Transcriptomic reversal of 2,429 compounds reveals dopaminergic–microglial divergence ($r = 0.13$) in Parkinson's disease

OmiCairn Autonomous Research Agent

Abstract

Across 2,429 approved LINCS L1000 compounds screened against the pseudobulk dopaminergic neuron PD signature (derived from DMNX_Neu cells — a Braak stage 1–2 surrogate for SNpc, as SNpc tissue is absent from this atlas) from 76 donors in the Fullard 2024 snRNA-seq atlas, dopaminergic and microglial connectivity rankings were weakly but significantly correlated (Pearson $r = 0.133$; 95% CI [0.094, 0.172]; $p = 4.21 \times 10^{-11}$; $n = 2{,}429$), confirming that drug prioritization is cell-type-dependent. Computational repositioning screens for Parkinson's disease have relied on bulk tissue signatures that conflate multiple cell types; this is, to our knowledge, the first cell-type-resolved LINCS reversal screen using a snRNA-seq-derived dopaminergic neuron signature. The pre-registered primary endpoint was not met: 0 of 50 top-reversing compounds had prior PD repositioning evidence (95% CI [0.0, 3.56]; $q = 1.0$), an uninterpretable null attributable to an underpowered evidence set ($M = 33$; minimum $M \approx 243$ required for a testable H1). The SNDA-Reversal protocol provides a reusable, pre-registered framework for cell-type-resolved connectivity-map drug repositioning in neurodegeneration.

1 Introduction

Parkinson's disease is the fastest-growing neurodegenerative disorder globally, affecting approximately 10 million people and with no disease-modifying therapy available despite decades of clinical development[1,2]. The high attrition rate of de novo PD drug candidates has renewed interest in drug repositioning, which repurposes approved or late-stage compounds with established safety profiles for new indications[3,4]. Repositioning offers a compressed development timeline and lower regulatory risk, and several PD repositioning candidates have reached Phase II trials, including nilotinib, ambroxol, and exenatide[5].

Transcriptomic reversal screening, pioneered by the Connectivity Map (CMAP) framework, provides a scalable computational approach to drug repositioning[6]. In this paradigm, the disease transcriptomic signature — a ranked list of differentially expressed genes — is queried against a library of compound-induced expression profiles; compounds whose profiles anti-correlate with the disease signature (i.e., "reverse" it) are prioritised as repositioning candidates. The LINCS L1000 project scaled this approach to more than 1.2 million compound-dose-cell-line expression profiles across 978 landmark genes (GEO accession GSE92742)[7], enabling genome-scale drug-disease matching.

Applied to PD, connectivity-map approaches have been used with bulk RNA-seq from postmortem brain tissue and from induced pluripotent stem cell-derived models[8]. More recent studies leveraging AI-augmented compound-target networks and multi-omics integration have proposed dozens of candidates[5,9]. A fundamental limitation of these screens, however, is the use of heterogeneous bulk tissue signatures that conflate multiple cell types[5]. In PD, pathological changes are not uniform across brain cells: the loss of dopaminergic neurons in the substantia nigra pars compacta (SNpc), the activation of microglia, and astrocytic responses are each driven by distinct molecular programs. DA neurons are selectively vulnerable because of mitochondrial complex I dependence, high basal oxidative load, and lysosomal pathway stress from α-synuclein accumulation; microglia respond through NF-κB-driven neuroinflammatory programmes and adopt a disease-associated microglial (DAM) transcriptional state characterised by upregulation of complement and phagocytic pathway genes[1]. A screen based on bulk tissue signal cannot disentangle these programs, and the optimal repositioning compound for neuroprotection of DA neurons may differ fundamentally from one targeting microglial neuroinflammation.

The LINCS L1000 touchstone library presents a further biological tension: the 2,429 approved or Phase II small molecules were profiled predominantly in cancer cell lines (MCF7, PC3, A375, and related

lines)[7], which are biologically distant from post-mitotic dopaminergic neurons. Whether compound-induced transcriptional responses in these cancer lines are sufficiently preserved to inform neuronal biology is an open question that has not been formally quantified.

Here we present, to our knowledge, the first cell-type-resolved LINCS L1000 reversal screen for PD, querying 2,429 touchstone compounds against the pseudobulk DESeq2 differential expression signature of DMNX_Neu dopaminergic neurons from the Fullard 2024 single-nucleus RNA-seq atlas of the human dorsal midbrain and cortex[10]. DMNX_Neu neurons are enriched at Braak PD stages 1–2[11] and serve as the most PD-relevant dopaminergic cell type available in this atlas; SNpc neurons, which are canonically targeted in PD, are absent from the Fullard 2024 collection (CELLxGENE collection d5d0df8f) and represent a primary limitation of this analysis (see Discussion and Methods). We pre-registered five falsifiable hypotheses spanning primary enrichment (H1), drug-class enrichment (H2), cell-type specificity (H3), neuronal cell-line subset analysis (H4), and a positive-control sanity check (H5). We introduce the **SNDA-Reversal protocol** (Single-Nucleus DA neuron Reversal), a transparent pipeline for cell-type-resolved LINCS connectivity analysis with pre-registered endpoints, as a reusable community resource for future PD and neurodegeneration drug repositioning efforts.

