

ABSTRACT

Background
 Alzheimer's disease (AD) is a heterogeneous disorder that may consist of disease subtypes unique in their underlying pathological features and disease progression. Studies to date have relied on cross-sectional data to investigate potential AD subtypes, yet it is difficult to disentangle disease subtype from disease stage from cross-sectional data. Consideration of both disease subtype and disease stage in subtype identification from cross-sectional data may provide increased biological validity of identified AD subtypes.

Methods
 We applied the Subtype and Stage Inference (SuStain) algorithm to identify AD subtypes while jointly considering disease stage using cerebrospinal fluid (CSF) multi-platform proteomic data (n=5242 protein measures) from n=296 individuals in the Emory Goizueta Alzheimer's Disease Research Center (ADRC) cohort. We validated the subtypes in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort (n=595 individuals) using the same CSF proteomic measures.

Results
 SuStain identified five subtypes. All subtypes showed progressively worse cognitive function with increasing disease stage. The subtypes identified by SuStain were not identified using five other clustering methods.

BACKGROUND & MOTIVATION

Prior AD subtyping studies
 Recent omics studies in brain and CSF have used data-driven approaches to characterize molecular alterations, identifying 3–5 subtypes in brain^{1,2} and 2–6 subtypes in CSF,^{3–6} differing in molecular phenotype, cognitive impairment, and demographic features but not clustering based on classic AD biomarkers. However, a key limitation shared by all these approaches is the inability to disentangle disease subtype from disease stage simultaneously.

Why SuStain?
 SuStain⁷ simultaneously infers subtypes AND stages. Each subtype follows a distinct ordered sequence of biomarker changes; Markov Chain Monte Carlo (MCMC) sampling estimates posterior distributions over subtype-specific event sequences and individual stage assignments, separating disease type from disease progression.

Key advantages

- Cross-sectional data only — no longitudinal samples required
- Joint subtype + stage inference in a single probabilistic model
- Uncertainty estimates per individual via MCMC posteriors

