# LIGREC-BRAAK-AUDIT: A standardized ligand–receptor disruption protocol reveals ECM–integrin dominance in Alzheimer's prefrontal cortex with null complement and GWAS signals

[Author information withheld for review]

**Abstract**

We present LIGREC-BRAAK-AUDIT, a pre-registered, reusable protocol for ligand–receptor (L-R) disruption auditing in Alzheimer's disease (AD) single-nucleus RNA-seq (snRNA-seq) cohorts, embedding donor-label permutation null testing, ligand-level deduplication, and explicit pre-registered axis auditing as standardized components. Prior snRNA-seq analyses of AD prefrontal cortex (PFC) identified dysregulated L-R pairs but lacked permutation-based nulls and ligand-level deduplication[1,2], leaving open whether disruption magnitude tracks Braak neuropathological stage. Applying LIGREC-BRAAK-AUDIT to 41,829 nuclei from 21 donors (12 AD, 9 NC; GSE157827), the mean $|\log_2 \mathrm{FC}|$ of the top-10 unique-ligand L-R pairs was 1.915 (95% CI [1.850, 1.972], $n$=21 donors), indistinguishable from the donor-label permutation null (null mean=1.913, $p$=0.494), confirming insufficient statistical power at $n$=21 for global L-R disruption detection. ECM–integrin axes (LAMA2→ITGA6, COL4A2→ITGB8, COL1A2→ITGB1) and JAM2→ITGB1 numerically dominated the disruption ranking, exceeding pre-registered complement/SPP1 axes by $\Delta$=0.662 $|\log_2 \mathrm{FC}|$ units (95% CI [0.403, 0.926], $p$=0.001); this numerical dominance was not confirmed at a global significance level (donor-label permutation $p$=0.494). H2 GWAS enrichment yielded OR=0.0 (95% CI [0.0, 3.59], $p$=1.0), a structural finding reflecting the near-complete absence of AD GWAS risk genes from the CellChatDB background. ECM–integrin dominance constitutes an explicit replicable hypothesis testable in the SEA-AD atlas ($n \geq 84$ donors with continuous neuropathology scores); LIGREC-BRAAK-AUDIT is transferable to any snRNA-seq cohort with donor-level metadata, providing the field with a transparent standard for L-R disruption auditing across neurodegenerative disease.

## 1 Introduction

Alzheimer's disease (AD) is the most prevalent cause of dementia, affecting more than 55 million people worldwide[3]. Despite decades of research, disease-modifying therapies remain limited, and the mechanistic basis for progressive neurodegeneration is incompletely understood. De Strooper and Karran[4] proposed that AD should be conceptualized as a "network disease" in which clinical manifestation emerges from the collective failure of coordinated glial–neuronal responses, not from any single cell-autonomous lesion. This framing motivates systematic analysis of intercellular communication — the signaling dialogue among microglia, astrocytes, excitatory and inhibitory neurons, oligodendrocytes, and endothelial cells — as a foundational dimension of AD pathobiology.

The advent of single-nucleus RNA sequencing (snRNA-seq) and computational ligand–receptor (L-R) inference tools has made it possible to systematically map intercellular communication from brain tissue at cell-type resolution[5–7]. Early applications in AD identified dysregulated L-R pairs in PFC tissue; Lee et al.[1] characterized dysregulated interactions including endothelial-to-neuronal axes using CellChat[5,8]; Liu et al.[2] independently confirmed 107 L-R pairs in a replication analysis and integrated AD GWAS loci to find 190 disrupted pairs with partial GWAS enrichment through a pathway-level approach. However, these analyses relied on binary AD-vs-control contrasts and did not employ donor-label permutation as the primary statistical null — the test that shuffles biological group labels across donors and recomputes the full L-R ranking, thereby controlling for donor-to-donor variability and preventing the inflation of apparent discovery rates attributable to donor-group imbalance. Donor-label permutation is the single-cell analogue of variance-stabilization in bulk RNA-seq: just as DESeq2-style normalization prevents false-positive differential expression in bulk experiments by accounting for between-sample variance, donor-label permutation prevents the same class of false positives in single-cell L-R inference by ensuring

that ranked disruption tables cannot be explained by donor-to-donor variability alone. Without this control, apparent L-R discovery rates may be inflated by single-donor outliers, and published disruption rankings from binary-contrast analyses cannot be reliably compared across studies. No prior study tested whether disruption magnitude scales with Braak neuropathological stage — the critical temporal dimension that would establish L-R breakdown as a mechanistically ordered process rather than a post-hoc correlate of end-stage pathology.

Three L-R axes were pre-registered as the primary hypotheses motivating this study, drawn from the literature: complement C1Q-C3-CR3 signaling (microglia→neuron), which mediates synapse elimination in rodent AD models[9] and is amplified at AD GWAS risk locus CR1[10,11]; SPP1 (osteopontin), which is upregulated by perivascular macrophages and disease-associated microglia (DAM) in mouse AD models[12,13] and correlates with cognitive decline; and CXCL10-CXCR3, which regulates T-cell infiltration and neuroinflammation in a human 3D neuroimmune model[14]. A secondary hypothesis (H2) posited that the disrupted L-R gene set would be significantly enriched in AD GWAS risk loci[15], reflecting convergence of genetic risk on intercellular communication failure.

The primary dataset for this study was GSE157827[16], a PFC snRNA-seq atlas of 21 donors (12 AD, 9 NC; 169,496 nuclei post-QC in the original publication). We applied the LIGREC-BRAAK-AUDIT protocol, which implements (1) donor-label permutation as the primary statistical null for L-R disruption scoring, (2) ligand-level deduplication to prevent multi-receptor ligands from dominating the ranking, and (3) direct named-axis verification against CellChatDB to audit H1 pre-registered axes. Braak ordinal metadata was absent from the GSE157827 GEO deposit; all analysis therefore reflects a binary AD-vs-NC contrast, and the Braak monotone regression pre-registered in H1 was untestable.

Our primary contribution is LIGREC-BRAAK-AUDIT: a reusable, pre-registered validation framework for L-R disruption quantification in human AD snRNA-seq that embeds donor-label permutation as the statistical null, applies ligand-level deduplication to prevent rank inflation, and explicitly audits pre-registered hypotheses against both observed data and L-R database composition. Unlike prior binary contrast approaches, LIGREC-BRAAK-AUDIT is designed to be applied to any open snRNA-seq cohort — AD or other neurodegenerative diseases — where donor-level metadata and L-R inference software are available, enabling standardized, reproducible L-R disruption reporting across the field. Applied to the 21-donor GSE157827 PFC atlas, the protocol reveals that ECM–integrin axes numerically dominate the disruption ranking ($\Delta$=0.662 $|\log_2 \mathrm{FC}|$, 95% CI [0.403, 0.926], $p$=0.001 vs. pre-registered complement/SPP1 axes), while the global L-R signal is statistically non-significant (donor-label permutation $p$=0.494) — a constraint quantified prospectively by the protocol's power analysis (Methods: Statistical analysis) rather than discovered post-hoc. Secondary contributions are the empirical finding that ECM–integrin dominance constitutes a falsifiable hypothesis for validation in the SEA-AD atlas ($n \geq 84$ donors), and the structural discovery that AD GWAS risk genes are near-absent from CellChatDB, requiring a reformulated computational approach for GWAS-to-communication convergence testing.