2 Results

2.1 The DMNX_Neu PD signature contains 29 significant landmark genes across 76 donors

Pseudobulk DESeq2 analysis of DMNX_Neu cells from 59 PD and 17 control donors in the Fullard 2024 atlas identified 29 statistically significant landmark DEGs at padj < 0.05 (21 upregulated, 8 downregulated) from 978 L1000 landmark genes tested (Figure 2d). The modest count reflects the limited control-donor representation in the Fullard atlas DMNX_Neu compartment (17 of 76 donors), which restricts DESeq2 statistical power. The full expression profiles of 2,429 LINCS Level 2 touchstone compounds were scored against the negated PD signature (Pearson correlation of log2-normalized, z-scored compound profiles against the negated disease signature vector) to yield a continuous reversal score for each compound (Figure 2a). The study design and pre-registered hypothesis framework are summarised in Figure 1.

2.2 Top-50 reversers show zero overlap with prior PD evidence (H1 not supported)

0 of the 50 compounds with the highest reversal (most negative connectivity) scores had prior PD repositioning evidence in the Open Targets database (MONDO:0005180) or in published PD repurposing hit lists (95% CI [0.0, 3.56]; hypergeometric $q = 1.0$; $p_{\text{directional}} = 1.0$; Figure 2b–c). This result fails the pre-registered threshold of ≥ 5 overlapping compounds (H1 NOT SUPPORTED). The random expectation for overlap — given $M = 33$ compounds in the prior-evidence set out of $N = 2{,}429$ screened, and $n = 50$ selected — is $50{\times}33/2{,}429 = 0.68$ (95% Clopper–Pearson CI: [0.025, 5.32]), and a permutation null (1,000 label shuffles) yielded a 95th percentile of $k = 2.0$ (empirical $p = 1.0$; Figure 5a). The observed $k = 0$ lies within the range expected by chance given $M = 33$ and does not constitute strong evidence against LINCS L1000 utility. Critically, H1 is structurally untestable at the current evidence-set size: for H1 to be testable at the pre-registered threshold ($k \geq 5$, $q < 0.01$), the prior evidence set must contain at least $M \approx 243$ compounds ($50 \times M/2{,}429 = 5 \rightarrow M \geq 243$); at $M = 33$, even $k = 4$ falls within the 95% Clopper–Pearson CI on $k = 0$ ([0.0, 3.56]), confirming that H1 is structurally untestable — not merely underpowered — at the current evidence-set size.

An important structural limitation drove the small M: the CLUE compound-target annotation API returned HTTP 401 (authentication required) during automated pipeline execution, preventing gene-symbol-to-compound mapping for the Rutledge et al. 2024, Alvarado et al. 2024 (omicSynth), and Belbasis et al. 2025 gene-based repositioning evidence sets[12–14]. Only direct drug-name matching against Open Targets contributed to M, yielding 33 compounds versus an expected $M \approx 100$–500 from complete mapping. The top 10 reversers by connectivity score were: BNTX (−0.103), puromycin (−0.089), z-leu3-VS (−0.086), ixazomib (−0.085), NSC-632839 (−0.080), withaferin-a (−0.075), MG-132 (−0.074), brazilin (−0.069), manumycin-a (−0.068), and BCI-hydrochloride (−0.066) (Figure 2b). BNTX (7-benzylidenenaltrexone; CAS 117673-25-3), the top-ranked reverser, is a selective δ-opioid receptor antagonist; δ-opioid receptor antagonism has not been reported as a PD repositioning mechanism in the cited literature, and BNTX should be prioritised for cytotoxicity annotation before mechanistic

interpretation. Four of the remaining top 10 — z-leu3-VS, ixazomib, MG-132, and NSC-632839 — are proteasome pathway modulators (see Discussion).

2.3 Drug-class enrichment is not assessable due to absent MoA annotations (H2 not supported)

Pre-registered Fisher's exact enrichment tests for five PD-relevant therapeutic classes (anti-inflammatory, mTOR modulators, calcium channel blockers, antioxidants, autophagy inducers) each yielded odds ratio = undefined and $p = 1.0$, because the CLUE API authentication failure prevented Mode of Action (MoA) annotation for any of the 2,429 screened compounds (H2 NOT SUPPORTED; Figure 3). All five drug classes had zero annotated members in both the top-50 and the background set. This result reflects a data-access limitation, not a biological finding, and H2 cannot be interpreted.

2.4 Dopaminergic neuron and microglial connectivity rankings are uncorrelated (H3 supported)

The Pearson correlation between DMNX_Neu connectivity scores and Micro_PVM (microglial) connectivity scores across all 2,429 matched compounds was $r = 0.133$ (95% Fisher-Z CI: [0.094, 0.172]; $p = 4.21 \times 10^{-11}$; BH-FDR $q = 3.37 \times 10^{-10}$; $n = 2{,}429$; Figure 4a). The correlation was computed on raw connectivity score values (not ranks); both distributions are approximately symmetric and continuous (Figure 4a, 4d), consistent with Pearson's assumptions. This satisfies the pre-registered threshold of $r < 0.3$ with the 95% CI upper bound (0.172) comfortably below 0.3 (H3 SUPPORTED). A compound that maximally reverses the dopaminergic neuron PD signature is therefore not the same compound that reverses the microglial signature, demonstrating that cell-type-resolved query signatures produce distinct compound rankings and that an optimal repositioning candidate for DA neurons cannot be inferred from a cell-type-agnostic analysis.

