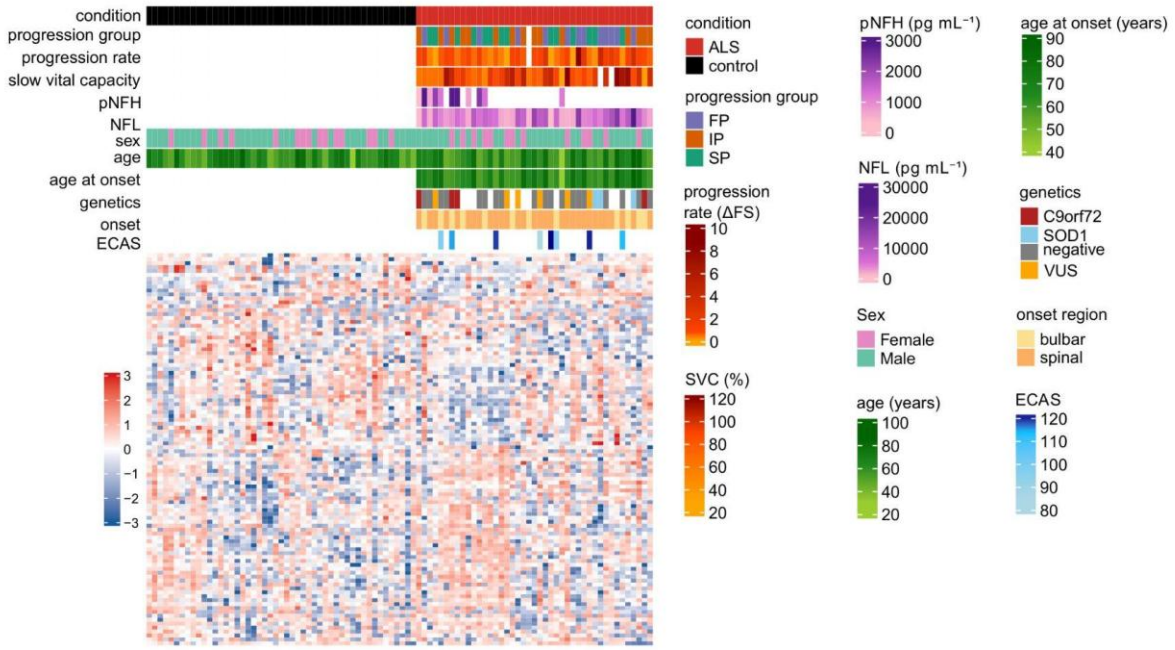
Inventory of Supporting Information

Proteomic Profiling of Cerebrospinal Fluid Identifies Immune and Synapto-Axonal ALS Subtypes

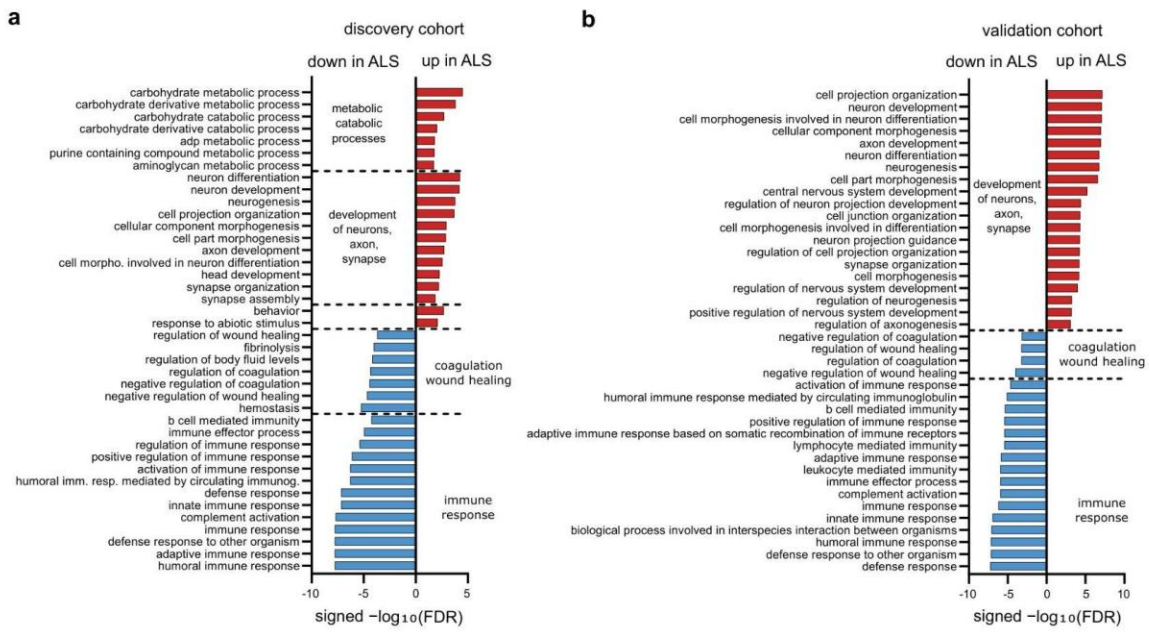
Tzeplaeff and Liu et al., 2025.

