.855 with PRS and 0.903 with ProRS (Table 2), echoing *Woerner et al.*'s findings. Absorption was 0.727 (95% CI 0.645–0.806) for the *APOE*-containing AD score, 0.394 (0.119–0.667) for AD with *APOE* excluded, 0.221 (0.009–0.509) for ALS and 0.187 (0.001–0.380) for PD. Because the 95% confidence interval for PD absorption reaches down to zero, the data cannot rule out a complete absence of proteomic absorption of genetic risk in PD. The residual polygenic increment's interval excluded zero in all four contrasts, with $\Delta R^2(\text{PRS} \mid \text{ProRS})$ at 0.0226 (0.0134–0.0341), 0.0057 (0.0014–0.0128), 0.0081 (0.0016–0.0184) and 0.0047 (0.0013–0.0101) for AD with APOE, without APOE, ALS and PD respectively.

Table 1. *Study design. Phenotype counts are for European-ancestry UK Biobank participants with Olink plasma proteomics and complete covariates, before restriction to participants with a polygenic score. SNP-heritability is on the liability scale from LD score regression.*

Trait	ICD-10	Discovery GWAS	Prevalent	Incident	Controls	Liability h^2 (SE)
AD, with *APOE*	G30	Kunkle 2019	73	432	41,358	0.068 (0.011)
AD, no *APOE*	G30	Kunkle 2019	73	432	41,358	0.061 (0.009)
PD	G20	FinnGen R13	105	652	42,372	0.069 (0.012)
ALS	G12.2	van Rheenen 2021	36	220	42,873	0.028 (0.003)
FTD*	G31.0	Ferrari 2014	20	69	43,040	0.026 (0.022)

*FTD is unstable (the standard error of $h^2 \approx$ the estimate) and is excluded from subsequent claims.

AD with APOE had the greatest absorption of any disease model at 0.73 whereas PD was the lowest at 0.19, and SNP heritability alone cannot explain the differences between them, since liability-scale h^2 is 0.068, 0.061 and 0.069 for the AD, AD-without-*APOE* and

PD scores, all within one standard error of each other (Table 1). Direct cross-disease comparisons should be interpreted cautiously because they involve different disease-specific case populations and PRS predictive increments, and because the PD estimate is imprecise, with its lower bound touching zero. The cleaner contrast is therefore within AD, where the outcome, analysis population, covariates, and ProRS were held constant. Excluding the APOE region reduced absorption from 0.73 to 0.39, isolating the APOE region as a major contributor to the unusually high absorption observed for AD.

An attempt to repeat the analysis in *All of Us* reproduces the point estimates but cannot address the contrast. Absorption for the *APOE*-containing AD score was 0.470 (0.248–0.862) on 39 prevalent cases, an interval overlapping the UK Biobank estimate, and PD gave 0.249 in the European-similar subset, likewise overlapping (Supplementary Table S3). Because these analyses rest on very small samples, this cohort provides an underpowered assessment of absorption effect sizes rather than an independent validation of the disease-specific dissociations observed in UK Biobank. AD without *APOE* was not estimable there, its denominator $\Delta R^2(\text{PRS} \mid \text{cov})$ falling to about 5×10^{-5}. The AD-versus-PD contrast rests on 20 and 35 cases in that subset and is under-powered. The frozen UK Biobank proteomic scores transferred to *All of Us* without refitting, giving an AUC of 0.865 for AD and 0.791 for PD in all ancestries (Supplementary Table S2).