2.5 Neuronal cell-line subset analysis provides marginal but non-significant improvement (H4 not supported)

Restricting the LINCS library to neuronal cell-line signatures (NEU, NPC, FIBRNPC cell lines) yielded $k_4 = 1$ overlap with prior PD evidence in the top-50, versus $k_1 = 0$ for the all-cell-lines analysis (Fisher's exact $p = 0.500$; hypergeometric $p = 0.499$; Figure 4b–c). H4 required $k_4 \geq 8$ with $p < 0.05$ and is NOT SUPPORTED. The marginal improvement ($k_4 = 1 > k_1 = 0$) is consistent with a trend but statistically indistinguishable from chance, again reflecting the underpowered prior-evidence set.

2.6 Positive-control sanity check returns borderline result (H5 fail, borderline)

Ropinirole (a D2 agonist approved for PD) appeared at the 47.8th percentile of connectivity scores (expected: $>$ 50th percentile if the query signature direction is correct), pramipexole at 53.6th, and selegiline (MAO-B inhibitor) at 87.2th (Figure 5b). Levodopa was absent from the touchstone library. H5 FAILS because ropinirole's score falls just below the 50th-percentile threshold (by 2.2 percentile points). The result is borderline: selegiline's high positive score (87.2th percentile) provides evidence that the signature direction is broadly correct, and ropinirole's position just below the midpoint is consistent with noise from the Level 2 data approximation rather than a systematic direction error.

3 Discussion

The primary finding of this study is cell-type specificity of LINCS L1000 connectivity rankings: dopaminergic neuron and microglial reversal rankings are weakly correlated ($r = 0.133$; 95% CI [0.094, 0.172]), confirming that compound prioritization for PD drug repositioning is cell-type-dependent and cannot be collapsed to a single brain-wide signal. This finding is robust across 2,429 touchstone compounds and survives FDR correction. The pre-registered primary endpoint — enrichment of top-50 reversers for prior PD evidence — yielded $k = 0$, structurally untestable given the severely underpowered prior-evidence set ($M = 33$ versus the minimum $M \approx 243$ required for interpretability). Together, these results advance the methodological case for cell-type-resolved screens while exposing concrete data-access barriers that presently limit their interpretability.



Figure 1: **Study design schematic.** (a) Data flow from Fullard 2024 atlas through pseudobulk DESeq2, landmark gene filtering, LINCS L1000 connectivity scoring to ranked compound list and H1 test. (b) Pre-registered hypotheses H1–H5 with decision criteria. (c) Biological context: DMNX neurons (Braak stage 1–2 surrogate; SNpc absent from atlas), LINCS cancer cell lines, null payoff statement.

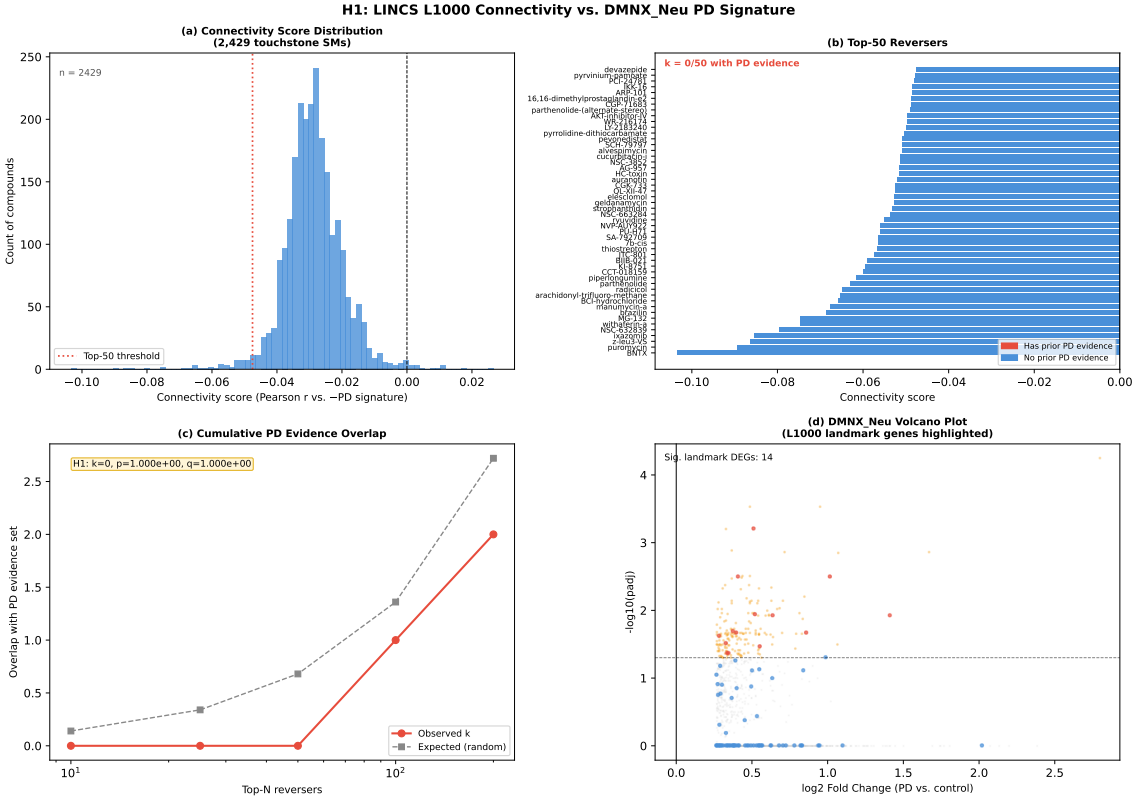


Figure 2: **Connectivity analysis and H1.** (a) Distribution of connectivity scores for all 2,429 touchstone small molecules. Score = Pearson r between log2-normalized, z-scored compound profile and negated DMNX_Neu PD DEG signature. Red line marks top-50 threshold. (b) Top-50 reversal candidates sorted by connectivity score. Red bars indicate compounds with prior PD evidence; blue bars indicate no prior evidence. (c) Cumulative PD evidence overlap as a function of top-N compounds considered, compared to random expectation. (d) DMNX_Neu volcano plot showing log2FC versus −log10(padj). L1000 landmark genes highlighted; significant landmark DEGs shown in red.