List of contents



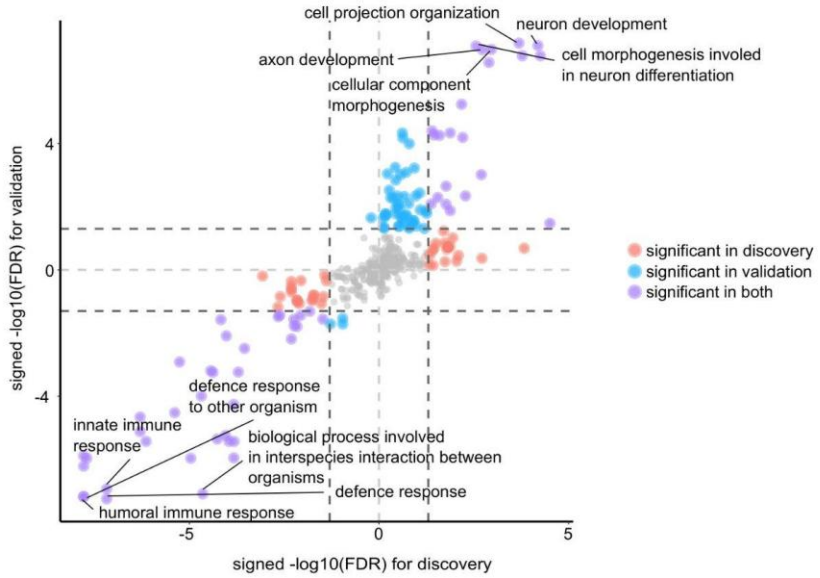
27

28 **Extended Data Fig. 1 | Heatmap and clinical characteristics of the validation cohort.**
29 Heatmap showing the 100 most variable cerebrospinal fluid (CSF) proteins in the validation
30 cohort. Rows represent proteins and columns represent individual samples. Demographic and
31 clinical variables are annotated, including condition, sex, age and age at onset (years), CSF
32 neurofilament light chain (NEFL) and phosphorylated neurofilament heavy chain (pNFH)
33 levels (pg ml^{-1}), slow vital capacity (SVC, %), Edinburgh Cognitive and Behavioural ALS
34 Screen (ECAS) score, site of onset, disease progression rate (Δ Functional score per month),
35 progression group, and genetic status. White indicates missing or unavailable data.