Table 2. *The four-model AUC and absorption results. ΔR^2 values are Nagelkerke increments. DeLong p reflects AUC significance of PRS given ProRS and covariates. Absorption intervals are percentile bootstrap ($B = 2000$, 3 cross-validation repeats). FTD is characterised in Table 1 but omitted here, its absorption denominator being below the stability threshold.*

	AUC				ΔR^2			
Trait	cov	+PRS	+ProRS	full	PRS|cov	PRS|ProRS	DeLong p	A [95% CI]
AD with *APOE*	0.819	0.855	0.903	0.908	0.0828	0.0226	0.056	0.73 [0.65, 0.81]
AD no *APOE*	0.819	0.825	0.903	0.905	0.0095	0.0057	0.125	0.39 [0.12, 0.67]
PD	0.745	0.750	0.805	0.809	0.0058	0.0047	0.010	0.19 [0.00, 0.38]
ALS	0.633	0.656	0.723	0.733	0.0104	0.0081	0.162	0.22 [0.01, 0.51]

3.2. *The proteins polygenic risk associates with and the proteins the proteomic score uses are generally distinct*

Regressing every assayed protein on each PRS in disease-free participants demonstrates the plasma's association with genomic signal on an individual-protein resolution. The *APOE*-containing AD score was associated with 7.9% of the proteome (227 of 2,873 proteins), against 1.2% for AD's polygenic background, 1.7% for ALS and 0.5% for PD (Fig. 1a; Supplementary Table S4). Proteins encoded in the MHC and in the 17q21.31 inversion are held out of that denominator and drawn separately, so percentages are computed over the 2,873 analytes outside those regions while the counts described below are over all 2,920. AD's polygenic back-

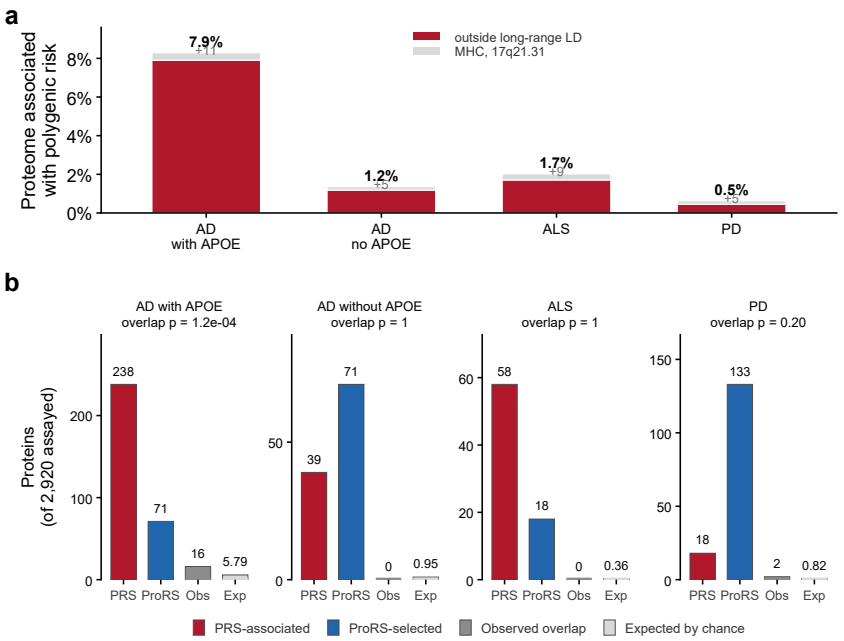


Fig. 1. *Two protein populations, in 38,602 to 39,807 participants without a recorded diagnosis at venepuncture, depending on trait. (a) Share of the assayed proteome each polygenic score is associated with at $q < 0.05$, over the 2,873 analytes outside the MHC and the 17q21.31 inversion, with the count contributed by those two regions drawn as a separate segment. (b) Absolute count of PRS-associated proteins, ProRS-selected proteins at 7 of 15 folds, their observed overlap and the overlap expected by chance, with the one-sided hypergeometric p for excess overlap above each panel.*

ground associates with the products of its own risk genes[16] (*TREM2* $q = 7.9 \times 10^{-10}$, *CD33* $q = 1.4 \times 10^{-5}$, *CR1* $q = 3.4 \times 10^{-3}$), whereas the PD score associates with neither of its main risk genes *SNCA* ($q = 0.86$) nor *GBA* ($q = 0.81$)[31] (Supplementary Table S5); *LRRK2* is not represented on the Olink panel and was not tested.