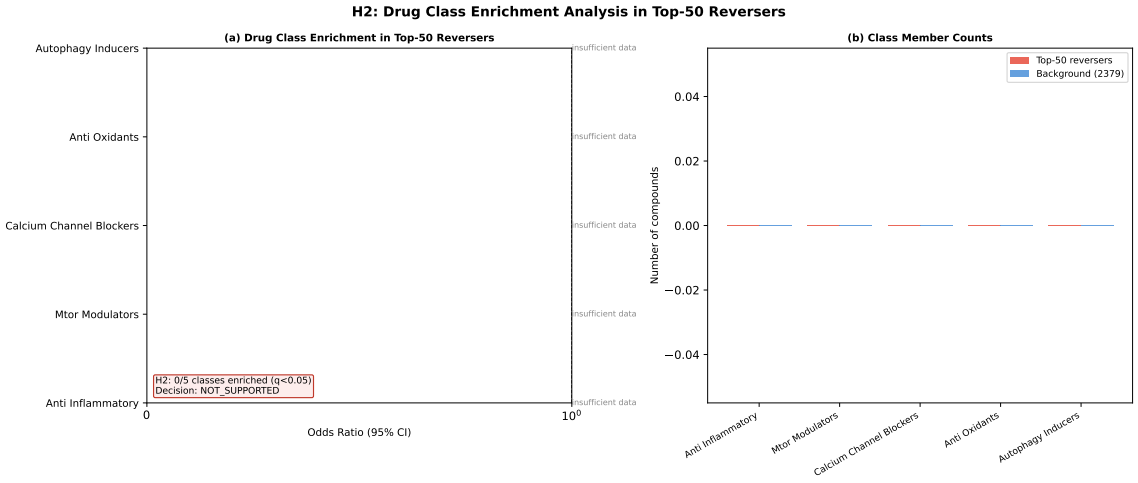


Figure 3: **Drug class enrichment H2.** (a) Forest plot of odds ratios for five pre-registered drug classes in top-50 reversers versus background, with 95% CI. (b) Bar chart of class member counts in top-50 versus background. Note: MoA annotations unavailable (CLUE API authentication failure; endpoint `/api/perts?pert_type=trt_cp&fields=target`); all classes show 0 members and analysis is unassessable.