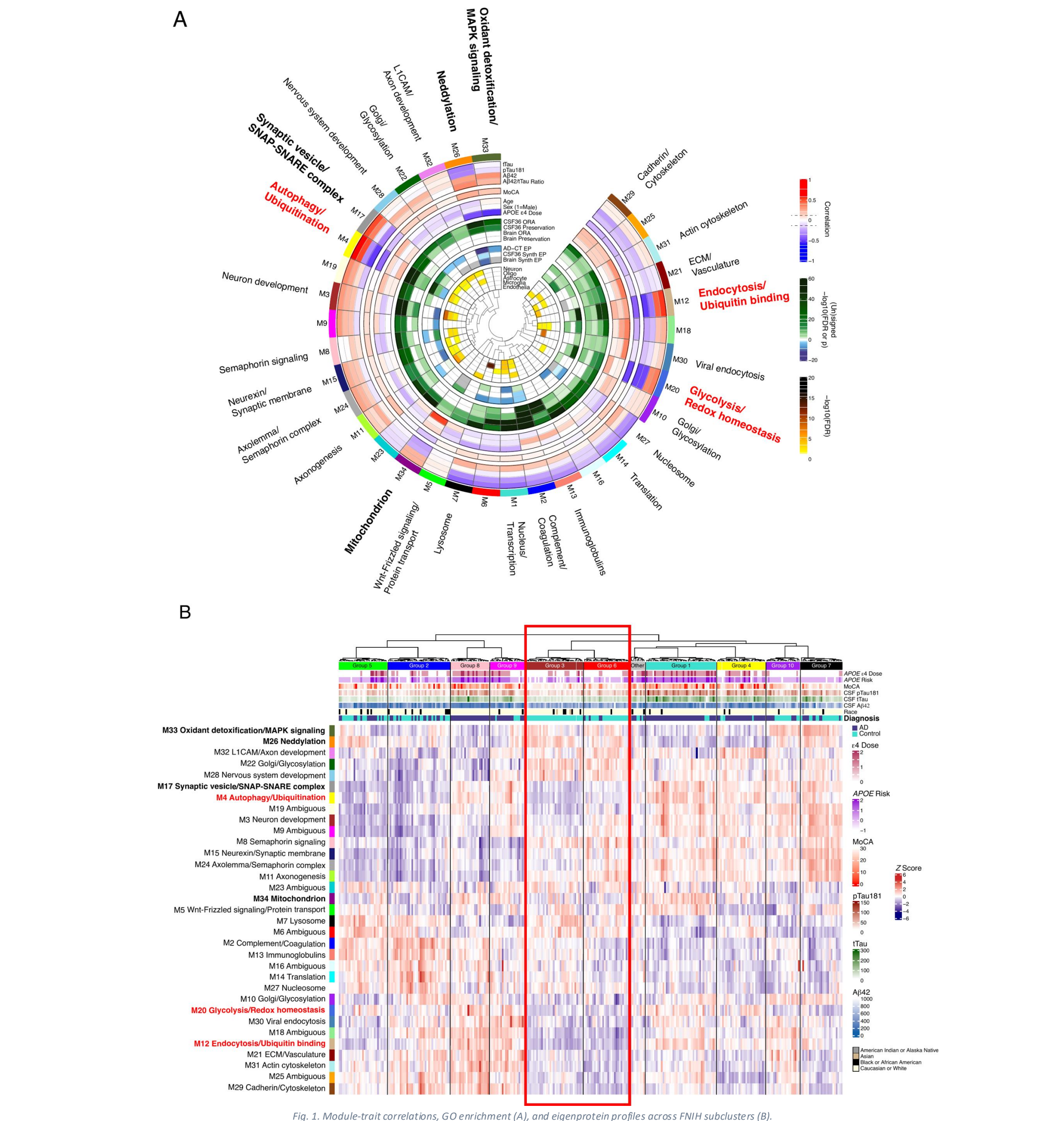
COHORTS & PROTEOMICS DATA

Training: FNIH / Emory ADRC (n=296)
 CSF: n=139 controls, n=157 AD (Aβ42/tau ratio criteria). TMT mass spectrometry + SomaScan 7k → 5,242 proteins. WGCNA: 34 modules → 33 eigenproteins (M1 excluded).

Validation: ADNI (n=595)
 Synthetic eigenproteins: top-20% kME hub proteins per FNIH module (min. 4). 30 ADNI controls selected by Euclidean distance to FNIH controls in t-SNE space. All 33 modules preserved in ADNI through a WGCNA preservation analysis.

Module-trait correlations & GO enrichment
 A) WGCNA co-expression network characterization showing module-trait correlations heatmap and GO enrichment for the 33 module eigenproteins used as SuStain input features. These 33 protein modules cover key biological processes-including synaptic function, protein degradation, immune response, vascular integrity, and metabolism-giving SuStain a broad molecular view to distinguish diverse disease trajectories.