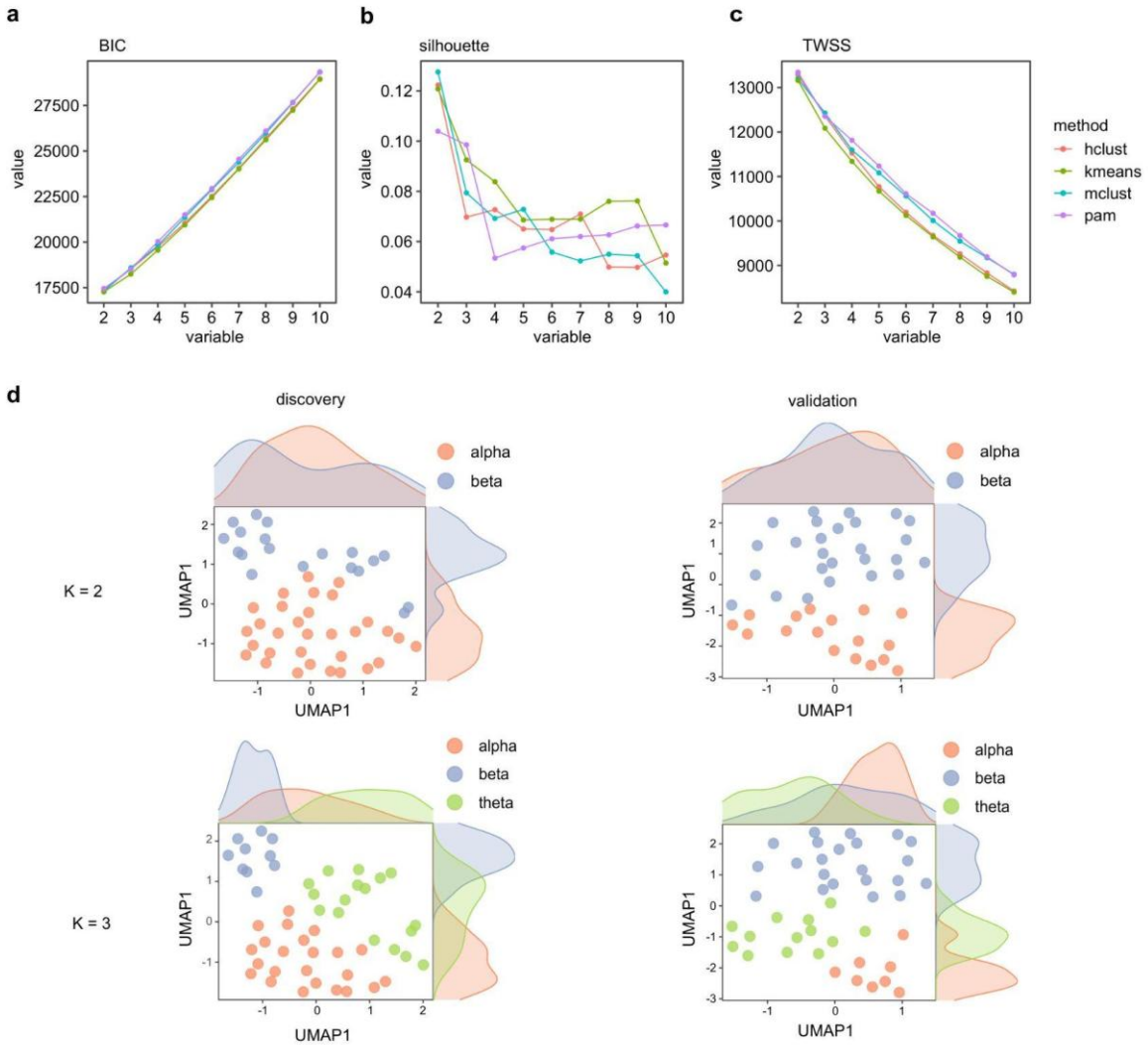
Extended Data Fig. 2 | Gene set enrichment analysis of ALS versus control in the discovery and validation cohorts.

a,b, Gene set enrichment analysis (GSEA) Gene Ontology Biological Process (GO BP) terms in the discovery cohort **(a)** and the validation cohort **(b)**. The top 20 significantly upregulated (red) and downregulated (blue) enriched GO BP terms (FDR < 0.05) in ALS versus controls are shown. Data are represented as signed $-\log_{10}$(FDR). Related GO terms were manually grouped into broader functional categories.