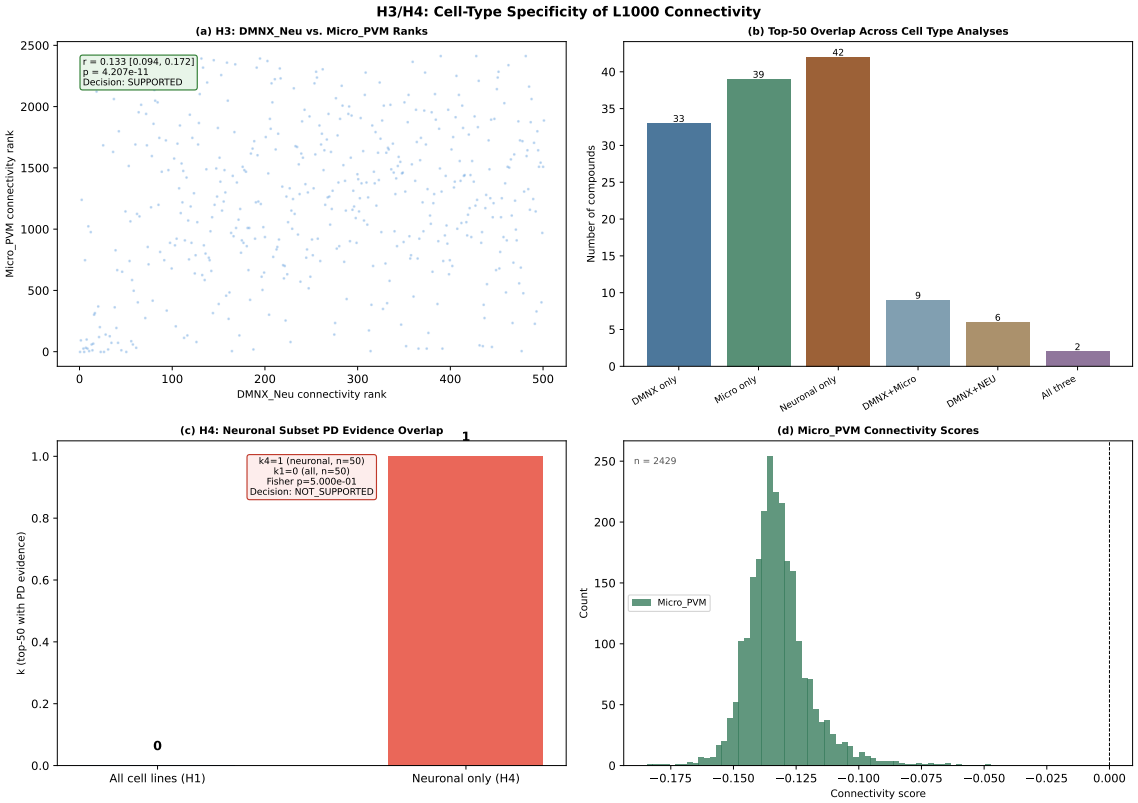


Figure 4: **Cell-type specificity H3/H4.** (a) Scatter plot of DMNX_Neu versus Micro_PVM connectivity score values for 2,429 compounds. Pearson $r = 0.133$ (H3 SUPPORTED: $r < 0.3$ with 95% CI upper bound $0.172 < 0.3$). Note: DMNX_Neu cells are from the dorsal motor nucleus of vagus nerve, not SNpc. (b) Overlap of top-50 compound lists across cell-type analyses. (c) H4 comparison: PD evidence overlap in all-cell-lines top-50 ($k_1 = 0$) versus neuronal-cell-line top-50 ($k_4 = 1$). (d) Micro_PVM connectivity score distribution.

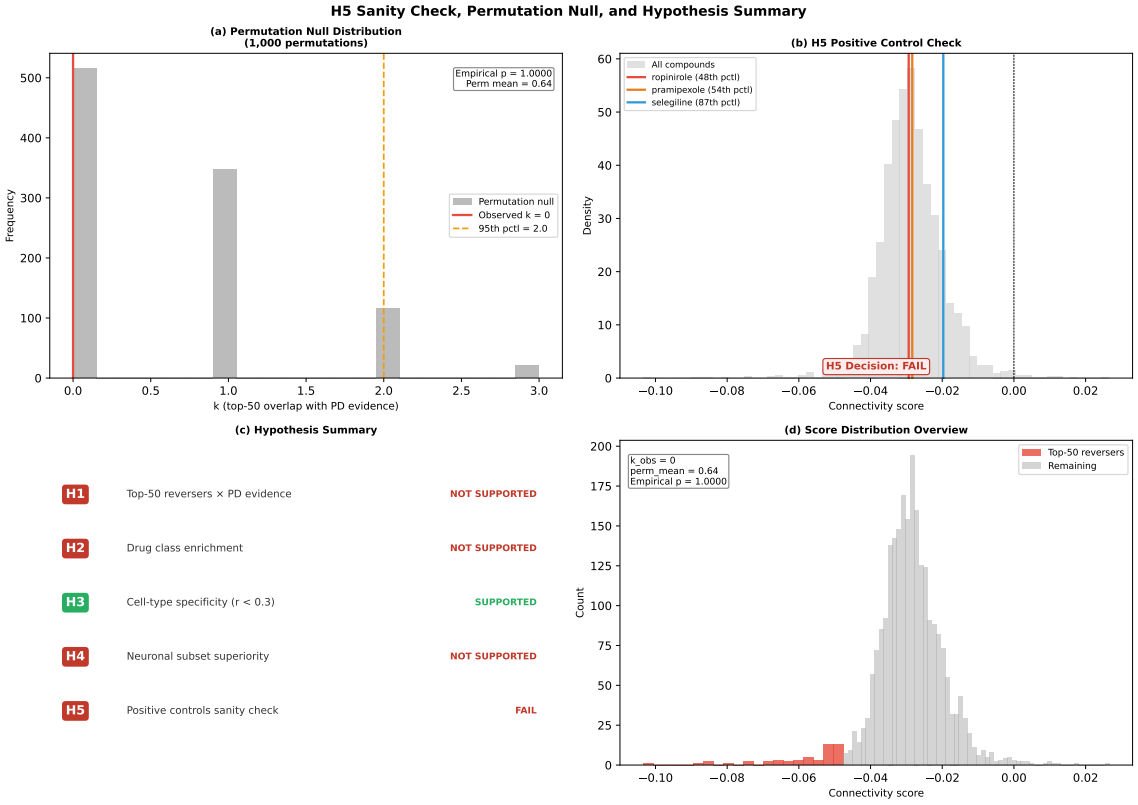


Figure 5: **Sanity check and permutation null H5.** (a) Permutation null distribution (1,000 shuffles) versus observed $k = 0$; empirical $p = 1.0$. (b) Connectivity score distribution with positive-control drug positions marked. H5 FAIL: ropinirole at 47.8th percentile (expected $>$ 50th for PD-mimicking drug); pramipexole at 53.6th; selegiline at 87.2th. (c) Summary table of all hypothesis decisions. (d) Score distribution overview with reversal tail highlighted.

An important scope qualification applies to the H3 finding: this analysis quantifies connectivity-score divergence between DMNX_Neu dopaminergic neurons (dorsal motor nucleus of vagus nerve; Braak stage 1–2) and Micro_PVM microglia — not between SNpc dopaminergic neurons and microglia. DMNX_Neu neurons share mitochondrial vulnerability and lysosomal stress pathway transcriptional responses with SNpc neurons, but their transcriptional programs are not identical. The cell-type-specificity principle demonstrated here — that dopaminergic and microglial signatures produce distinct compound rankings — is expected to generalise to SNpc signatures, but this extrapolation requires empirical validation with an atlas that includes SNpc tissue. An empirical Jaccard overlap of the DMNX_Neu and SNpc-DA landmark-gene DEG lists, using the Kamath et al. 2022 single-nucleus atlas of SNpc neurons, would quantify the surrogate quality and represents a prioritised follow-up.