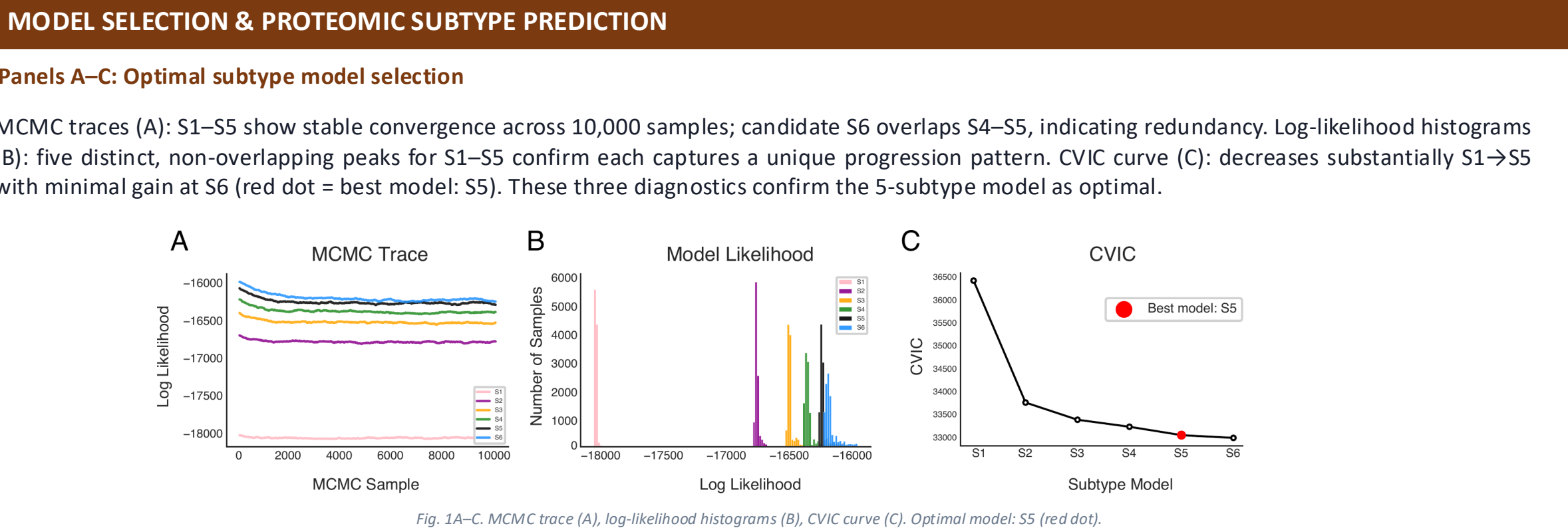
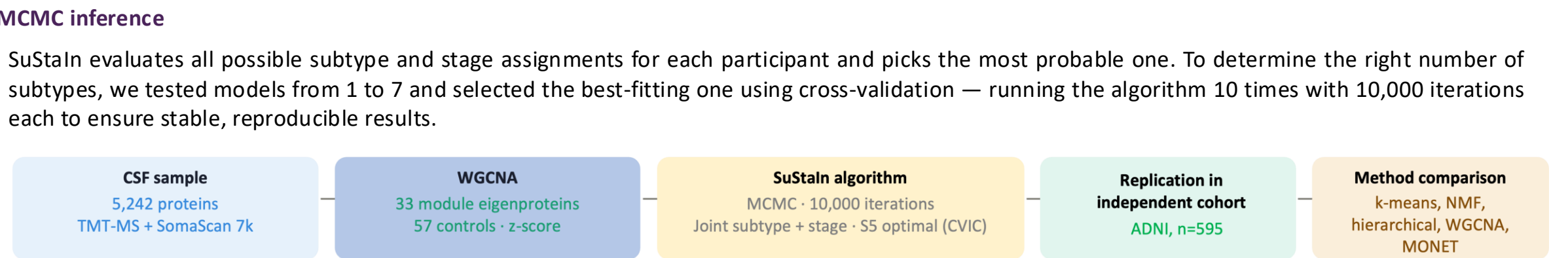
B) Eigenprotein z-score profiles across FNIH participant subclusters, with the red box indicating the 57 clean controls group selected for z-score normalization prior to SuStain modeling based on belonging to the same WGCNA supercluster group and low disease-relevant variability (they are cognitively normal, had CSF Aβ/tau negativity, and low APOE ε4 frequency)⁸.



PROBLEM DEFINITION

SuStain formal framework
 SuStain takes each participant's protein data, compares it to healthy controls, and assigns two things: (1) which disease subtype they belong to, and (2) how far along that subtype's progression they are — separating *what kind of AD* from *how advanced it is*.

MCMC inference
 SuStain evaluates all possible subtype and stage assignments for each participant and picks the most probable one. To determine the right number of subtypes, we tested models from 1 to 7 and selected the best-fitting one using cross-validation — running the algorithm 10 times with 10,000 iterations each to ensure stable, reproducible results.



Proteomic heatmaps — Emory (n=296) & ADNI (n=595)
 Individuals sorted by modal SuStain subtype (S1–S5, color bars at top) and stage within subtype. Rows: 33 module eigenproteins (z-score: red=↑, blue=↓ vs. controls). Clinical metadata tracks: AD status, pTau, Aβ42, pTau/Aβ42, APOE4 dose, age, sex, MoCA (Emory)/ADAS13 (ADNI).