AD with APOE is the outlier here as it was for absorption. Its PRS associates with the most plasma proteins at 238, and it is the only model whose overlap with the ProRS-selected set exceeds chance, at 16 shared proteins against 5.79 expected (Supplementary Tables S6 and S7; Fig. 1b). The three other models are largely disjoint. AD without *APOE* and ALS have separate ProRS, PRS sets, and PD shares two against 0.82 expected, which is not statistically distinguishable from chance. PD has the smallest PRS-associated footprint at 18 proteins, and its low absorption further suggests that the PD score is associated with few plasma proteins, which together with its low absorption suggests a limited plasma footprint, although the PD estimates are imprecise.

3.3. *Germline signal and the proteomic score's vocabulary separate at module resolution*

Unsupervised clustering of the residualized proteome in UK Biobank yielded 33 co-expression modules after two pairs being merged from 35 modules, ranging from 17 to 717 proteins (Supplementary Table S8). Modules recovered recognizable biology, including immune cell-

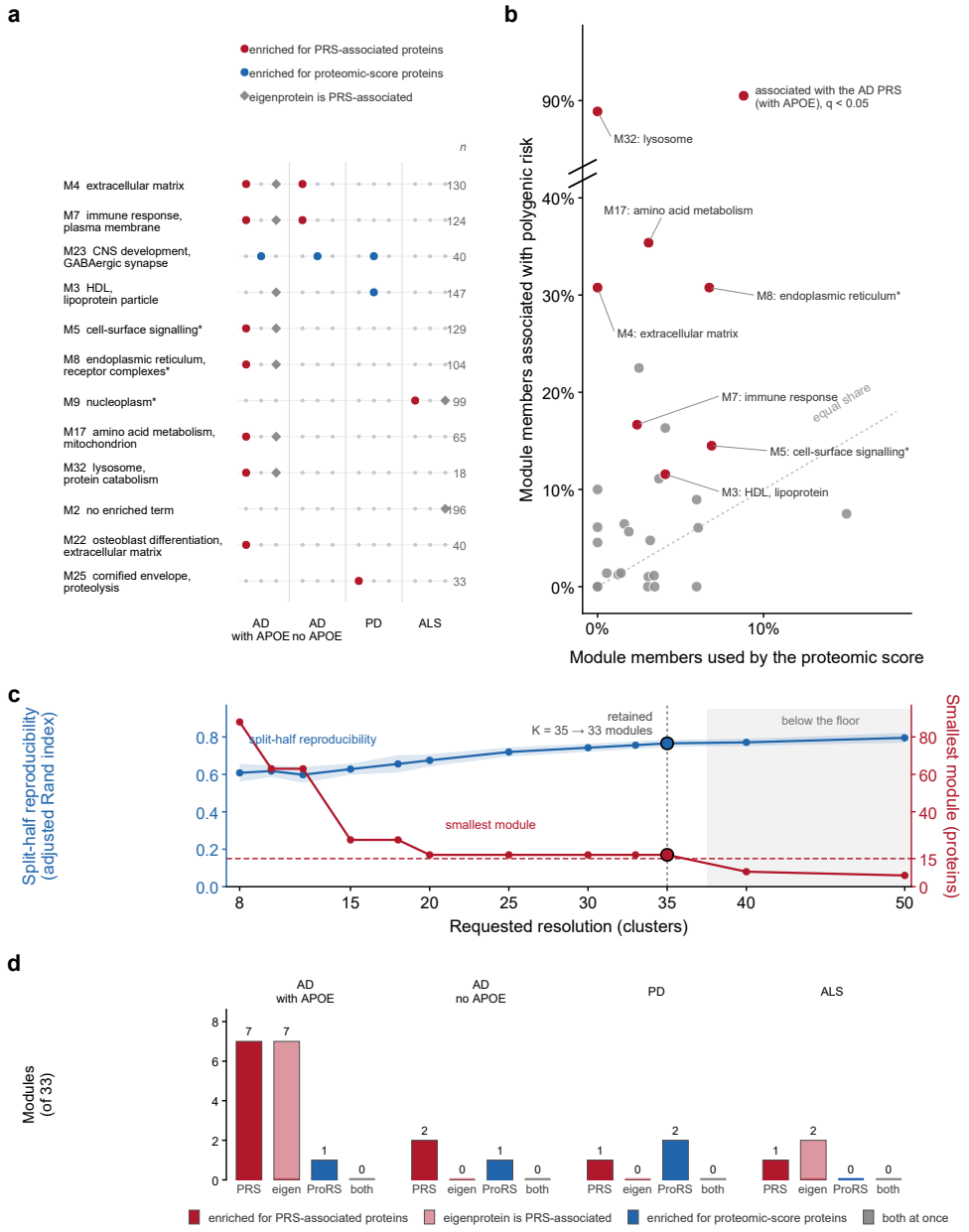


Fig. 2. *Module-resolution analysis. Modules are defined once in the 41,358 participants with no recorded AD diagnosis; the association axes are evaluated in each disease's own analysis population. (a) The 12 modules significant on at least one of the three axes, of the 33 tested. An asterisk marks a module named by its most enriched gene-ontology term where that term does not clear $q < 0.05$. (b) Per module, fraction of ProRS-selected proteins against the fraction of AD with APOE PRS-associated proteins, all 33 modules, filled red where the eigenprotein itself is associated with the AD PRS. (c) Split-half reproducibility and smallest module size across twelve resolutions, with the dashed line marking the floor of 15 proteins per module. (d) Barplot showing the counts of how many modules have eigenprotein associated with PRS, is enriched for ProRS-selected proteins (in $t \geq 7/15$ folds), is enriched for PRS-associated proteins, and is enriched for both ProRS and PRS proteins.*