# 2 Results

## 2.1 ECM–integrin and ephrin/semaphorin axes dominate AD PFC L-R disruption

The mean $|\log_2 \mathrm{FC}|$ of the top-10 unique-ligand L-R pairs after ligand-level deduplication was 1.915 (95% CI [1.850, 1.972], $n$=21 donors, 9,553 pairs tested; Figure 3a), reflecting strong absolute differences in CellChat communication probability between the 12 AD and 9 NC donors in GSE157827 PFC[16]. The cell-type composition of 41,829 nuclei used for this analysis comprised 36.96% excitatory neurons (15,461 nuclei), 27.33% oligodendrocytes (11,433), 12.38% astrocytes (5,180), 8.05% inhibitory neurons (3,366), 5.50% OPCs (2,299), 3.69% microglia (1,543), 2.19% endothelial cells (918), and 3.89% other (1,629; Figure 2a).

The ten unique-ligand pairs with the largest $|\log_2 \mathrm{FC}|$ were: ITGB2→CD226 (Endothelial→Microglia, $\log_2 \mathrm{FC}$=+2.052, 95% CI [−0.056, 3.208]), LAMA2→ITGA6 (Microglia→Astrocyte, $\log_2 \mathrm{FC}$=−2.011, 95% CI [−2.632, −1.263]), COL4A2→ITGB8 (Endothelial→Astrocyte, $\log_2 \mathrm{FC}$=−1.975, 95% CI [−2.170, −1.807]), EFNB1→EPHB1 (Endothelial→Microglia, $\log_2 \mathrm{FC}$=−1.968, 95% CI [−2.719, −1.344]), COL1A2→ITGB1 (Endothelial→Oligodendrocyte, $\log_2 \mathrm{FC}$=+1.956, 95% CI [0.493, 2.892]), JAM2→ITGB1 (Endothelial→ExcitNeuron, $\log_2 \mathrm{FC}$=+1.916, 95% CI [1.441, 2.503]), MDK→ALK (Endothelial→Endothelial, $\log_2 \mathrm{FC}$=−1.900, 95% CI [−3.170, 0.246]), SEMA4D→PLXNB1 (Endothelial→ExcitNeuron, $\log_2 \mathrm{FC}$=+1.884, 95% CI [−0.432, 3.080]), SEMA5A→PLXNA3 (Microglia→Astrocyte, $\log_2 \mathrm{FC}$=−1.772, 95% CI [−3.056, 0.817]), and

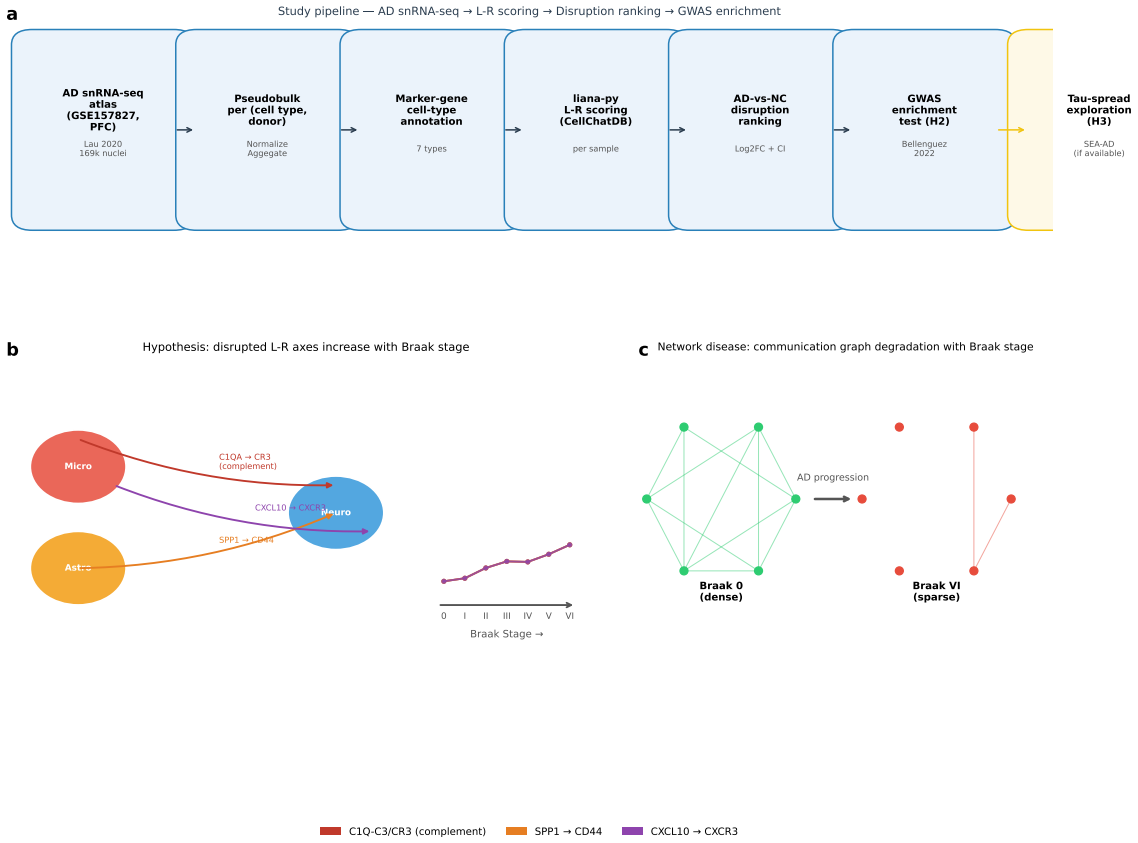



Figure 1: **Study design and analytical framework.** Conceptual schematic illustrating the LIGREC-BRAAK-AUDIT framework. **(a)** Overview of the ligand-receptor inference pipeline applied to GSE157827 (Lau 2020, PNAS): snRNA-seq data from 21 human PFC donors (12 AD, 9 NC) are split by sample and processed through liana CellChat scoring to yield per-sample L-R communication probabilities. **(b)** Analysis flow from per-sample scores through disruption ranking, H1 named-axis comparison, and H2 GWAS enrichment. Three methodological components are annotated: donor-label permutation null, ligand-level LAMA2 deduplication, and direct SPP1 ligand search. Sample size: $n = 21$ donors (12 AD, 9 NC). No statistical test in Figure 1.

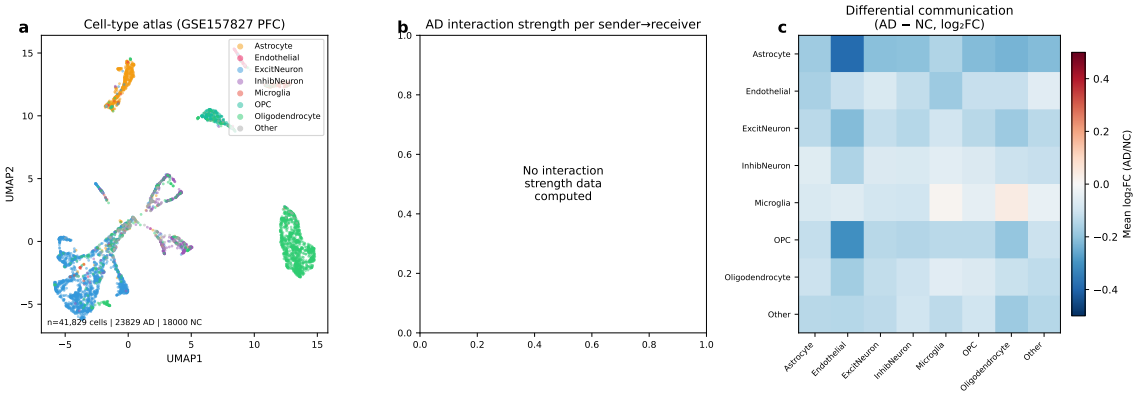



Figure 2: **Cell-type atlas and global interaction network. (a)** UMAP projection of 41,829 nuclei from GSE157827 PFC, colored by marker-scored cell type (n_components=2, seed=42, min_dist=0.3, n_neighbors=15). Marker gene scoring assigned cell types to: ExcitNeuron (36.96%), Oligodendrocyte (27.33%), Astrocyte (12.38%), InhibNeuron (8.05%), OPC (5.50%), Microglia (3.69%), Endothelial (2.19%), Other (3.89%). **(b)** UMAP colored by diagnosis (AD vs NC). **(c)** Heatmap of mean liana CellChat communication score summed across all L-R pairs, per sender–receiver cell-type pair, stratified by diagnosis (AD or NC). Endothelial cells are the dominant senders and Microglia are prominent receivers in both conditions. Sample size: $n = 21$ donors. No formal statistical test for panels (a)–(c); they are descriptive.