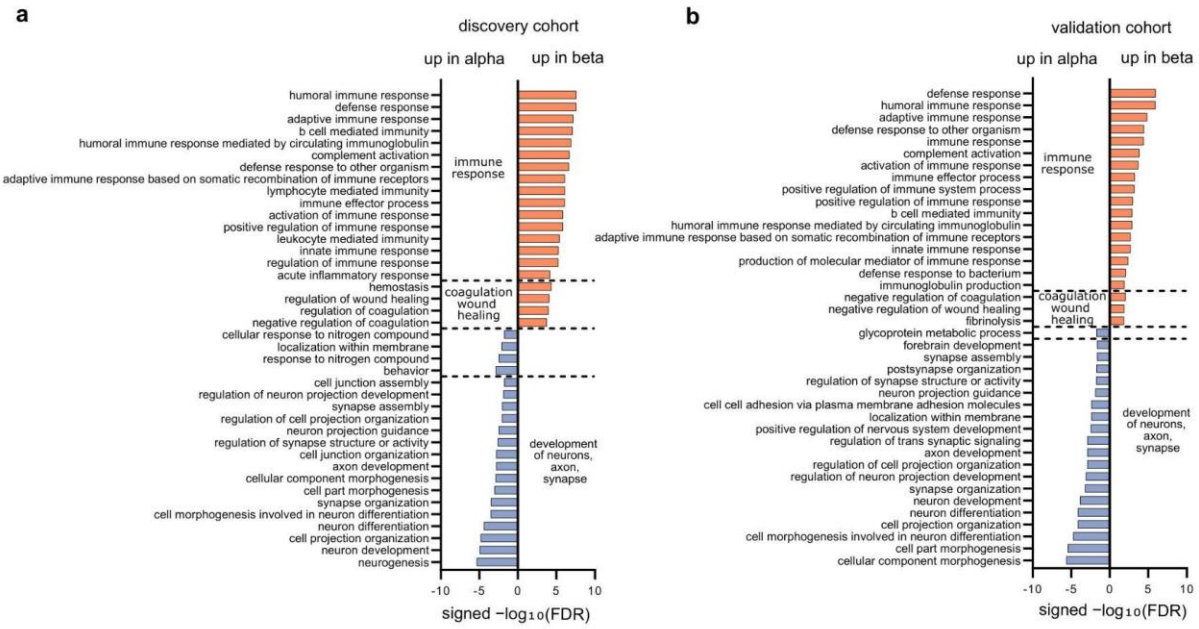


Extended Data Fig. 3 | Concordance of GSEA pathway enrichment between the discovery and validation cohorts.

Scatter plot showing the correlation of commonly enriched Gene Ontology Biological Process (GO BP) terms between the discovery and validation cohorts following comparison of ALS versus controls. Data are represented as signed $-\log_{10}$(FDR), with downregulated biological processes in ALS shown as values < 0 and upregulated processes shown as values > 0. GO BP terms dysregulated exclusively in the discovery cohort are shown in pink, terms dysregulated exclusively in the validation cohort in blue, and terms dysregulated in both cohorts in purple. Correlation and statistical significance were assessed using Pearson's product–moment correlation (r = 0.85; *P* < 2.2e-16).

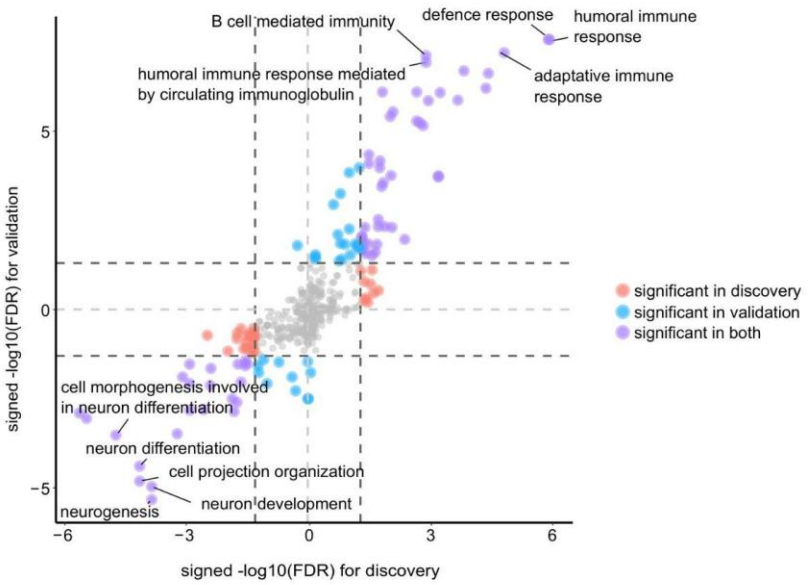


Extended Data Fig. 4 | Clustering selection and visualization of proteomic subtypes. **a-c,** Clustering performance metrics evaluated in the discovery cohort. Bayesian Information Criterion (BIC) **(a)**, silhouette score **(b)**, and total within-cluster sum of squares (TWSS) **(c)** are shown for four clustering algorithms: hierarchical clustering (hclust), k-means clustering (k-means), model-based clustering (mclust), and partitioning around medoids (PAM). Cluster numbers ranging from two to ten (k = 2–10) were tested. The k-means algorithm with two to three clusters achieved an optimal balance between high silhouette scores and biological interpretability and was therefore selected as the optimal clustering approach.
d, Uniform Manifold Approximation and Projection (UMAP) showing the distribution of proteomic samples using k = 2 (top panels) or k = 3 (bottom panels) clustering in the discovery cohort (left) and validation cohort (right). Samples are labeled according to their assignment to the alpha, beta, or theta cluster.