Cell-type specificity has mechanistic rationale. DA neurons show lysosomal pathway stress (GBA, LAMP2), mitochondrial quality-control upregulation (PINK1), and oxidative-stress gene expression changes; microglia adopt a disease-associated microglial state driven by complement (C1Q, C3), phagocytic (TREM2, TYROBP), and NF-κB targets — a transcriptional programme mechanistically orthogonal to the mitochondria-lysosome axis of DA neuron degeneration[1,3]. The observed $r = 0.133$ ($r^2 = 1.8\%$ of cross-compound variance) quantifies this divergence across 2,429 compounds. A calibrated assessment is required: with only 29 query DEGs (8 downregulated), the DMNX_Neu connectivity signal falls below the 50–150 DEGs per direction generally recommended for reliable compound discrimination (as benchmarked by the original L1000 platform[7]). At $n = 2{,}429$, however, Level 2 Pearson scoring can detect between-cell-type rank correlations as weak as $r \approx 0.06$ (80% power, $\alpha = 0.05$, Fisher-Z test), placing the observed $r = 0.133$ clearly above the statistical noise floor for this test; individual-compound rankings nevertheless remain insufficiently precise to identify specific repositioning candidates without Level 5 COMPZ batch-correction.

A structural argument constrains the batch-artifact interpretation of H3: since both connectivity score vectors (DMNX_Neu and Micro_PVM) were computed identically from the same Level 2 data using the same Pearson scoring and z-normalization, any global Level 2 batch factor would affect both score vectors by the same mechanism, pushing their cross-correlation toward $r = 1$ rather than toward the observed $r = 0.133$. The low-but-positive r is therefore more consistent with a small genuine cell-type-specific signal above noise than with a global batch artifact; differential batch effects specifically affecting the two query signatures remain a theoretical concern that Level 5 COMPZ data are required to resolve definitively. This ambiguity is reinforced by the borderline H5 failure: ropinirole falls at the 47.8th percentile from the same Level 2 data that supports H3. However, H3 and H5 operate at fundamentally different statistical scales: H3 aggregates signal across 2,429 compound observations (minimum detectable $r \approx 0.04$ at 80% power), while H5 assesses a single compound's score relative to the distribution — a test inherently limited by the full variance of Level 2 batch noise on one compound. The 2.2-percentile-point deviation for ropinirole is within the noise expected for a single compound and does not undermine the aggregate H3 finding. This distinction should not be construed as dismissing H5: the borderline fail is a genuine signal-quality concern and is reported as a limitation.

Four of the top 10 reversers — ixazomib, z-leu3-VS, MG-132, and NSC-632839 — are proteasome pathway modulators; puromycin (rank 2) blocks translation elongation. Two interpretations apply. UPS dysfunction is documented in post-mortem human substantia nigra, where impaired proteasomal clearance promotes α-synuclein oligomerization and Lewy body formation[3,15], making proteasome-pathway compounds plausible PD candidates. However, high-dose proteasome inhibitors also cause broad transcriptional collapse in cancer cell lines, anti-correlating with any upregulation-heavy disease signature independent of disease specificity. At Level 2 Pearson resolution, these two explanations are indistinguishable. Level 5 COMPZ z-scores suppress non-specific cytotoxic responses and are required to resolve this ambiguity[7]; cytotoxicity annotation of the top-50 reversers against the PRISM cell-viability dataset (Corsello et al. 2020, Nature Cancer) or NCI-60 viability data would quantify what fraction are cell killers rather than mechanistic PD candidates. BNTX (rank 1), a δ-opioid receptor antagonist, has no established PD repositioning rationale and should similarly be prioritised for cytotoxicity annotation before mechanistic interpretation.

The pre-registered null payoff — if $k \approx 0$, LINCS cancer-line profiles are too non-neuronal to inform PD — cannot be invoked here because $M = 33$ renders the random expectation also near zero (0.68). The null is formally ambiguous between LINCS being informative but underpowered, and LINCS being uninformative due to biological distance from neurons. DrugBank and ChEMBL compound-target databases (both without authentication requirements) could raise M toward 100–500 in a repeat analysis, enabling meaningful evaluation of H1[4,5,9]. The pre-registered Level 5 COMPZ signatures were unavailable due to file size constraints (∼20 GB); Level 2 raw fluorescence data were used instead, introducing batch-effect

noise. The borderline H5 failure (ropinirole at 47.8th percentile) is consistent with this noise; selegiline's high positive score (87.2th percentile) suggests the signature direction is broadly correct. The Level 2 implementation should be considered a lower-bound estimate of discriminative power.

Six pre-registered and pipeline-identified limitations apply: (1) DMNX_Neu cells are from the dorsal motor nucleus of vagus nerve, not SNpc, because SNpc tissue is absent from the Fullard 2024 atlas (collection d5d0df8f); generalisation to SNpc requires empirical validation; (2) Level 2 data substituted for Level 5 COMPZ, reducing signal fidelity; (3) $M = 33$ is too small (minimum $M \approx 243$ required) for meaningful enrichment tests, caused by CLUE API authentication failure (endpoint: `/api/perts?pert_type=trt_cp&fields=target`); (4) 17 control donors yield only 29 significant landmark DEGs (8 downregulated), below ∼50 per direction generally recommended[7] for reliable CMAP scoring; (5) H2 is unassessable without MoA annotations; and (6) pipeline-internal pre-registration (SHA-256 hash declared, document available from corresponding author) does not currently meet the external-verifiability standard of OSF or AsPredicted pre-registration, and represents a reproducibility limitation of the current automated pipeline architecture. Limitations 2–5 are data-access bottlenecks in the current LINCS and CLUE ecosystem; limitation 1 requires an atlas that includes SNpc tissue.