surface signalling ($q = 4\times10^{-23}$), lysosomal hydrolases ($q = 3\times10^{-23}$), sarcomere ($q = 6\times10^{-10}$) and lipoprotein particles ($q = 2 \times 10^{-3}$) (Fig. 2a). Each module was analyzed for whether its membership was enriched for PRS-associated proteins, whether it was enriched for ProRS-selected proteins (selected in $t \geq 7/15$ folds), and whether its eigenprotein was associated with PRS (Fig. 2a). No module was enriched for both the ProRS-selected and the PRS-associated set in any disease, suggesting that network-level convergence of the two sets remains limited (Fig. 2d). This holds at $K = 40$; at $K = 30$ one module became significant for AD with *APOE* (ProRS enrichment $q = 0.049$ against 0.058 at $K = 35$), reflecting weaker correction over fewer modules rather than new biology. M32 and M23 are unchanged at both resolutions (Supplementary Table S9).

For PD, while 2 modules were enriched for ProRS-selected proteins, 1 was enriched for PRS-associated proteins, and no module's eigenprotein was PRS-associated. Importantly, because only 18 proteins in total met the PRS-association threshold for PD, module-level overlap tests are severely underpowered. Most modules contain 0 or 1 PRS-associated protein, so the absence of significant dual enrichment may reflect limited statistical power rather than definitive evidence of biological divergence at the module level. This contrasts with AD with APOE, where 7 modules were enriched for PRS-associated proteins and 7 module eigenproteins were themselves associated, dominated by the 18-protein lysosomal module M32, 89% of whose members are individually PRS-associated. This enrichment is absent for AD without *APOE* (PRS enrichment $q = 1.0$, eigenprotein $q = 0.058$), which is consistent with a concentration of the *APOE* signal in this module. *APOE* itself belongs to M14, and M32 holds 16 of the 238 PRS-associated proteins. None of the 18 members is selected by the ProRS at $t \geq 7/15$ folds. The two sets of seven overlap in six modules, M4, M5, M7, M8, M17 and M32, which are enriched for PRS-associated members and also associated as a module eigenprotein with PRS, but this pattern dissociates once *APOE* is removed from the score.