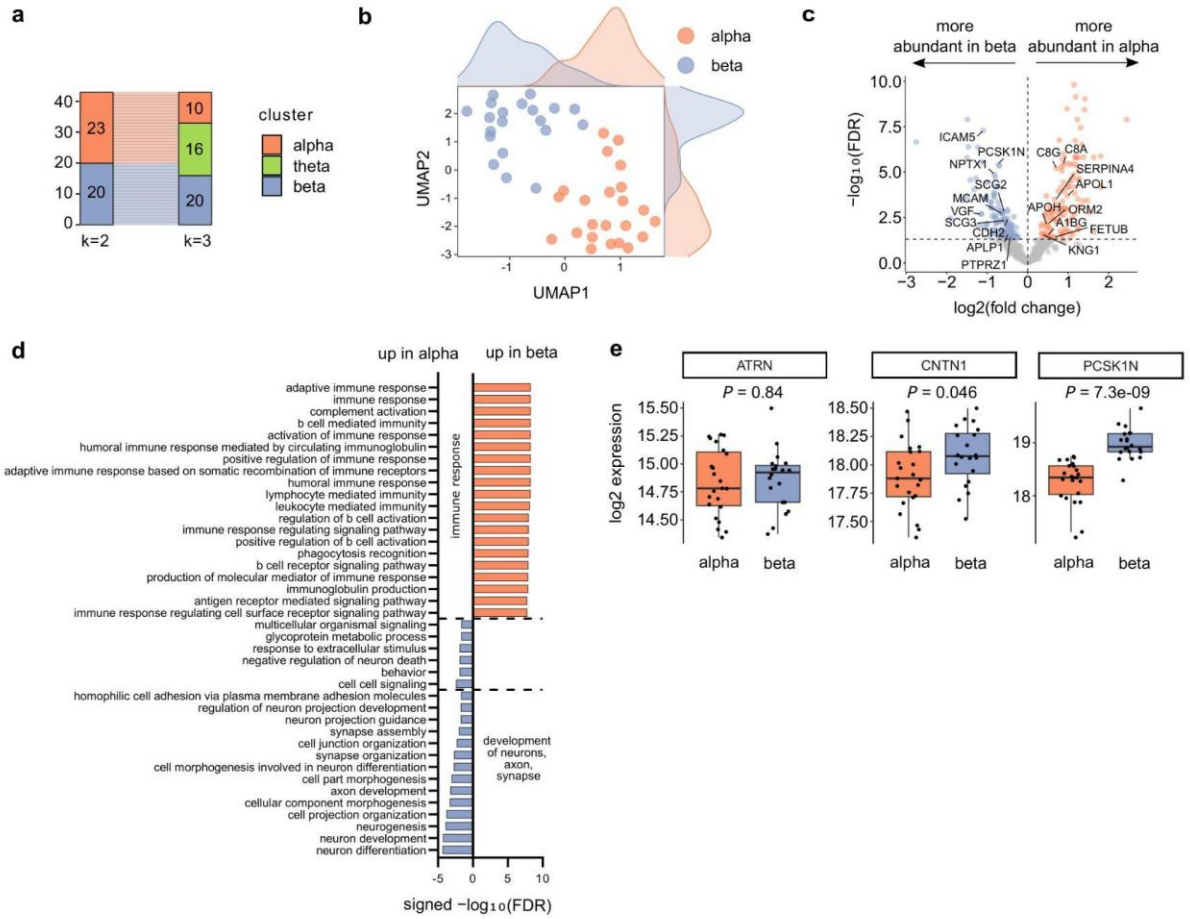


Extended Data Fig. 5 | Gene set enrichment analysis comparing alpha and beta subtypes in the discovery and validation cohorts.

a,b, Gene set enrichment analysis (GSEA) of Gene Ontology Biological Process (GO BP) terms identifying the top 20 significantly enriched pathways (FDR < 0.05) distinguishing the alpha and beta clusters in the discovery cohort **(a)** and validation cohort **(b)**. GO BP terms more enriched in the beta cluster are shown in blue, whereas those more enriched in the alpha cluster are shown in orange. Data are represented as signed $-\log_{10}$(FDR). Related GO terms were manually grouped into broader functional categories.



Extended Data Fig. 6 | Concordance of GSEA pathway enrichment between alpha and beta subtypes in the discovery and validation cohorts.
Scatter plot showing the correlation of commonly enriched Gene Ontology Biological Process (GO BP) terms between the discovery and validation cohorts following comparison of alpha versus beta subtypes. Data are represented as signed $-\log_{10}$(FDR), with GO BP terms more enriched in the beta cluster shown as values < 0 and those more enriched in the alpha cluster shown as values > 0. GO BP terms dysregulated exclusively in the discovery cohort are shown in pink, those dysregulated exclusively in the validation cohort in blue, and those dysregulated in both cohorts in purple. Correlation and statistical significance were assessed using Pearson's product–moment correlation (r = 0.85; *P* < 2.2e-16).