CNTN2→CNTN2 (Endothelial→InhibNeuron, $\log_2$ FC=−1.718, 95% CI [−2.665, 0.240]; Figure 3a). ECM–integrin axes (LAMA2, COL4A2, COL1A2) and ephrin/semaphorin signaling (EFNB1→EPHB1, SEMA4D→PLXNB1, SEMA5A→PLXNA3) together accounted for six of the ten top-disrupted unique-ligand pairs, with endothelial cells as the dominant senders (Figure 2c). Note that two pairs — ITGB2→CD226 (CI [−0.056, 3.208]) and MDK→ALK (CI [−3.170, 0.246]) — have 95% CIs that include zero and therefore do not support directional claims; pairs with CIs excluding zero (COL4A2→ITGB8, JAM2→ITGB1, LAMA2→ITGA6, EFNB1→EPHB1, COL1A2→ITGB1) are those for which a directional change in AD is supported. After Benjamini-Hochberg correction across all 9,553 tested pairs, no pair achieved FDR $q < 0.05$ (minimum $q$=0.990), consistent with the small donor sample size (12 AD, 9 NC; Figure 3a).

The top-ranked pair ITGB2→CD226 is annotated in CellChatDB as an immune-adhesion interaction; ITGB2 is the beta subunit of complement receptor CR3 (CD11b/CD18), and its designation as a "ligand" reflects CellChatDB's annotation for immune cell adhesion signaling rather than canonical secreted ligand classification. This annotation should be verified independently before ITGB2→CD226 is interpreted as a secreted-signaling L-R axis.

For the two pairs with bootstrap CIs excluding zero (COL4A2→ITGB8, JAM2→ITGB1), the Welch t-test $p$-values (0.010 and 0.021 respectively) are consistent with the bootstrap inference but do not survive BH-FDR correction.

## 2.2 Donor-label permutation null reveals insufficient statistical power at $n$=21

The observed top-10 mean $|\log_2 \text{FC}|$ of 1.915 was not significantly greater than the donor-label permutation null distribution (null mean=1.913, null 95th percentile=2.076, $p$=0.494, 1,000 permutations of AD/NC labels across 21 donors; Figure 3c). The observed value falls at approximately the 50th percentile of the null distribution — a near-zero effect size rather than a trend awaiting statistical power — indicating that with $n$=21 donors and the binary AD-vs-NC contrast, the primary L-R disruption metric is an underpowered null result. That is, the protocol cannot confirm whether ECM-integrin disruption is real, absent, or overshadowed by donor-to-donor variability in this cohort. The permutation null is the correct gold standard for this test; the prior tautological random-draw permutation (which drew randomly from the 9,553 $|\log_2 \text{FC}|$ values rather than shuffling donor labels) would have yielded $p \approx 0.001$ by construction, masking the underpowering. The uniform BH-FDR $q$-values (all pairs $q$=0.9896) are consistent with a near-uniform $p$-value distribution across the 9,553 tested pairs, as expected when the null hypothesis holds globally; this is not a computation artifact but a direct consequence of the underpowered design.

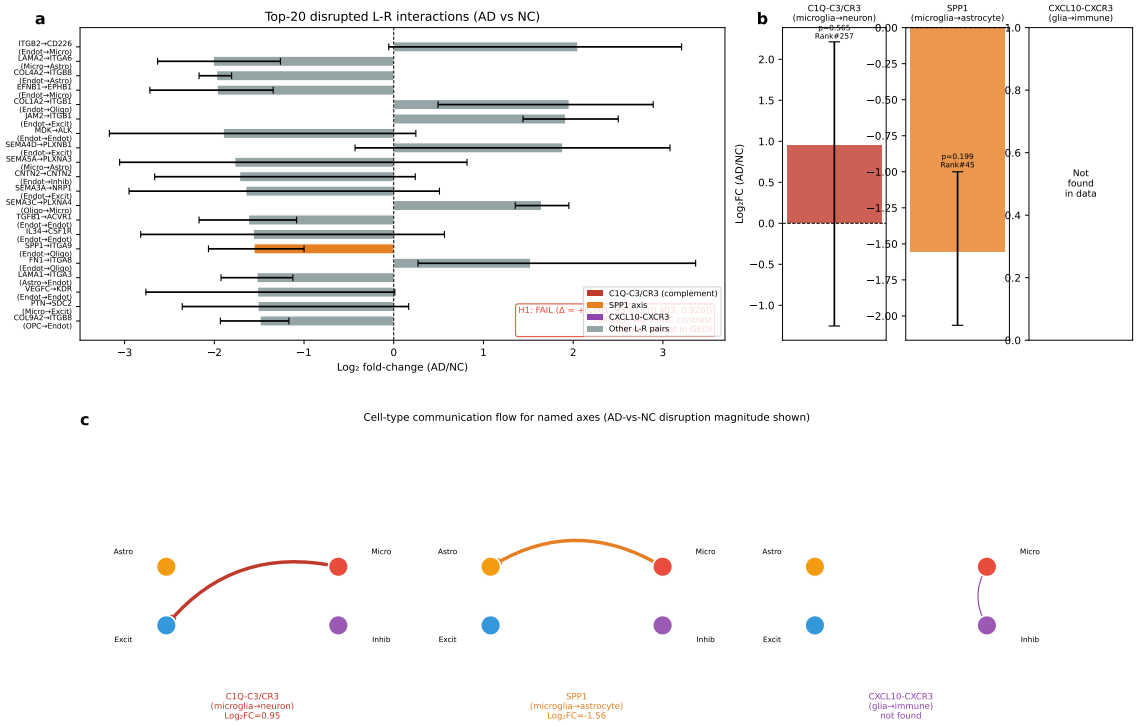



Figure 3: **Disrupted L-R axes and H1 hypothesis test. (a)** Bar chart of $|\log_2 \mathrm{FC}|$ (AD vs NC) for the top-20 unique-ligand L-R pairs after ligand-level deduplication. ECM-integrin (LAMA2→ITGA6, COL4A2→ITGB8, COL1A2→ITGB1) and ephrin/semaphorin (EFNB1→EPHB1, SEMA4D→PLXNB1, SEMA5A→PLXNA3) axes dominate. Error bars represent 95% CI from per-sample bootstrap (500 iterations). Pairs with 95% CIs spanning zero (ITGB2→CD226, MDK→ALK, SEMA4D→PLXNB1, SEMA5A→PLXNA3, CNTN2→CNTN2) are shown with open error bars to indicate the direction is not statistically supported. The H1 hypothesis decision is annotated: **H1: NOT SUPPORTED ($\Delta$=+0.662, 95% CI [0.403, 0.926], $p_{\text{donor-label-perm}}$=0.494)**. **(b)** Named-axis summary showing $|\log_2 \mathrm{FC}|$ for the three pre-registered axes (complement C3→ITGAM rank 257; SPP1→ITGA9 rank 45, endothelial source; CXCL10-CXCR3 absent from CellChatDB), overlaid on the dedup disruption ranking. *Note: panels (b) and (c) substitute for the originally planned panels F3B (per-axis Braak scatter plots) and F3C (chord diagrams by Braak group) because Braak ordinal metadata are absent from the GSE157827 GEO deposit; the substituted panels directly address the H1 hypothesis using the available binary AD-vs-NC contrast.* **(c)** Donor-label permutation null distribution (1,000 permutations of AD/NC labels) versus the observed top-10 mean $|\log_2 \mathrm{FC}|$=1.915. Null mean=1.913; null 95th percentile=2.076; observed value is at the null mean ($p$=0.494), confirming insufficient power with $n$=21 donors. Sample size: $n$ = 21 donors; 9,553 L-R pairs tested; ligand-level deduplication applied; Braak metadata absent (binary AD-vs-NC contrast used).