Although LASSO may shrink correlated proteins and thus potentially reduce the number of ProRS, PRS protein overlap, their separation here does not depend on the penalty. Refitting every ProRS by elastic net enlarged the selected sets by 17% to 48%, yet the overlap grew no faster than chance, the observed-to-expected ratio for AD with *APOE* moving from 2.76 to 2.22 while the other three models remained indistinguishable from chance, and no module was enriched for both sets under either penalty (Supplementary Figs. S2 and S3, Supplementary Table S10).

3.4. *Module definitions transfer across cohort, and coherence tracks which ones*

Upon defining $K = 33$ modules in UK Biobank, two ways to define module construction in *All of Us* at $K = 33$ were tested. First, modules were built *de novo* in *All of Us* on 2,819 shared proteins in 9,211 disease-free participants gave an adjusted Rand index of 0.582 against the UK Biobank partition. Read against the internal ceiling of 0.686, the median over ten random split-half rebuilds within *All of Us* (range 0.628 to 0.779), that recovers 84.8% of the attainable agreement (Fig. 3a).

Under the independent partitions, co-location was also tested to determine whether the

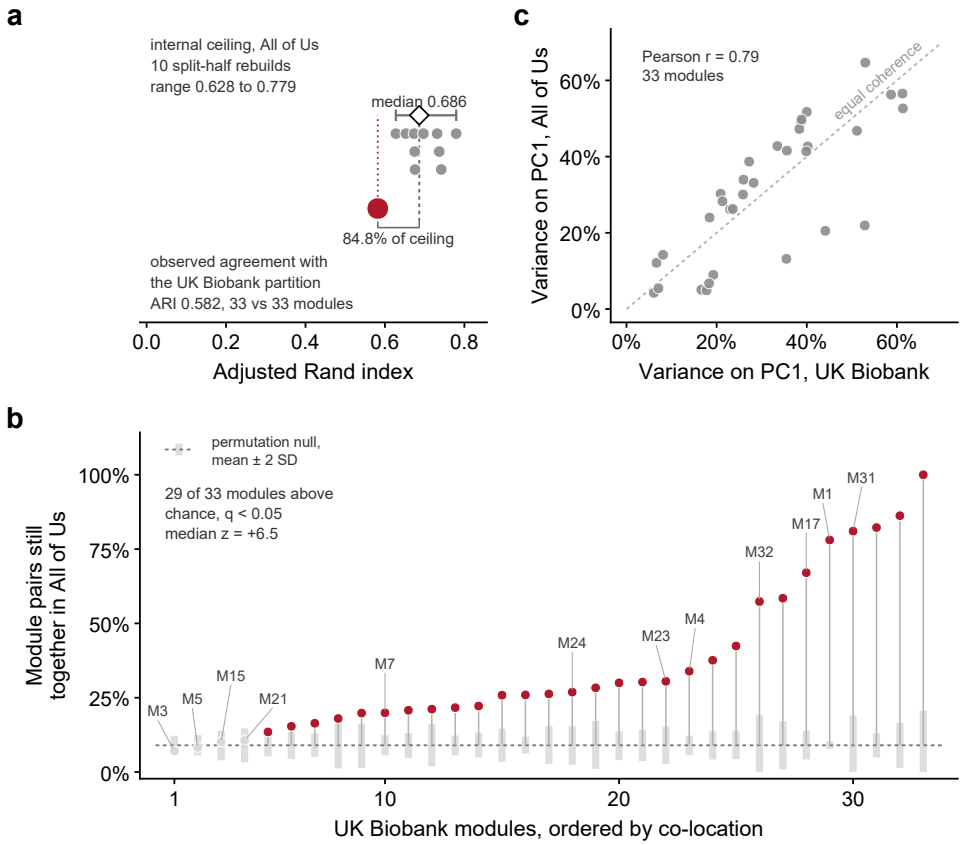


Fig. 3. *Transfer of the UK Biobank module definition to All of Us. All of Us partitioned* ***independently*** *from UK Biobank at* $K = 33$*: (a) Partition agreement between the two cohorts with the ten split-half rebuilds drawn individually and module counts annotated; (b) Per module, the fraction of within-module protein pairs still assigned together by the de novo All of Us partition, against a size-matched permutation null from the same 2,819 shared proteins. All of Us partitioned* ***identically*** *as UK Biobank: (c) Variance on the first principal component, membership held at the UK Biobank definition, measured once in each cohort.*

members of one UK Biobank module are still partitioned together in *All of Us*, and 29 of 33 modules beat a size-matched permutation null at $q < 0.05$, with a median z of +6.5 (Fig. 3b). The lysosomal module M32 keeps 57.4% of its member pairs together against a chance level near 9%, a z of +9.57. M3, the HDL and lipoprotein module, is one of the four failures at $z = -1.19$ (Supplementary Table S11).