Extended Data Fig. 7 | Proteomic clustering and subtype characterization in the external cohort.

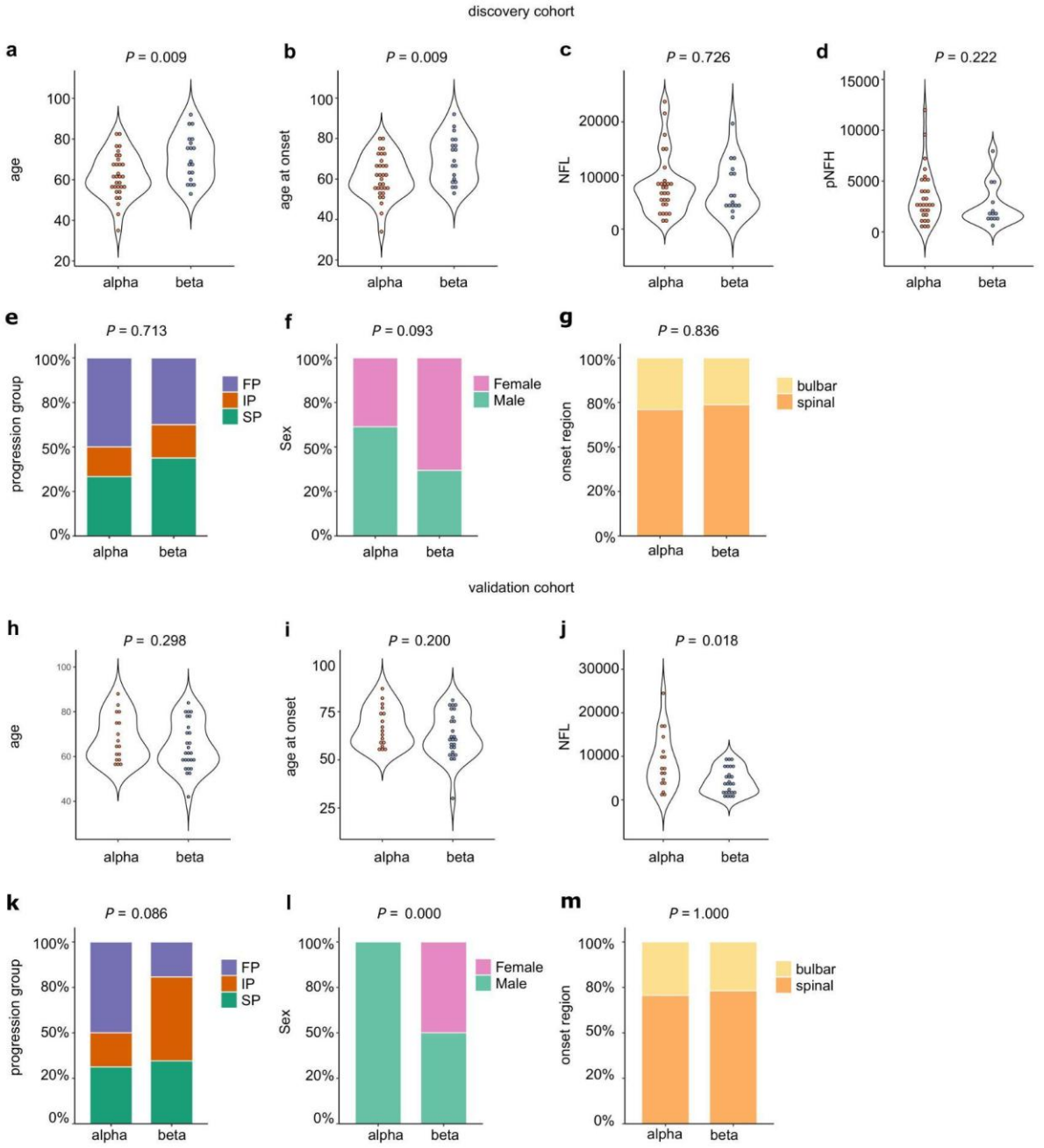
a, Number of ALS patients assigned to the alpha, beta, or theta cluster in the external cohort (EC; *n* = 43 ALS).

b, Uniform Manifold Approximation and Projection (UMAP) showing the distribution of proteomic samples using k = 2 clustering in the external cohort. Samples are labeled according to assignment to the alpha (orange) or beta (blue) cluster.

c, Volcano plot of differentially expressed proteins between the alpha and beta clusters in the external cohort. The x-axis shows $\log_2$ fold change and the y-axis shows $-\log_{10}$(FDR). Proteins more abundant in the beta cluster are shown in blue and those more abundant in the alpha cluster in orange. Statistical testing was performed using a two-sided test with Benjamini–Hochberg correction for multiple comparisons.