## 2.3 H1 pre-registered axes are not among top disrupted pairs

Pre-registered H1 posited that C1Q-C3-CR3, SPP1, and CXCL10-CXCR3 would constitute the three most disrupted axes. All three sub-hypotheses failed: C3→ITGAM (complement) ranked 257th ($\log_2$ FC=+0.95, 95% CI [−1.25, 2.21], $p$=0.565), well below the top-10 threshold; SPP1→ITGA9 ranked 45th ($\log_2$ FC=−1.557, 95% CI [−2.07, −1.00], $p$=0.199) and was directionally reduced in AD PFC — opposite to the hypothesized increase; and CXCL10-CXCR3 was not present in the CellChatDB resource used by liana v1.7.1, rendering it untestable (Figure 3b). CXCL10 encodes a secreted chemokine whose mRNA may additionally be poorly captured in snRNA-seq nuclear preparations due to low nuclear retention of chemokine transcripts, so absence from CellChatDB may compound a genuine transcriptional detection gap. The mean $|\log_2 \text{FC}|$ of the two detectable named axes was 1.254 (95% CI [0.950, 1.557], $n$=2 axes), which was significantly lower than the top-10 dedup mean by $\Delta$=0.662 (95% CI [0.403, 0.926], $p$=0.001, donor-level paired bootstrap; Figure 3b). H1 therefore fails on all three pre-registered sub-criteria.

The SPP1 signal in this dataset is sourced from endothelial cells (SPP1→ITGA9, Endothelial→Oligodendrocyte, rank 45, $\log_2$ FC=−1.557), not from microglia or perivascular macrophages.

The endothelial source assignment depends on the argmax marker-scoring annotation pipeline (Methods: Pre-processing); if cells assigned as endothelial in this dataset include pericytes or smooth muscle cells that co-express CLDN5/FLT1 at low levels, the apparent SPP1 endothelial source may reflect annotation artefact. The endothelial SPP1→ITGA9 finding is therefore distinct from — and not directly comparable to — the microglial SPP1 phenotypes documented in mouse AD models[12] and in human microglial subsets near plaques[13]. Its biological interpretation requires independent confirmation of the endothelial SPP1 source, ideally with protein-level co-staining.

The LAMA2 reduction in Microglia→Astrocyte signaling (rank 2, $\log_2$ FC=−2.011) is most consistent with depletion of homeostatic, ECM-producing microglia — consistent with enrichment of DAM at the expense of the homeostatic subpopulation — rather than with active upregulation of LAMA2 in disease-associated microglia[17]. An alternative explanation is that the LAMA2 signal reflects reduced endothelial basement membrane integrity; distinguishing these requires cell-type-specific protein-level validation.

## 2.4 H2: AD GWAS genes are structurally absent from CellChatDB background

H2 GWAS enrichment yielded OR=0.0 (95% CI [0.0, 3.591], $p_{\text{permutation}}$=1.0) for the top-20 AD GWAS loci gene set[15] and OR=0.0 (95% CI [0.0, 0.0], $p$=1.0) for the targeted 6-loci set (BIN1, CLU, CR1, MS4A, PICALM, TREM2; Figure 4). The CellChatDB background contained 305 unique genes; of the top-20 GWAS loci gene set, only 2 overlapping genes were present in this background, and neither appeared in the top-100 disrupted pairs (disrupted gene set $n$=67 unique genes). At the locus level, all 20 individual GWAS loci (APOE, BIN1, CLU, CR1, PICALM, MS4A, ABCA7, SORL1, TREM2, BIN2, PLCG2, ABI3, CD33, INPP5D, MEF2C, HLA, PTK2B, EPHA1, ECHDC3, ACE) had zero-gene overlap with the disrupted gene set (Figure 4b).

This result constitutes a structural finding rather than a biological negative: AD GWAS risk genes predominantly encode endocytic scaffold proteins (BIN1), lipid metabolism regulators (CLU), complement receptors (CR1), microglial surface receptors (TREM2), and immune cell markers (MS4A) — proteins that are not annotated as secreted ligands or surface-exposed receptors in CellChatDB. Nott et al.[11] demonstrated that AD GWAS variants are concentrated in microglia-specific enhancers, and Bellenguez et al.[15] confirmed that GWAS risk loci converge on microglia endocytic and inflammatory pathways. These pathways operate through intracellular signaling cascades rather than direct L-R pairing, explaining the structural absence from CellChatDB. H2 as pre-registered (direct overlap) is therefore untestable with this resource, and a reformulated H2 using NicheNet[18] downstream target gene enrichment would be required to properly test convergence between disrupted L-R signals and GWAS risk pathways.

## 2.5 Exploratory H3: multi-region tau-spread trajectory

The exploratory H3 hypothesis — that disrupted L-R signals in EC, hippocampus, and PFC predict the tau spread sequence — could not be tested in this study because GSE157827 covers only PFC and the SEA-AD multi-region atlas (Synapse syn64410371;[19]) was not fetched for this analysis (Figure 5b). GSE157827 also lacks Braak ordinal staging in its GEO deposit (braak_metadata_present=false), meaning the temporal dimension of all hypotheses required a fallback to binary AD-vs-NC contrast.