S1 Classic (n=96, pink): progressive ↑M20 (glycolysis/redox homeostasis), M17 (synaptic vesicle/SNAP-SNARE), M4 (autophagy/ubiquitination) with stage; highest APOE ε4 burden; worst MoCA. S2 Metabolic-sparing (n=82, purple): sparing of M20 (glycolysis/redox homeostasis — key differentiator); fewest APOE ε4; highest Aβ42; slowest cognitive decline. S3 Blood-brain barrier (n=75, orange): ↑M2 (complement/coagulation) and M13 (immunoglobulins); pattern consistent with BBB dysfunction and CSF immune activation. S4 ECM/vascular-sparing (n=17, green): sparing of M21 (ECM/vacuature), M31 (actin cytoskeleton); highest tau; most microhemorrhages. S5 Neuronal/synaptic-sparing (n=26, black): sparing of M17 (synaptic vesicle); APOE ε4 homozygote-enriched; highest amyloid burden; most Lewy body co-pathology; fastest cognitive decline.

ADNI right panel: all five subtype patterns reproduced with harmonized synthetic eigenproteins, confirming cross-cohort replicability of the full proteomic subtype structure.

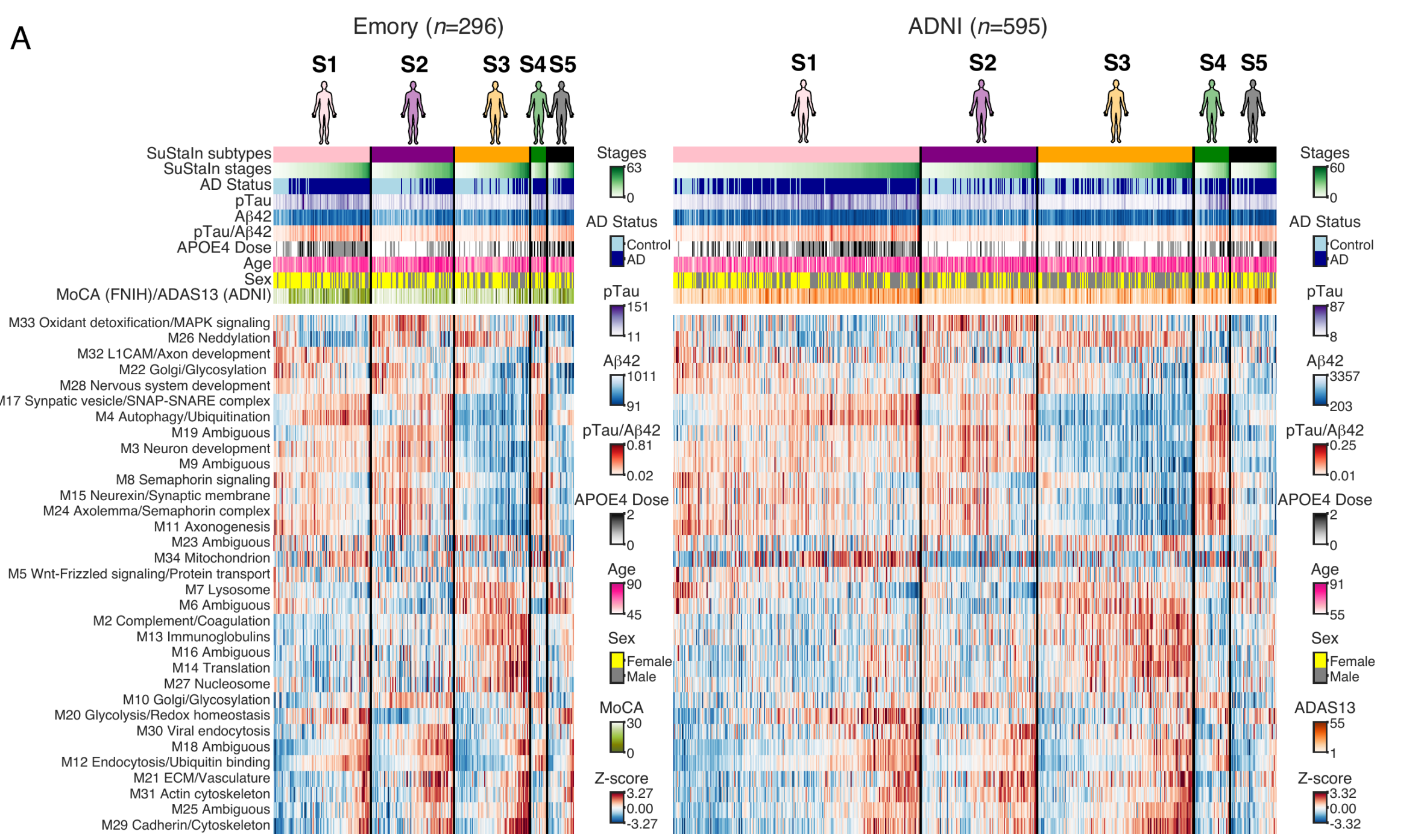


Fig. 2. Proteomic subtype heatmaps for Emory (n=296, left) and ADNI (n=595, right) across 33 module eigenproteins and 5 subtypes. Stage correlates with AD biomarker positivity and worse cognition in all subtypes.

SUBTYPE CLINICAL & MOLECULAR CHARACTERISTICS

Panel B: Stage by AD/control status