The second approach applied UK Biobank's $K = 33$ partition unchanged to *All of Us* and asked whether the modules are equally coherent there. Comparing the variance on the first principal component of the same 33 modules across the two cohorts gave a Pearson correlation of 0.79. Coherence is therefore stable across these two Olink-assayed cohorts. Since both use the same assay platform, this does not establish that modules are free of technical structure, but it does show that internal consistency is reproducible rather than being particular to one sample. (Fig. 3c).

4. Discussion

Proteomic substitution for polygenic risk is never complete, and its size is not explained by heritability alone. Three scores whose liability-scale heritability agrees within one standard error give absorption estimates spanning most of the available range, from substantial and bounded away from zero in AD to estimates in PD that are compatible with little or no absorption (95% CI 0.00–0.38), ordered by whether the score carries a large-effect locus whose product is itself measured in plasma.[32] APOE is heavily implicated in our findings: the absorption gap, the number of PRS-associated proteins and the single lysosomal module all depend on inclusion of the *APOE* region in the score.

M32 is enriched for lysosomal proteins, holding 18 proteins across four enzyme classes. A plasma lysosomal signature of patients carrying *APOE* $\varepsilon 4$ has been reported,[6] and lysosomal *ABCA1* has been placed downstream of cholesterol accumulation in *APOE4* models, which offers a candidate explanation that our associational data do not test.[33] The association of M32 with the score for AD with *APOE* is therefore the expected one and, because it is absent without *APOE*, is consistent with a concentration of the *APOE* signal in this module; however, the proteomic score selects none of its 18 members.

M23 was consistently concentrated for ProRS-selected proteins across AD and PD. It is annotated to central nervous system development and the GABAergic synapse, and several members are established fluid markers of neurodegeneration, including NEFL, GFAP, SYT1 and NPTXR.[34–36] AD and PD were associated with this module through different proteins. AD selected BCAN, GFAP, NPTXR, SYT1 and VGF, PD selected F3, LRTM2, NCAN, PTPRN2, SCG2 and SLITRK1, and NEFL was the only member both scores use. This agreement between AD and PD presents evidence that module-level analysis may capture shared biology that individual proteins cannot.

This study defined two protein sets, those selected by LASSO in the ProRS model and those significantly associated with PRS, and showed the two are largely disjoint. However, the L_1 penalty has no grouping effect.[37,38] Among correlated predictors LASSO retains a subset and shrinks the rest to zero; which member survives varies across resamples, so a protein both PRS-associated and predictive could be dropped for a correlated partner, deflating the overlap.[37,38] The elastic-net refit rules this out. When the ProRS was refit with elastic net, which keeps correlated proteins together instead of dropping some of them, the selected sets of proteins grew in size, but their overlap with the PRS-associated proteins did not rise relative to chance, suggesting that the small overlap observed should not be attributed to the mechanism of LASSO dropping correlated proteins. The two sets may therefore be largely different groups of proteins that carry complementary information, a possibility our data do not directly test. CR1, for example, is shared by both sets in AD, and external Mendelian randomization work has supported its causal role.[39,40]