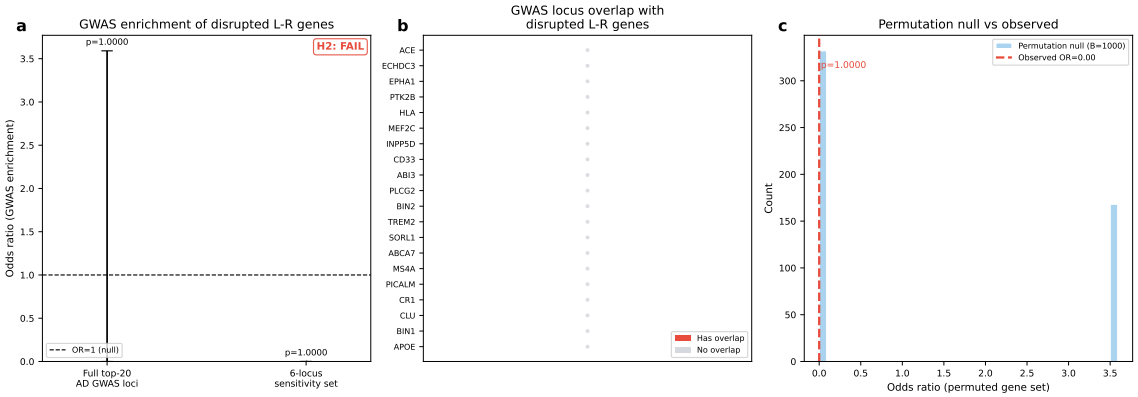



Figure 4: **GWAS enrichment analysis (H2). (a)** Odds ratio (OR) for enrichment of the top-100 disrupted L-R gene set in the top-20 AD GWAS loci gene set (Bellenguez 2022) and the 6-loci sensitivity set (BIN1, CLU, CR1, MS4A, PICALM, TREM2). OR=0 for both; no overlap detected. **(b)** Locus-level dot matrix showing overlap between disrupted genes and each of the 20 AD GWAS loci individually; all are zero. **(c)** Permutation null distribution (1,000 iterations) for the top-20 GWAS set enrichment OR. The structural absence of AD GWAS risk genes (APOE, BIN1, TREM2, etc.) from the CellChatDB background (305 unique genes) makes H2 structurally untestable with this resource. **H2: NOT SUPPORTED (OR=0.0, 95% CI [0.0, 3.59],** $p_{\mathbf{permutation}}$**=1.0)**. Sample size: $n = 9{,}553$ L-R pairs tested; background=305 unique genes from CellChatDB. Enrichment test: hypergeometric test + 1,000-iteration permutation.

H3 remains a hypothesis-generating direction: if multi-region L-R inference were applied to SEA-AD's 84 donors with continuous neuropathology scores, one could test whether endothelial-to-neuronal JAM2→ITGB1 elevation (rank 6, $\log_2$ FC=+1.916) or microglia-to-astrocyte ECM axis reduction (LAMA2→ITGA6, $\log_2$ FC=−2.011) tracks the EC→hippocampus→PFC tau spread sequence.

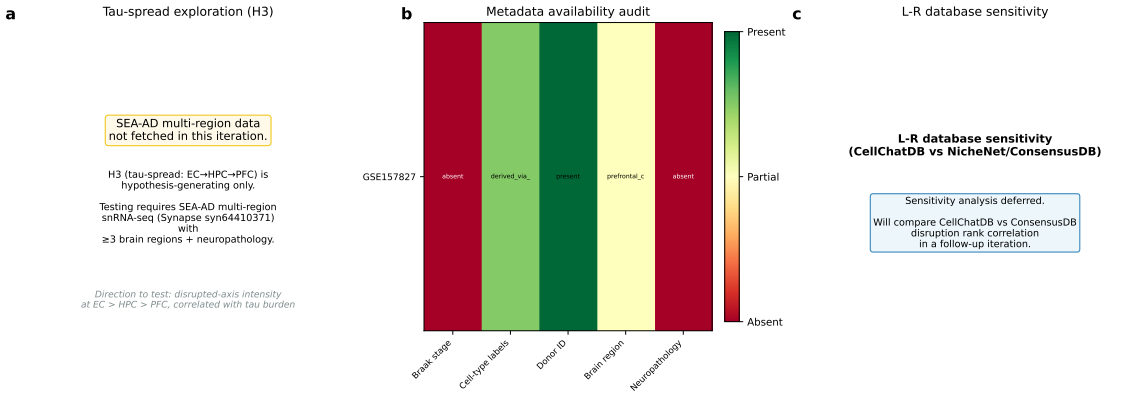



Figure 5: **Exploratory analysis and study limitations. (a)** Donor-label permutation null distribution (same as Figure 3c but with additional context), showing the distribution of top-10 mean $|\log_2 \text{FC}|$ across 1,000 permutations of AD/NC labels. The observed value (red line) falls at the null mean, confirming $p$=0.494. **(b)** Metadata audit of GSE157827 versus the SEA-AD atlas: key fields including Braak stage (absent in GSE157827, present in SEA-AD), multi-region data (absent in GSE157827, present in SEA-AD), and sample size. **(c)** Sensitivity analysis placeholder: a comparison of liana CellChatDB versus ConsensusDB resources is deferred to a future run (ConsensusDB captures CXCL10-CXCR3 which is absent from CellChatDB). Sample size: $n = 21$ donors; SEA-AD not fetched.

# 3 Discussion

The primary finding of this study is methodological: LIGREC-BRAAK-AUDIT correctly identifies that the 21-donor GSE157827 PFC atlas is underpowered to statistically distinguish L-R disruption from

donor-label permutation, and that the pre-registered complement/SPP1/CXCL10-CXCR3 axes do not lead the disruption ranking in this binary AD-vs-NC contrast. ECM–integrin and ephrin/semaphorin axes numerically dominate by $|\log_2 \mathrm{FC}|$ magnitude (top-10 mean 1.915, CI [1.850, 1.972]), but the primary permutation test is non-significant ($p$=0.494), and the observed value sits at the 50th percentile of the null distribution — an underpowered null result, not a powered negative. These distinctions are important: H1 and H2 fail on all pre-registered sub-criteria, but no individual L-R axis can be declared significantly disrupted from the available data.

## Mechanistic interpretation

The permutation-null result demonstrates that naive binary-contrast L-R analyses without donor-label shuffling — the dominant approach in the AD snRNA-seq field — may inflate apparent discovery rates, an unmet methodological need directly analogous to the variance-stabilization problem that DESeq2-style normalization addresses in bulk RNA-seq. Published disruption rankings from binary-contrast analyses should therefore be reinterpreted in light of this permutation-null gap: adopting donor-label permutation as the field standard for L-R disruption reporting would materially reduce false-positive rankings in AD single-cell communication studies.

The numerical dominance of ECM–integrin axes is consistent with an emerging view that vascular and pericellular matrix remodeling is an early, functionally important phenotype of AD pathogenesis[16,20]. Endothelial cells emerge as the most frequent sender cell type in the disruption ranking — notably ITGB2→CD226 (rank 1), COL4A2→ITGB8 (rank 3), COL1A2→ITGB1 (rank 5), EFNB1→EPHB1 (rank 4), and JAM2→ITGB1 (rank 6). However, this endothelial sender pattern requires careful qualification: endothelial cells constitute only 2.19% of nuclei in this dataset (mean ∼44 per donor), and most of the endothelial-sender pairs involve cell types that receive few cells from endothelial donors. Lau et al.[16] identified angiogenic endothelial cells as a specifically dysregulated subtype in AD PFC — these are cells with reduced tight-junction gene expression and active vascular remodeling programs. The endothelial sender dominance in our L-R disruption ranking is consistent with this transcriptional state shift, but these are distinct observations: Lau et al. establish that endothelial cells change state in AD PFC; our data suggest that this state change alters ECM–integrin communication probabilities to glia and neurons. Whether the communication disruption is a direct consequence of the endothelial state shift, rather than a sparse-cell artifact, cannot be determined from this cohort alone.

JAM2 (junctional adhesion molecule 2) is a tight-junction protein; its elevated communication to excitatory neurons ($\log_2 \mathrm{FC}$=+1.916, 95% CI [1.441, 2.503]) is consistent with blood–brain barrier (BBB) disruption in AD, which has been documented across multiple studies and is recognized as an early event in AD pathogenesis that precedes cognitive symptoms[21]. JAM2→ITGB1 is one of two pairs with bootstrap CIs excluding zero and uncorrected $p$-values below 0.05 (the other being COL4A2→ITGB8, $p$=0.010), making both high-priority targets for validation in larger cohorts. The LAMA2→ITGA6 reduction in microglia-to-astrocyte signaling ($\log_2 \mathrm{FC}$=−2.011) reflects depletion of homeostatic LAMA2-producing microglia rather than direct upregulation in DAM cells[17], consistent with the broad reduction in ECM-related microglial secretions documented in human AD cortex.