 •**Stage vs. diagnosis:** AD individuals show significantly higher SuStain stages than controls within nearly every subtype in both cohorts (p<0.005 for 9/10 comparisons), validating that disease stage tracks diagnostic status — the one exception (ADNI S3, p=0.21) likely reflects cohort composition differences

Panels C–D: Demographics & genetics
 •**Demographics:** AD prevalence, age differ significantly across subtypes in both cohorts (p<0.001–0.002). S1, S4, and S5 are more enriched in AD.
 •**Genetics:** S1 has the highest APOE ε4 burden (73%) and S2 the lowest (30%); S4 in Emory and S5 in ADNI shows reduced APOE ε2 (protective allele) and, along with S1, the highest polygenic hazard score — indicating S1, S4, and S5 carry the greatest genetic AD risk

Panel E: CSF molecular biomarkers
 •**Tau & amyloid biomarkers:** S4 shows the high tau levels (Ttau/pTau) in Emory and ADNI; S2 has the least amyloid burden (highest Aβ42) and S5 has the most (lowest Aβ42, highest AV45-PET; all p<0.001)
 •**Other markers:** α-SAA seeding activity is enriched in S5 (p=0.003), while plasma NfL does not differ significantly across subtypes (p=0.11)

Panels F–G: Imaging & cognitive outcomes
 •**Imaging:** S1 and S5 show the lowest hippocampal volume atrophy (p<0.001).
 •**Cognition:** S2 shows the best cognitive performance and slowest decline, while S5 shows the worst performance and fastest decline (MMSE, CDR-SB, ADAS13 all p<0.001) — confirming the five subtypes capture clinically meaningful differences in disease severity and trajectory.