4.1. *Limitations*

Finally, every estimate rests on very small polygenic signal once *APOE* is removed, where the scores explain under 1% of liability and the three remaining absorption intervals reach lower bounds of 0.12, 0.01 and 0.00. Larger discovery GWAS would resolve this, and phenotype

definition compounds the problem. The FTD model is uninterpretable at this GWAS size and was excluded rather than reported as null, and ALS cases were scored as a single trait without *C9orf72* or *SOD1* stratification, treating a heterogeneous liability as homogeneous. Genotype-stratified replication would separate these. Modules were built without plate adjustment, so some module covariance may be technical. Preservation in *All of Us*, which was processed independently, argues against a cohort-specific artifact, but shared Olink platform structure cannot be excluded, particularly for M2, which has no enriched annotation. Adjusting for age and age squared does not let the relationship between the PRS and the outcome change with age, so a single baseline PRS from a case-control GWAS may distort absorption if genetic effects vary with age, and age-resolved summary statistics were not available. Prevalent cases enter every ProRS training fold, so medication such as cholinesterase inhibitors or levodopa could contribute to the score. They are a minority of cases, and only 5 prevalent AD cases were on a cholinesterase inhibitor at baseline, but we did not examine treatments in PD or ALS or refit without prevalent cases. The number of disease cases in *All of Us* is small, so the disease-specific absorption estimates cannot be interpreted as strongly replicated. The minimal overlap between PRS-associated and ProRS-selected proteins may also reflect a comparison of marginal univariate regressions with conditional LASSO selection rather than distinct biology. Cross-cohort module concordance (ARI = 0.582) should be read in light of k-nearest-neighbor imputation within a European subset in UK Biobank versus mean imputation across a multi-ancestry sample in *All of Us*, and fully harmonized, ancestry-stratified comparisons await larger *All of Us* releases. All PRS-protein and PRS-module relationships are associations in observational data and do not establish that genetic risk acts through these proteins.

Acknowledgments

This work was supported by the National Institutes of Health [R01 AG071470] and [U01 AG068057].

Supplementary Data

The supplementary files are available at https://github.com/czbw/Absorption-and-Coexpression-Modules-Show-Where-PRS-and-ProRS-Diverge-in-Neurodegenerative-Diseases

References

1. B. B. Sun *et al.*, Plasma proteomic associations with genetics and health in the UK Biobank, *Nature* **622**, 329 (2023).
2. E. Assarsson *et al.*, Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability, *PLoS ONE* **9**, p. e95192 (2014).
3. J. Woerner *et al.*, Large-scale evaluation of proteomic and polygenic risk scores reveals complementary contributions to incident disease prediction, *medRxiv* (2025), Preprint.
4. A. Shvetcov *et al.*, APOE ε4 carriers share immune-related proteomic changes across neurodegenerative diseases, *Nature Medicine* **31**, 2590 (2025).
5. K. A. Walker *et al.*, Proteomic analysis of APOEε4 carriers implicates lipid metabolism, complement and lymphocyte signaling in cognitive resilience, *Molecular Neurodegeneration* **19**, p. 81 (2024).