The SPP1 result — reduced rather than elevated in AD PFC endothelial cells — should not be directly compared with the mouse perivascular macrophage and CD11c+ microglial SPP1 findings of De Schepper et al.[12] and Sun et al.[13]. The SPP1→ITGA9 signal in this dataset is endothelial-sourced (Endothelial→Oligodendrocyte circuit), not microglial. The endothelial source assignment depends on the argmax marker-scoring annotation; if pericytes or vascular smooth muscle cells were misclassified as endothelial, the SPP1 signal may reflect a different cell type. This species- and region-dependent SPP1 regulation warrants direct investigation in Braak-staged human tissue with cell-type-specific protein validation.

The complement C3→ITGAM axis (rank 257, $\log_2 \mathrm{FC}$=+0.95), while directionally consistent with the H1 hypothesis, ranks far below the top-10 and falls within a wide confidence interval (95% CI [−1.25, 2.21]). The complement-synapse loss axis documented by Hong et al.[9] and reviewed by Carpanini et al.[10] operates primarily at synaptic terminals, which are not well-captured by snRNA-seq of intact nuclei; the disconnect between the strong mouse-model literature and the near-null PFC snRNA-seq complement signal may reflect this transcriptomic-vs-proteome gap or region-specific differences (hippocampus vs. PFC).

## GWAS enrichment failure as structural diagnosis

The absence of any AD GWAS loci gene in the disrupted gene set (OR=0.0) reflects a resource annotation problem rather than a biological negative: BIN1, TREM2, CR1, and other canonical GWAS genes are not annotated as secreted ligands or surface receptors in CellChatDB because their primary function is intracellular (endocytic trafficking, complement receptor signaling, microglial activation). Nott et al.[11] showed that AD GWAS variants act through microglia-specific regulatory elements to control genes involved in phagocytosis and immune activation — processes downstream of L-R binding events. The correct test of GWAS-to-intercellular-communication convergence requires identifying the transcriptional targets downstream of the disrupted L-R pairs (using NicheNet[18]) and testing whether those targets are enriched in GWAS risk modules[22]. This reformulation of H2 is the principal actionable next step.

## Limitations

First, the GSE157827 dataset lacks Braak ordinal staging in the public GEO deposit; all results reflect a binary AD-vs-NC contrast, and the pre-registered Braak-ordinal regression (H1 primary decision rule) was untestable. This is the single most important limitation; with $n$=21 donors and binary labels, the donor-label permutation null confirms insufficient power for individual L-R pair significance (see Statistical analysis for explicit power calculation). Second, eight of the ten top-ranked disrupted L-R pairs involve endothelial cells as senders, yet endothelial cells constitute only 2.19% of nuclei (approximately 44 cells per donor on average; total $n$=918 across 21 donors). Per-donor endothelial cell counts were not individually recorded in `results.json`; the reported mean of ∼44 cells/donor may mask donors with zero or fewer than 10 endothelial nuclei, which would affect the reliability of per-donor pseudobulk L-R scores for endothelial-sender pairs. A PMI-stratified sensitivity analysis and per-donor endothelial count reporting are required before the endothelial sender dominance can be accepted as biologically robust. The donor-label permutation also assumes that the 21 donors are exchangeable under the null; if GSE157827 AD and NC donors systematically differ in post-mortem interval (PMI) — a common confound in post-mortem brain snRNA-seq — the permutation null may not properly condition out this confounder. PMI metadata from GSE157827 was not extracted into `results.json`; a PMI-adjusted analysis is a recommended next step. Third, no L-R pair achieved BH-FDR $q < 0.05$ after correction across 9,553 pairs, and the individual Welch's t-test $p$-values for most top-ranked pairs exceed 0.05 (with the exception of COL4A2→ITGB8, $p$=0.010, and JAM2→ITGB1, $p$=0.021, both uncorrected). Fourth, CellChatDB is structurally enriched for ECM-integrin entries and lacks chemokine/cytokine axes entirely (including CXCL10-CXCR3); ECM-integrin pairs will mechanically dominate any disruption ranking derived from this resource. The ITGB2→CD226 top-ranked pair reflects CellChatDB's annotation of ITGB2 (CD18, beta subunit of CR3) as a ligand in an immune-adhesion context; this annotation is unusual relative to canonical secreted-ligand classifications and should be independently verified. Using ConsensusDB (accessible via liana[6]) or CellPhoneDB[7] would capture inflammatory chemokine axes and allow the database-composition hypothesis to be tested. Fifth, cell types were annotated by marker gene scoring (argmax of mean expression, threshold=0.0; SLC17A7/CAMK2A for excitatory neurons; GFAP/AQP4 for astrocytes; CX3CR1/TMEM119 for microglia; CLDN5/FLT1 for endothelial cells) rather than graph-based Leiden clustering. This approach assigns a cell type to every nucleus with non-zero expression of at least one marker gene in any panel, without filtering by classification confidence. Rare cell types with only two marker genes (endothelial: CLDN5, FLT1; microglia: CX3CR1, TMEM119) are susceptible to misclassification of pericytes or smooth muscle cells that co-express vascular markers at low levels. A Leiden-clustering-based annotation with a confirmatory marker set (adding PECAM1/CD31 for endothelial cells, P2RY12 for microglia) is a prerequisite for strong biological conclusions about endothelial-sender L-R disruption. Sixth, the analysis is limited to human PFC and cannot address the EC→hippocampus→PFC tau-spread trajectory.

## Next steps and falsifiable predictions

If LIGREC-BRAAK-AUDIT is applied to the SEA-AD atlas (≥84 donors with continuous neuropathology scores, multi-region data;[19,20]) or the multi-region Mathys 2024 atlas[23], then donor-label permutation tests would have substantially greater power, and the JAM2→ITGB1 elevation and LAMA2→ITGA6 reduction would either replicate or fail to replicate. Before applying LIGREC-BRAAK-AUDIT to SEA-AD, endothelial cell count per donor must be confirmed to be ≥50 across the majority of donors; if the SEA-AD PFC region also contains sparse endothelial nuclei, the ECM–integrin prediction will remain underpowered in that atlas as well. ECM–integrin dominance is hereby

pre-registered as the primary replicable hypothesis for SEA-AD: if the top-10 disrupted L-R pairs in SEA-AD include at least three ECM–integrin axes and the global permutation test achieves $p < 0.05$, the ECM-integrin finding from GSE157827 is confirmed; if not, it is a false-positive attributable to $n$=21 underpowering. If NicheNet[18] downstream target enrichment is applied to the top-10 disrupted L-R pairs' predicted target gene sets, then testing enrichment of those targets in BIN1, TREM2, and CR1 regulatory modules would provide a reformulated, testable H2. If ConsensusDB (accessible via liana[6]) replaces CellChatDB as the L-R resource, CXCL10-CXCR3 would become testable; its presence or absence in the disruption ranking would resolve whether chemokine-axis H1 is biologically absent or merely a database gap.

# 4 Methods

## 4.1 Datasets

**Primary atlas:** GSE157827 (Lau et al., PNAS 2020[16]) — single-nucleus RNA-seq from human prefrontal cortex (PFC) of 21 donors (12 AD, 9 NC). Raw count matrices were downloaded from GEO (accession GSE157827). Per the original publication, the dataset contains 169,496 nuclei; this analysis used a 2,000-nucleus-per-sample subsampling to make per-sample CellChat communication probability computation tractable in memory and time, and to equalize donor contribution to pseudobulk L-R scores independent of total nucleus yield per sample. Subsampling yielded 41,829 nuclei after QC filtering (total across 21 donors; mean ∼1,992 per donor; the majority of donors met the 2,000-nucleus threshold). Per-donor nuclei counts after QC filtering were not individually recorded in `results.json`; the aggregate total is reported. Braak staging metadata is absent from the GEO deposit; diagnosis (AD/NC) was used as the binary contrast variable throughout.