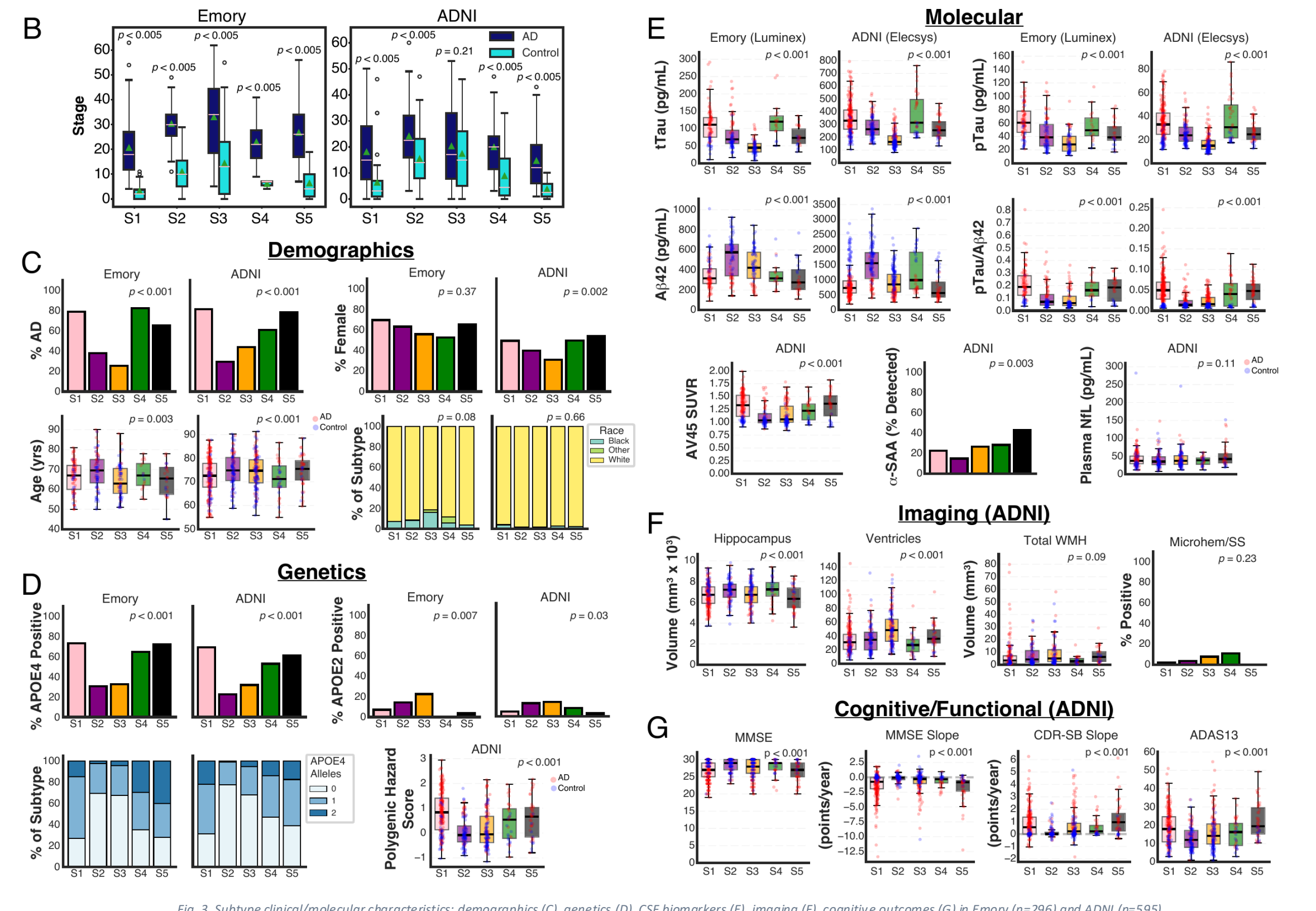


Fig. 3. Subtype clinical/molecular characteristics: demographics (C), genetics (D), CSF biomarkers (E), imaging (F), cognitive outcomes (G) in Emory (n=296) and ADNI (n=595).