6. N. Amin *et al.*, Interplay between age, APOE $\varepsilon 4$ and the metabolome in plasma and brain in Alzheimer's disease, *Translational Psychiatry* **15**, p. 460 (2025).
7. P. Langfelder and S. Horvath, WGCNA: an R package for weighted correlation network analysis, *BMC Bioinformatics* **9**, p. 559 (2008).
8. E. C. B. Johnson *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, *Nature Neuroscience* **25**, 213 (2022).
9. E. B. Dammer *et al.*, Multi-platform proteomic analysis of Alzheimer's disease cerebrospinal fluid and plasma reveals network biomarkers associated with proteostasis and the matrisome, *Alzheimer's Research & Therapy* **14**, p. 174 (2022).
10. R. Saloner *et al.*, Large-scale network analysis of the cerebrospinal fluid proteome identifies molecular signatures of frontotemporal lobar degeneration, *Nature Aging* **5**, 1143 (2025).
11. V. Emilsson *et al.*, Co-regulatory networks of human serum proteins link genetics to disease, *Science* **361**, 769 (2018).
12. C. Bycroft *et al.*, The UK Biobank resource with deep phenotyping and genomic data, *Nature* **562**, 203 (2018).
13. C. Sudlow *et al.*, UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age, *PLoS Medicine* **12**, p. e1001779 (2015).
14. The All of Us Research Program Investigators, The "All of Us" Research Program, *New England Journal of Medicine* **381**, 668 (2019).
15. The All of Us Research Program Genomics Investigators, Genomic data in the All of Us Research Program, *Nature* **627**, 340 (2024).
16. B. W. Kunkle *et al.*, Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing, *Nature Genetics* **51**, 414 (2019).
17. M. I. Kurki *et al.*, FinnGen provides genetic insights from a well-phenotyped isolated population, *Nature* **613**, 508 (2023).
18. W. van Rheenen *et al.*, Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology, *Nature Genetics* **53**, 1636 (2021).
19. R. Ferrari *et al.*, Frontotemporal dementia and its subtypes: a genome-wide association study, *The Lancet Neurology* **13**, 686 (2014).
20. T. Ge *et al.*, Polygenic prediction via Bayesian regression and continuous shrinkage priors, *Nature Communications* **10**, p. 1776 (2019).
21. C. C. Chang *et al.*, Second-generation PLINK: rising to the challenge of larger and richer datasets, *GigaScience* **4**, p. 7 (2015).
22. G. Leonenko *et al.*, Identifying individuals with high risk of Alzheimer's disease using polygenic risk scores, *Nature Communications* **12**, p. 4506 (2021).
23. B. K. Bulik-Sullivan *et al.*, LD Score regression distinguishes confounding from polygenicity in genome-wide association studies, *Nature Genetics* **47**, 291 (2015).
24. J. Friedman, T. Hastie and R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software* **33**, 1 (2010).
25. N. J. D. Nagelkerke, A note on a general definition of the coefficient of determination, *Biometrika* **78**, 691 (1991).
26. A. Gelman and H. Stern, The difference between "significant" and "not significant" is not itself statistically significant, *The American Statistician* **60**, 328 (2006).
27. E. R. DeLong, D. M. DeLong and D. L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, *Biometrics* **44**, 837 (1988).
28. M. C. Zody *et al.*, Evolutionary toggling of the MAPT 17q21.31 inversion region, *Nature Genetics*

40, 1076 (2008).
29. J. Trowsdale and J. C. Knight, Major histocompatibility complex genomics and human disease, *Annual Review of Genomics and Human Genetics* **14**, 301 (2013).
30. P. Langfelder *et al.*, Is my network module preserved and reproducible?, *PLoS Computational Biology* **7**, p. e1001057 (2011).
31. M. A. Nalls *et al.*, Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies, *The Lancet Neurology* **18**, 1091 (2019).
32. M. Uhlén *et al.*, Tissue-based map of the human proteome, *Science* **347**, p. 1260419 (2015).
33. S. Wang *et al.*, Cellular senescence induced by cholesterol accumulation is mediated by lysosomal ABCA1 in APOE4 and AD, *Molecular Neurodegeneration* **20**, p. 15 (2025).
34. M. Khalil, C. E. Teunissen, M. Otto *et al.*, Neurofilaments as biomarkers in neurological disorders, *Nature Reviews Neurology* **14**, 577 (2018).
35. L. Heimfarth, F. R. S. Passos, B. S. Monteiro *et al.*, Serum glial fibrillary acidic protein is a body fluid biomarker, a valuable prognostic for neurological disease, *International Immunopharmacology* **107**, p. 108624 (2022).
36. E. Camporesi, J. Nilsson, A. Brinkmalm *et al.*, Fluid biomarkers for synaptic dysfunction and loss, *Biomarker Insights* **15**, p. 1177271920950319 (2020).
37. R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 267 (1996).
38. H. Zou and T. Hastie, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301 (2005).
39. L. Belbasis *et al.*, Mendelian randomization identifies proteins involved in neurodegenerative diseases, *Brain* **148**, 2412 (2025).
40. H. Zhan *et al.*, Biomarker identification for Alzheimer's disease through integration of comprehensive Mendelian randomization and proteomics data, *Journal of Translational Medicine* **23**, p. 278 (2025).