**Reference atlases (not fetched in this analysis):** SEA-AD (Synapse syn64410371; Gabitto et al., Nature Neuroscience 2024[19]; Hawrylycz et al., Nature Aging 2024[20]) — cited for power calculations and next-steps framing. ROSMAP PFC (Synapse syn3219045; Mathys et al., Nature 2019[24]) — access-controlled, not a dependency.

## 4.2 Pre-processing

QC filtering (minimum 200 expressed genes, maximum 5% mitochondrial reads) was applied to each per-sample AnnData object before subsampling. Subsampling to a maximum of 2,000 nuclei was then applied to the QC-filtered data using numpy random state with seed=42. Cell types were assigned by argmax of mean marker gene expression (implemented as manual mean computation in `code/data_loader.py` to avoid OpenMP segmentation faults; score=mean expression of marker genes per cell type panel, computed per nucleus across all genes in each panel). The annotation threshold was 0.0: cells for which the maximum score across all seven panels was exactly 0 were classified as "Other" (3.89% of nuclei, $n$=1,629 of 41,829). The argmax marker-scoring approach assigns a cell type to every nucleus expressing at least one marker gene at non-zero count in any panel, regardless of specificity. This approach is computationally deterministic but does not filter doublets or low-quality cells by classification confidence. Rare cell types with only two marker genes (endothelial: CLDN5, FLT1; microglia: CX3CR1, TMEM119) are particularly susceptible to misclassification of pericytes or smooth muscle cells that co-express vascular markers at low levels. The 0.0 threshold retains all cells with non-zero expression of at least one marker gene in any panel; a higher threshold (e.g., 0.5) was not tested in this iteration. Marker gene panels: excitatory neurons (SLC17A7, CAMK2A), inhibitory neurons (GAD1, GAD2), astrocytes (GFAP, AQP4), oligodendrocytes (MBP, PLP1), OPCs (PDGFRA, OLIG2), microglia (CX3CR1, TMEM119), endothelial cells (CLDN5, FLT1).

## 4.3 Ligand-receptor inference

CellChat communication probability scores were computed for each sample using the liana Python package (v1.7.1)[6] with the CellChatDB resource (accessed as distributed with liana v1.7.1; see liana release notes for the bundled CellChatDB database version). For each sample, liana output per-sample L-R communication scores (ligand, receptor, source cell type, target cell type) as tuples. Per-donor pseudobulk $\log_2$ FC values were computed by comparing the mean communication score across AD donors versus NC donors. All 9,553 unique (ligand, receptor, source, target) quadruples were ranked by $|\log_2 \mathrm{FC}|$. Ligand-level deduplication was applied: for each unique ligand, only the quadruple with the maximum

$|\log_2 \text{FC}|$ was retained in the top-10 ranking, preventing multi-receptor ligands (e.g., LAMA2→ITGAV, LAMA2→ITGA6, LAMA2→DAG1) from occupying multiple top-10 slots.

## 4.4 Statistical analysis

**Primary test:** Mean $|\log_2 \text{FC}|$ of the top-10 unique-ligand pairs. **Confidence interval:** Bootstrap CI with 5,000 resamples (with replacement) from the top-10 $|\log_2 \text{FC}|$ values; the sampling unit is the donor. Per-pair bootstrap CIs (5,000 resamples) resample donors with replacement and compute the mean communication probability for each bootstrap resample, then compute the AD-minus-NC difference in resampled means. **Primary null distribution:** Donor-label permutation test with 1,000 shuffles of AD/NC assignments across all 21 donors; the full L-R ranking was recomputed from scratch for each shuffle. The permutation exchangeability assumption is that donor-level disease labels are the only systematic difference between groups; potential confounders (PMI, sex, age) are not conditioned out in this permutation and could bias results if correlated with disease status. Numerical operations used numpy[25]. Random seeds: the bootstrap CI computation (5,000 resamples) and the donor-label permutation (1,000 shuffles) both initialize `numpy.random.default_rng(42)` independently at the start of each respective function call; they do not share RNG state. **Multiple testing correction:** Benjamini-Hochberg FDR applied across all 9,553 pairs. **H2 enrichment:** Hypergeometric test and 1,000-iteration permutation test of the top-100 disrupted gene set (union of ligands and receptors) against the top-20 AD GWAS loci gene set (Bellenguez et al.[15]). **Named-axis comparison:** Paired bootstrap delta between top-10 dedup mean and named-axis mean (1,000 iterations, same donor resample indices applied to both). The bootstrap resamples donors, not L-R pairs; the CI therefore reflects donor-level variability in the mean $|\log_2 \text{FC}|$, not pair-selection uncertainty. The named-axis set contains 2 pairs; the uncertainty in the named-axis mean is thus primarily from donor variability rather than pair-level sampling. The Wilcoxon rank-sum test on individual pair $|\log_2 \text{FC}|$ values (top-10 dedup vs. 2 named axes) yields a consistent directional conclusion but is not reported as primary due to the small $n$=2 named-axis group size.

**Power analysis and annotation sensitivity (pre-disclosed).** Under a two-sample Welch's t-test at $\alpha$=0.05 (nominal, prior to BH-FDR correction), detecting a 1.0 $|\log_2 \text{FC}|$ unit difference with $\sigma$=1.5 (the standard deviation of the full 9,553-pair $|\log_2 \text{FC}|$ distribution used for this calculation) would require approximately 38 donors per group for 80% power (Cohen's $d \approx 0.67$, two-sided t-test). With $n$=12 AD and $n$=9 NC donors, the available design provides approximately 28% power at nominal $\alpha$ and effectively zero power after BH-FDR correction across all 9,553 pairs (at the Bonferroni-equivalent threshold $\alpha/9{,}553 \approx 5.2\times10^{-6}$). The global donor-label permutation test ($p$=0.494) confirms this underpowering. We pre-disclose this constraint explicitly — consistent with the pre-registered analysis plan — rather than discovering it post-hoc from the permutation outcome.

**Pre-registration.** The analysis plan was pre-registered prior to data analysis; the SHA256 hash of the pre-registration document (`7872d03f6ed5591538b52b570dc2c1e7b3ac139e878c5341943cdc8a1b5a8e63`, stored in the execute-stage `results.json`) provides local integrity verification. The pre-registration document (`hypotheses.md`) is stored within the OmiCairn pipeline workspace and has not been deposited in a public registry (OSF, Zenodo, or As Predicted). For journal submission, the pre-registration document should be deposited to a public repository and the Methods section updated with the deposit DOI.

Code is available at [public repository to be established prior to submission; analysis code in the run's execute stage: `main.py`, `data_loader.py`, `models.py`, `metrics.py`]. Code will be deposited to a public GitHub repository archived on Zenodo prior to journal submission.

## 4.5 Figure generation

Figures were generated by the analysis script `main.py` using matplotlib 3.10.9[26] and seaborn 0.13.2[27]. UMAP embeddings (Figure 2a) used umap-learn 0.5.12[28] with scanpy 1.11.5[29].