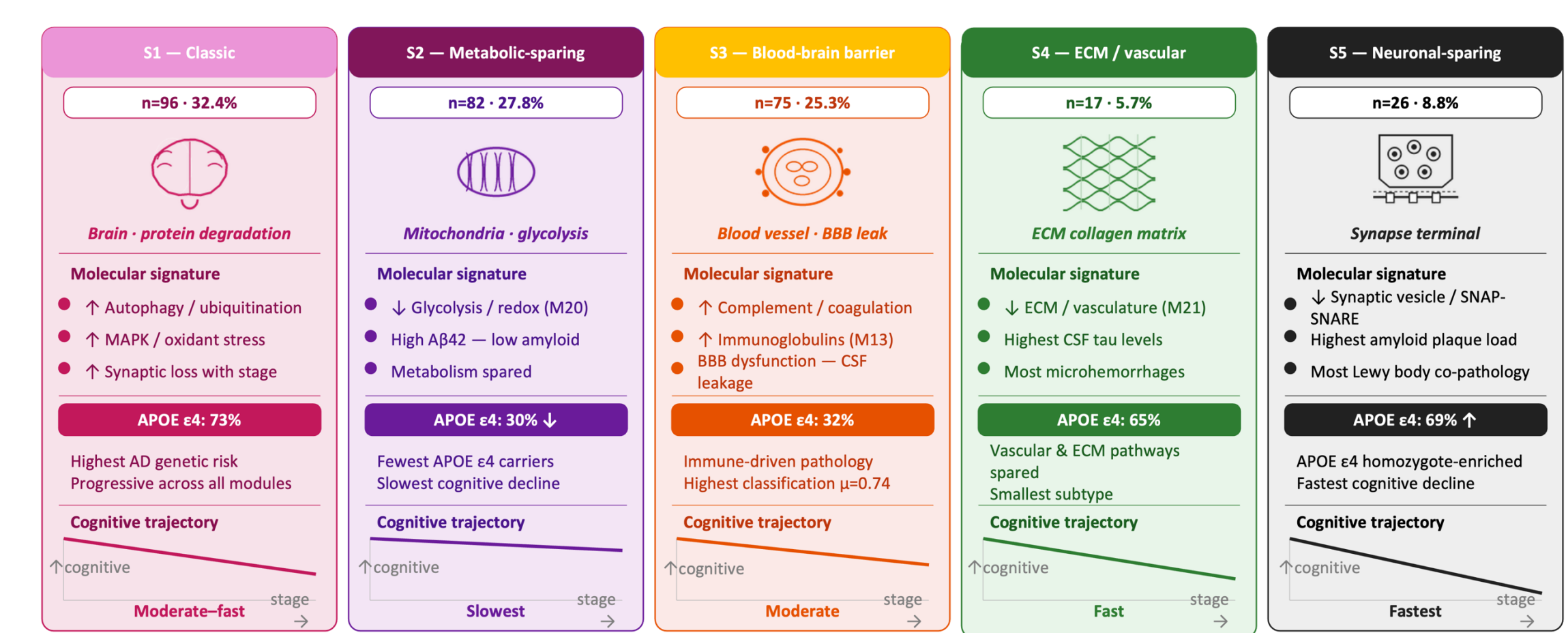


Fig. 4. (A) Pentagon plot (n=296 FNIH), (B) Max classification probability. Dashed = uniform prior (0.2).

SUBTYPE PREDICTION CONFIDENCE

Pentagon plot & classification probability
 Pentagon (A): each vertex = one subtype; position reflects posterior probability. Individuals plotted closer to a vertex have higher probability of belonging to that subtype, indicating greater classification confidence. S3 (orange, μ=0.74): tightest clustering, most molecularly distinct. S4 (green, μ=0.48): widest spread given small n=17 and partial overlap with S1. S5 (black, μ=0.69): tight, consistent with unique APOE ε4-enriched profile.

Boxplots (B): maximum posterior probability per subtype. S3 highest confidence (μ=0.74); S4 lowest (μ=0.48); S1/S2/S5 intermediate. These values reflect the molecular distinctiveness of each subtype. All subtypes show mean confidence well above uniform prior (dashed line, p=0.2), confirming all assignments are above chance.