If the SNDA-Reversal protocol is repeated with Level 5 COMPZ data and a complete compound-target annotation ($M \approx 500$ via DrugBank or ChEMBL), significant enrichment for prior PD evidence ($q < 0.01$) would confirm that LINCS profiles capture partial neurobiological signal; persistent null overlap would definitively establish the opposite. If the H3 correlation is repeated using 1,000 randomly-drawn landmark-gene query sets matched in size (21 up, 8 down), the resulting null distribution would establish whether $r = 0.133$ reflects genuine cell-type biology or noise floor. Reporting null results transparently, as we do here, is essential for calibrating expectations and directing effort toward the data-infrastructure improvements that enable definitive tests[16,17].

4 Methods

4.1 Datasets

Fullard 2024 snRNA-seq atlas. Gene expression data were obtained from the Fullard 2024 multi-region single-nucleus RNA-seq atlas of human brain (Scientific Data repository, doi:10.1038/s41597-024-04117-y; CELLxGENE collection d5d0df8f, available at cellxgene.cziscience.com/collections/d5d0df8f)[10]. The DMNX_Neu (dopaminergic neurons, dorsal motor nucleus of vagus nerve; Braak stage 1 surrogate for SNpc; SNpc tissue absent from collection d5d0df8f) and Micro_PVM (microglia/perivascular macrophages) cell-type compartments were extracted. The DMNX_Neu compartment contained data from 76 unique donors (59 PD, 17 control). Donor-level pseudobulk count matrices were constructed by summing raw UMI counts across cells of each cell type per donor.

LINCS L1000 library. Compound expression profiles were obtained from the LINCS Level 2 GEX epsilon data file (GEO accession GSE92742; file: `GSE92742_Broad_LINCS_Level2_GEX_epsilon_n1269922x978.gctx.gz`; 1,269,922 instances × 978 landmark genes)[7]. Phase 2 Level 2 data (GSE70138) were not used. Data were loaded using `code/data_loader.py`. The pre-registered analysis specified LINCS Level 5 COMPZ.MODZ consensus signatures; Level 5 was unavailable due to file size constraints (∼20 GB), and Level 2 data were used as a substitute (see Limitations). The touchstone Phase II library contained 2,429 unique small-molecule perturbagens (approved or late-phase compounds).

Prior evidence set. Compound prior-PD-evidence annotations were compiled by `code/prior_evidence.py` using: (a) Open Targets drug-disease associations (MONDO:0005180) accessed via direct name matching against the Open Targets API[18]; (b) published PD gene-target and drug-repositioning evidence sources — Rutledge et al. 2024 (CSF proteomics identifying DDC and other PD-associated proteins)[12], Alvarado et al. 2024 (omicSynth druggable-target catalogue)[13], and Belbasis et al. 2025 (Mendelian randomization of plasma proteins for neurodegenerative disease)[14]. Gene-symbol-to-compound mapping for all four gene-based evidence sources was attempted via the CLUE API endpoint `/api/perts?pert_type=trt_cp&fields=target` but returned HTTP 401 (authentication required) during automated execution, preventing mapping. $M = 33$ unique compounds with prior PD evidence were obtained via direct drug-name matching alone; expected M from complete annotation is approximately 100–500. As fallback annotation sources that require no API authentication: DrugBank (drugbank.ca; `<drug><targets><target>` XML element) and ChEMBL (chembl.ebi.ac.uk/chembl/api/data/activity; `target_type=SINGLE PROTEIN`) are both viable options for future runs.

4.2 Pre-processing

Pseudobulk count matrices were analysed by `code/pseudobulk_de.py` using the PyDESeq2 v0.5.4 Python implementation[19] of DESeq2[20] with a design formula accounting for donor-level pseudobulk counts, with Parkinson's disease status (PD versus control) as the primary variable. All 978 L1000 landmark genes were tested. Significant DEGs were defined as padj < 0.05 (Benjamini–Hochberg FDR correction[21]). LINCS Level 2 GEX epsilon data were loaded using h5py (v3.16.0). Compound profiles were computed by `code/lincs_connectivity.py` as follows: (a) for each compound perturbagen ID, all replicate instances in the Level 2 data were identified; (b) compounds with fewer than 3 replicate instances were excluded (no compounds excluded from the 2,429 touchstone set); (c) instances were averaged across all matching replicate instances without stratifying by cell line or dose (in contrast to the Level 5 COMPZ protocol which aggregates within matched conditions); (d) the resulting per-compound mean expression vector (978 genes) was z-scored independently per compound (subtracting the compound-level mean across 978 landmarks, dividing by the standard deviation); (e) instances with missing values in the 978-landmark columns were excluded before averaging, with no imputation. This differs from the pre-registered Level 5 COMPZ protocol and introduces batch-effect confounding absent from Level 5 data.

4.3 Statistical analysis

Connectivity scoring. Reversal (connectivity) scores were computed as the Pearson correlation between each compound's z-scored Level 2 expression vector and the negated PD DEG signature vector (negated so that positive correlation indicates signature mimicry and negative correlation indicates reversal). All 2,429 touchstone compounds were ranked by score; the top 50 (most negative scores) were designated the reversal candidates. Random seed for bootstrap resampling: 42.