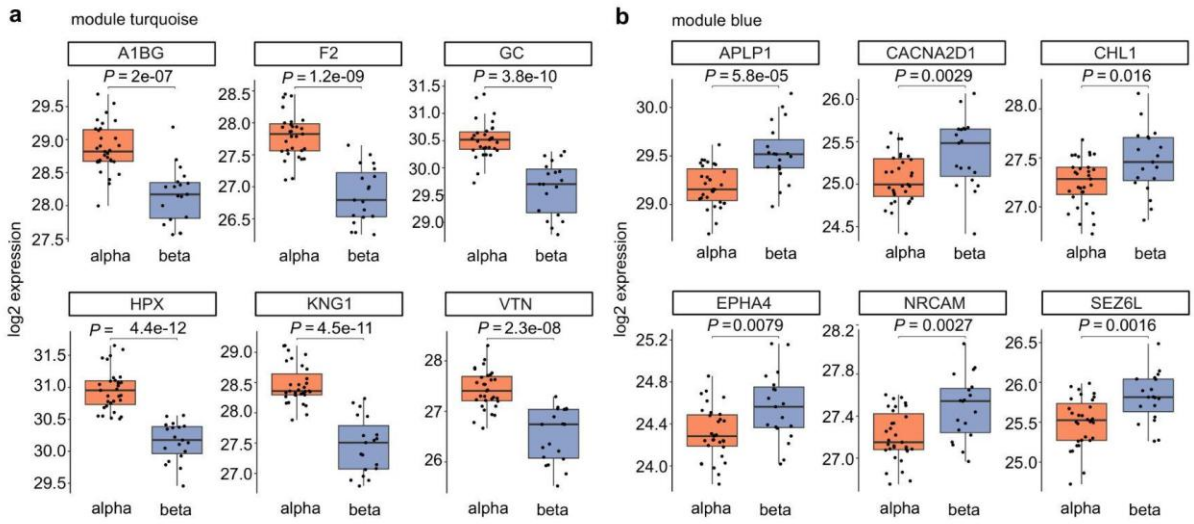
d, Gene set enrichment analysis (GSEA) of Gene Ontology Biological Process (GO BP) terms identifying the top 20 significantly enriched pathways (FDR < 0.05) distinguishing the alpha and beta clusters in the external cohort. GO BP terms more enriched in the beta cluster are shown in blue, whereas those more enriched in the alpha cluster are shown in orange. Data are represented as signed $-\log_{10}$(FDR). Related GO terms were manually grouped into broader functional categories.

e, Box plots showing expression levels of ATRN, CNTN1, and PCSK1N in the external cohort.