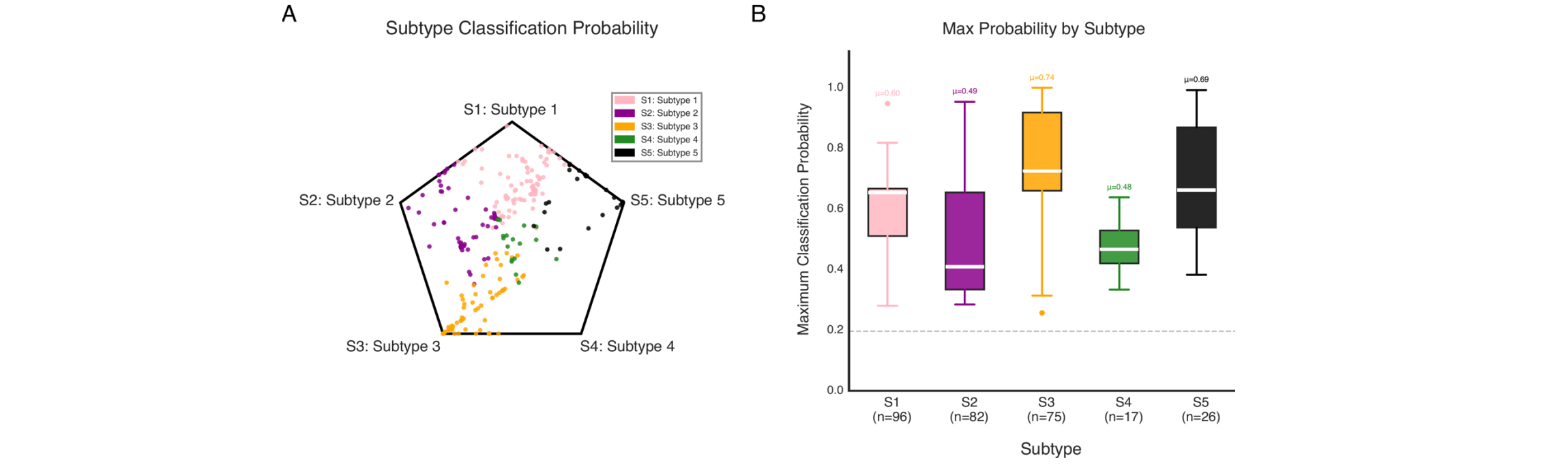


Fig. 5. (A) Stacked bars: subtype overlap across methods, (B) ARI matrix: SuStain row ARI=0.13–0.19 vs. all methods.

CLUSTERING METHOD COMPARISON

Adjusted Rand Index (ARI) matrix and subtype overlap stacked bars
 To assess whether SuStain identifies subtypes unique to its framework, we applied five additional clustering methods to the same data: MONET M1 (K=3), k-means (K=5), NMF (K=5), hierarchical clustering (K=5), WGCNA (K=11). (A): SuStain subtypes are spread across multiple clusters of every other method; most alternative clusters contain individuals from multiple SuStain subtypes. ARI matrix (B): ARI was used to measure the similarity between two clustering methods, adjusted for chance agreement. SuStain shows consistently low ARI with all other methods (range 0.13–0.19). Highest cross-method agreement: MONET M1 vs. hierarchical (ARI=0.52), confirming conventional methods cluster similarly to each other but not to SuStain. Stacked bars confirm that SuStain's joint subtype-and-stage optimization identifies different subtypes compared to all other clustering methods.

CONCLUSIONS
 Five biologically distinct subtypes identified. SuStain identified Classic (S1), Metabolic-sparing (S2), Blood-brain barrier (S3), ECM/vascular-sparing (S4), and Neuronal/synaptic-sparing (S5) in an unsupervised manner, differing across all 33 modules in molecular phenotype, genetic risk, and neuropathology.
 Disease stage separates AD from controls within each subtype. Higher stage → AD biomarker positivity and worse cognition in all subtypes. Stage correlates with pTau (r=0.50–0.65), Aβ42 (r=–0.30–0.52), MMSE (r=–0.19–0.38), hippocampal volume (r=–0.12–0.38).
 All five subtypes replicated in ADNI (n=595). All five subtype patterns reproduced with harmonized synthetic eigenproteins, confirming cross-cohort replicability of the full proteomic subtype structure.
 SuStain captures unique structure missed by other methods. ARI vs. k-means, NMF, hierarchical, WGCNA, MONET M1: 0.13–0.19. confirms SuStain produces different subtyping results.

FUTURE DIRECTIONS

Longitudinal validation: Confirm subtypes track disease progression and biomarker change rates over time.

Treatment stratification: Test differential response to experimental therapeutics (atomoxetine, anti-amyloid) by subtype.

ACKNOWLEDGEMENTS

This work was supported by the National Institutes of Health (NIH) grant R01AG089497. We gratefully acknowledge all participants and clinical staff who contributed to the FNIH/Emory and ADNI cohorts.

REFERENCES

- 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).
- Neff, R.A., *et al.* Molecular subtyping of Alzheimer's disease using RNA sequencing data reveals novel mechanisms and targets. *Sci Adv* **7** (2021).
- Bangs, M.C., *et al.* Proteomic subtyping of Alzheimer's disease CSF links blood-brain barrier dysfunction to reduced levels of tau and synaptic biomarkers. *Alzheimers Dement* **21**, e70830 (2025).
- Moutinho, D., *et al.* Pathophysiological subtypes of mild cognitive impairment due to Alzheimer's disease identified by CSF proteomics. *Trans Neurodegener* **13**, 19 (2024).
- Tijms, B.M., *et al.* Pathophysiological subtypes of Alzheimer's disease based on cerebrospinal fluid proteomics. *Brain* **143**, 3776–3792 (2020).
- 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).
- Young, A.L., *et al.* Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nat Commun* **9**, 4273 (2018).
- Dammer, E.B., *et al.* Proteomic analysis of Alzheimer's disease cerebrospinal fluid reveals alterations associated with APOE epsilon4 and atomoxetine treatment. *Sci Transl Med* **16**, eadh3504 (2024).