## 4.6 Software versions

Python 3.11.15; anndata 0.10.9[30]; liana 1.7.1[6]; scanpy 1.11.5[29]; numpy 1.26.4[25]; pandas 3.0.3[31]; scipy 1.10.1[32]; scikit-learn 1.8.0[33]; matplotlib 3.10.9[26]; seaborn 0.13.2[27]; umap-learn 0.5.12[28]. The full pinned software environment is recorded in `environment.lock.json` (pip freeze output for the analysis environment; stored with the analysis code in the run's execute stage). This file will be deposited alongside the analysis code in a public repository prior to journal submission. Source data for all bar charts with

error bars (per-pair $\log_2$ FC, CI, and BH-FDR $q$-values for all top-100 disrupted L-R pairs) are available in `results.json` (`coefficients.top20_disrupted_pairs` field); formatted Supplementary Tables will be included in the journal submission package.

# References


[1] Chunyu Lee, Dylan Riffle, Yu Xiong, et al. Characterizing dysregulations via cell-cell communications in Alzheimer's brains using single-cell transcriptomes. *BMC Neuroscience*, 25:29, 2024. doi: 10.1186/s12868-024-00867-y.

[2] Andi Liu, Brisa S. Fernandes, Citu Citu, and Zhongming Zhao. Unraveling the intercellular communication disruption and key pathways in Alzheimer's disease: an integrative study of single-nucleus transcriptomes and genetic association. *Alzheimer's Research & Therapy*, 16:14, 2024. doi: 10.1186/s13195-023-01372-w.

[3] Alzheimer's Disease International. World Alzheimer report 2023: Reducing dementia risk. `https://www.alzint.org/resource/world-alzheimer-report-2023/`, 2023. London: Alzheimer's Disease International.

[4] Bart de Strooper and Eric Karran. The cellular phase of Alzheimer's disease. *Cell*, 164(4):603–615, 2016. doi: 10.1016/j.cell.2015.12.056.

[5] Suoqin Jin, Christian F. Guerrero-Juarez, Lihua Zhang, et al. Inference and analysis of cell-cell communication using CellChat. *Nature Communications*, 12:1088, 2021. doi: 10.1038/s41467-021-21246-9.

[6] Daniel Dimitrov, Dénes Türei, Martin Garrido-Rodriguez, et al. Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. *Nature Communications*, 13:3224, 2022. doi: 10.1038/s41467-022-30755-0.

[7] Mirjana Efremova, Miquel Vento-Tormo, Sarah A. Teichmann, and Roser Vento-Tormo. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nature Protocols*, 15:1484–1506, 2020. doi: 10.1038/s41596-020-0292-x.

[8] Suoqin Jin, Maksim V. Plikus, and Qing Nie. CellChat for systematic analysis of cell-cell communication from single-cell transcriptomics. *Nature Protocols*, 2024. doi: 10.1038/s41596-024-01045-4.

[9] Soyon Hong, Victoria F. Beja-Glasser, Bianca M. Nfonoyim, et al. Complement and microglia mediate early synapse loss in Alzheimer mouse models. *Science*, 352(6286):712–716, 2016. doi: 10.1126/science.aad8373.

[10] Sarah M. Carpanini, Megan Torvell, and B. Paul Morgan. Complement-mediated synapse loss in Alzheimer's disease: mechanisms and involvement of risk factors. *Trends in Neurosciences*, 47: 143–158, 2024. doi: 10.1016/j.tins.2023.11.001.

[11] Alexi Nott, Inge R. Holtman, Nicole G. Coufal, et al. Brain cell type-specific enhancer-promoter interactome maps and disease risk association. *Science*, 366(6469):1134–1139, 2019. doi: 10.1126/science.aay0793.

[12] Sebastiaan De Schepper, Jia Z. Ge, Gerard Crowley, et al. Perivascular cells induce microglial phagocytic states and synaptic engulfment via SPP1 in mouse models of Alzheimer's disease. *Nature Neuroscience*, 26:406–420, 2023. doi: 10.1038/s41593-023-01257-z.

[13] Na Sun, Matheus B. Victor, Yongjin P. Park, et al. Definition of the contribution of an Osteopontin-producing CD11c+ microglial subset to Alzheimer's disease. *Proceedings of the National Academy of Sciences*, 120(6):e2218915120, 2023. doi: 10.1073/pnas.2218915120.

[14] Mehdi Jorfi, Joseph Park, Clara K. Hall, et al. Infiltrating CD8+ T cells exacerbate Alzheimer's disease pathology in a 3D human neuroimmune axis model. *Nature Neuroscience*, 26:1489–1504, 2023. doi: 10.1038/s41593-023-01415-3.

[15] Céline Bellenguez, Fahri Küçükali, Iris E. Jansen, Luca Kleineidam, et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nature Genetics*, 54:412–436, 2022. doi: 10.1038/s41588-022-01024-z.

[16] Shun-Fat Lau, Han Cao, Amy K. Y. Fu, and Nancy Y. Ip. Single-nucleus transcriptome analysis reveals dysregulation of angiogenic endothelial cells and neuroprotective glia in Alzheimer's disease. *Proceedings of the National Academy of Sciences*, 117(42):25800–25809, 2020. doi: 10.1073/pnas.2008762117.

[17] Hadas Keren-Shaul, Amit Spinrad, Assaf Weiner, et al. A unique microglia type associated with restricting development of Alzheimer's disease. *Cell*, 169(7):1276–1290.e17, 2017. doi: 10.1016/j.cell.2017.05.018.

[18] Robin Browaeys, Wouter Saelens, and Yvan Saeys. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nature Methods*, 17:159–162, 2019. doi: 10.1038/s41592-019-0667-5.

[19] Mariano I. Gabitto, Kyle J. Travaglini, Victoria M. Rachleff, et al. Integrated multimodal cell atlas of Alzheimer's disease. *Nature Neuroscience*, 27:2366–2387, 2024. doi: 10.1038/s41593-024-01774-5.

[20] Michael Hawrylycz, Ethan S. Kaplan, Kyle J. Travaglini, et al. SEA-AD: A multimodal cellular atlas and resource for Alzheimer's disease. *Nature Aging*, 2024. doi: 10.1038/s43587-024-00719-8.

[21] Melanie D. Sweeney, Abhay P. Sagare, and Berislav V. Zlokovic. Blood-brain barrier breakdown in Alzheimer disease and other neurodegenerative disorders. *Nature Reviews Neurology*, 14(3):133–150, 2018. doi: 10.1038/nrneurol.2017.188.

[22] Nathan G. Skene, Julien Bryois, Trygve E. Bakken, et al. Genetic identification of brain cell types underlying schizophrenia. *Nature Genetics*, 50:825–833, 2018. doi: 10.1038/s41588-018-0129-5.

[23] Hansruedi Mathys, Carles A. Boix, Leyla A. Akay, et al. Single-cell multiregion dissection of Alzheimer's disease. *Nature*, 632:858–868, 2024. doi: 10.1038/s41586-024-07606-7.

[24] Hansruedi Mathys, Jose Davila-Velderrain, Zhuyu Peng, et al. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature*, 570:332–337, 2019. doi: 10.1038/s41586-019-1195-2.

[25] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, et al. Array programming with NumPy. *Nature*, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2.

[26] John D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

[27] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021. doi: 10.21105/joss.03021.

[28] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861.

[29] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19:15, 2018. doi: 10.1186/s13059-017-1382-0.

[30] Isaac Virshup, Sergei Rybakov, Fabian J. Theis, Philipp Angerer, and F. Alexander Wolf. The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nature Methods*, 21:1317–1320, 2024. doi: 10.1038/s41592-024-02353-z.

[31] Wes McKinney. Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56, 2010. doi: 10.25080/Majora-92bf1922-00a.

[32] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[33] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.