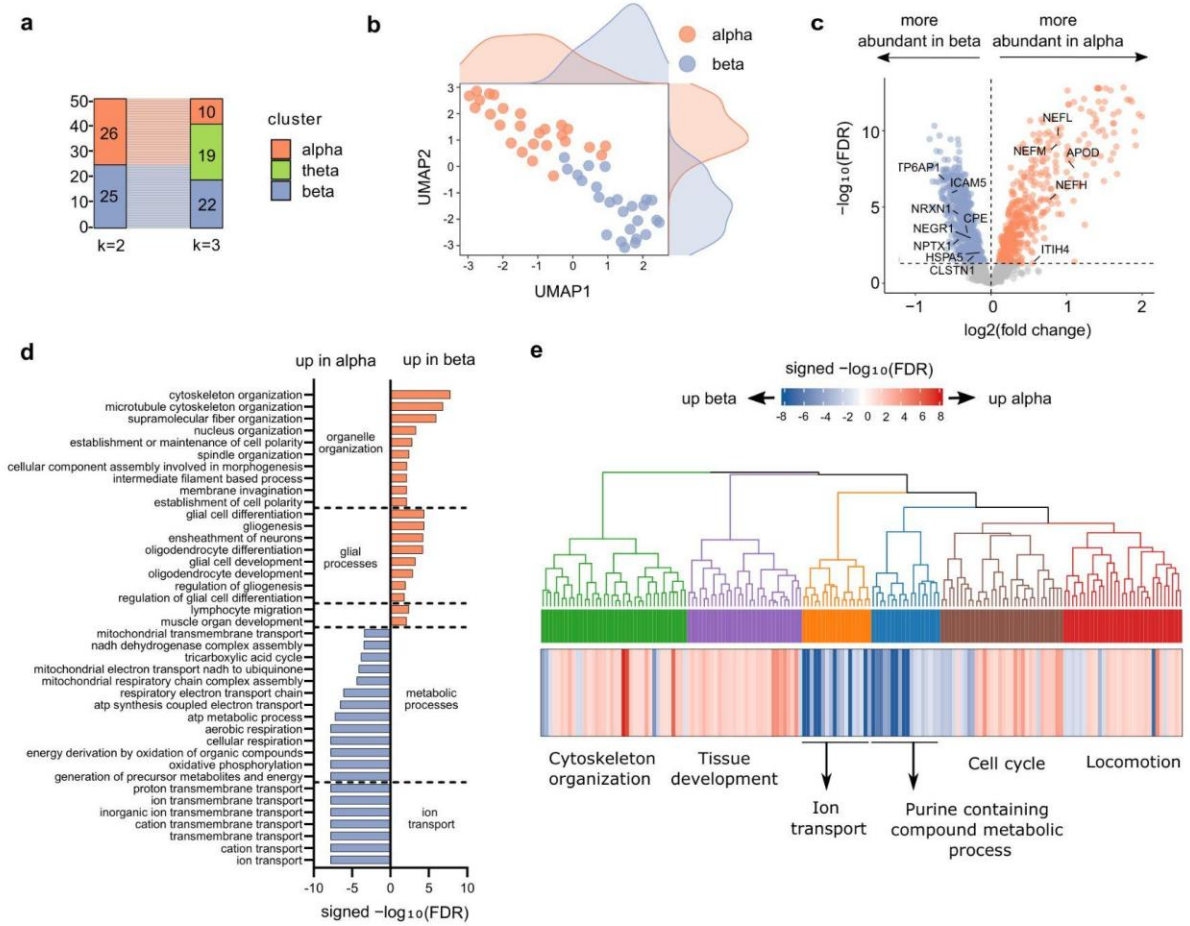
Extended Data Fig. 8 | Clinical characteristics stratified by ALS alpha and beta subtypes in the discovery and validation cohorts.

a-g, Clinical characteristics stratified by alpha and beta subtypes in the discovery cohort. **h-m,** Clinical characteristics stratified by alpha and beta subtypes in the validation cohort. Shown are mean age (years) (a,h), mean age at onset (years) (b,i), cerebrospinal fluid neurofilament light chain (NEFL, pg ml^{-1}) (c,j), cerebrospinal fluid phosphorylated neurofilament heavy chain (pNFH, pg ml^{-1}) (d), disease progression group (e,k), sex (f,l), and site of onset (g,m). Statistical analyses are described in the Methods. $P < 0.05$ was considered statistically significant.



Extended Data Fig. 9 | Top 6 proteins of the turquoise and blue modules in the discovery cohort.

a,b, Boxplot showing the expression levels of the top 6 proteins in the turquoise **(a)** and the blue **(b)** module. Each dot represents an individual sample. The box indicates the interquartile range (IQR), the line within the box represents the median, and the whiskers extend to 1.5×IQR. $P < 0.05$ was considered statistically significant.



Extended Data Fig. 10 | Clustering of brain proteomic samples.
a, Number of ALS patients within the alpha, beta, or theta cluster in the brain proteomic dataset (n = 51 ALS).
b, UMAP representation of the human brain proteomic samples for k = 2 clustering. Labels indicate samples assigned to the alpha (orange) or beta (blue) cluster.
c, Volcano plot showing dysregulated proteins between the alpha and beta clusters in the external cohort. X-axis: $\log_2$ fold change; Y-axis: -$\log_{10}$ (FDR). Proteins significantly more abundant in beta are shown in blue, and those more abundant in alpha in orange. Statistics were calculated using a two-sided test with Benjamini-Hochberg correction. Proteins previously identified in the CSF alpha and beta clusters are labeled.
d, Gene Set Enrichment Analysis (GSEA) of the top 20 significant Gene Ontology Biological Processes (GO BP; FDR < 0.05) in the alpha and beta clusters of the external cohort. Blue indicates enrichment in beta, orange in alpha. Data are shown as -$\log_{10}$ (FDR). Related GO terms were manually grouped under general categories.
e, GSEA of GO BP for alpha vs beta comparisons in the discovery, validation and external cohort. Significantly enriched GO BP terms (FDR < 0.05) were grouped based on semantic similarity, consolidating related terms into six categories named after the highest similar term. Data are represented as signed -$\log_{10}$(FDR), with blue indicating up in alpha and red up in beta.