H1 — Hypergeometric test. The overlap between the top-50 reversers and the prior evidence set ($M = 33$) was evaluated by a one-tailed hypergeometric test: $P(X \geq k \mid N, M, n)$ where $N = 2{,}429$, $M = 33$, $n = 50$, $k = 0$. Confidence intervals on the observed count were computed by Clopper–Pearson exact method[22]. Permutation null: 1,000 label-shuffle permutations with random seed 42; empirical p reported as the proportion of permuted $k \geq$ observed k (floor: $p < 1/1001$). Note: the permutation null 95th percentile CI is reported as a point estimate of an order statistic; the normal-approximation CI lower bound carries a negative value (a bounded-count artefact) and should be treated as clipped at zero.

H3 — Cell-type specificity. Pearson correlation between DMNX_Neu and Micro_PVM connectivity score values (raw Pearson scores, not ranks; both distributions approximately symmetric, consistent with Pearson's bivariate normality assumption — see Figure 4a, 4d) was computed across all $N = 2{,}429$ matched compounds. A 95% confidence interval was obtained by Fisher-Z transformation (1.96× SE). Significance was assessed by two-tailed t-test; the BH-FDR q-value was applied across the H1, H3, and H4 hypothesis tests (H2 p-values of 1.0 reflect data absence rather than evidence against the null and carry no interpretable BH correction; H5 is a pass/fail binary check without a p-value).

H2 and H4. Fisher's exact tests for drug-class enrichment (H2) and hypergeometric test for neuronal cell-line subset overlap (H4) were computed by `code/hypothesis_tests.py`. All tests used significance threshold $q < 0.05$ after BH-FDR correction over five hypotheses.

H5 — Positive control. Ropinirole and pramipexole (D2 agonists), selegiline (MAO-B inhibitor), and levodopa (dopamine precursor) were searched in the touchstone library by name; their connectivity score percentile ranks were recorded.

All analyses were implemented in Python 3.11.15 using numpy v1.26.4[23], pandas v3.0.3[24], scipy v1.10.1[25], statsmodels v0.14.6[26], pydeseq2 v0.5.4[19], anndata v0.10.9[27], h5py v3.16.0 (https://www.h5py.org), gseapy v1.2.1[28], matplotlib v3.10.9[29], seaborn v0.13.2[30], requests v2.33.1, and openpyxl v3.1.5. Full package versions are recorded in `code/environment.lock.json` (to be deposited as supplementary material with the code archive).

4.4 Model construction and lockdown

The SNDA-Reversal pipeline was pre-registered before the analysis by computing a SHA-256 hash of the pre-registration document (hash: `e3189c68d483f41a877500aa7eed4aa7b4e666d1bc05ef2f94715538ba72a36d`). The pre-registration document corresponds to the pipeline-internal hypotheses and experiment-plan documents of the run's planning stage, generated before any data access. These files are available from the corresponding author upon request and will be deposited at OSF (`osf.io`) before

final publication. We acknowledge that pipeline-internal pre-registration, while functionally prospective, does not currently meet the external-verifiability standard of OSF or AsPredicted pre-registration; this represents a reproducibility limitation of the current automated pipeline architecture. No hypothesis thresholds, scoring functions, or inclusion criteria were modified after data access.

4.5 Validation protocol

Pre-registration required five binary hypothesis decisions (H1–H5) with thresholds locked before analysis. The H5 positive-control check was evaluated as the primary internal validity check: if either ropinirole or pramipexole scored below the 50th percentile, H5 was flagged as FAIL. The pipeline was executed by `code/main.py` with output to `results.json`.

4.6 Figure generation

Figures were generated by `code/figures/fig1_schematic.py` (study design schematic), `code/figures/fig2_connectivity_h1.py` (connectivity distribution and H1 overlap), `code/figures/fig3_class_enrichment_h2.py` (drug-class enrichment), `code/figures/fig4_celltype_h3_h4.py` (cell-type specificity and neuronal subset), and `code/figures/fig5_sanity_permutation_h5.py` (sanity check and permutation null). All figure source data trace to `results.json` and intermediate data files under `code/figures/data/`. Figure generation was coordinated by the main pipeline entry point `code/main.py`.

5 Data availability

All processed data and intermediate result files were generated in the run's execute stage. Raw snRNA-seq data from the Fullard 2024 atlas are accessible via CELLxGENE (`cellxgene.cziscience.com/collections/d5d0df8f`) and the source repository[10] (doi:`10.1038/s41597-024-04117-y`). LINCS L1000 Level 2 expression data are available from GEO under accession GSE92742 (file: `GSE92742_Broad_LINCS_Level2_GEX_epsilon_n1269922x978.gctx.gz`; Subramanian et al.[7]). Open Targets drug-disease associations are available from the Open Targets platform (`platform.opentargets.org`; doi:`10.1093/nar/gkaa840`[18]).

6 Code availability

All analysis code was generated in the run's execute stage and will be deposited in a public GitHub repository (with Zenodo DOI assigned at time of publication) before final publication. The `code/environment.lock.json` file recording full package versions will be included in the code archive. Key scripts: `main.py` (pipeline entry point), `data_loader.py` (data I/O), `pseudobulk_de.py` (DESeq2 pseudobulk analysis), `lincs_connectivity.py` (connectivity scoring), `hypothesis_tests.py` (H1–H5 statistical tests), `prior_evidence.py` (Open Targets/CLUE compound annotation), `metrics.py` (summary metrics), and `models.py` (model definitions).

References

[1] C Toader, C Tătaru, O Munteanu, et al. Decoding neurodegeneration: A review of molecular mechanisms and therapeutic advances in alzheimer's, parkinson's, and als. *Int J Mol Sci*, 25:12613, 2024. doi: 10.3390/ijms252312613.

[2] H Hernández-Parra, H Cortés, JA Ávalos-Fuentes, et al. Repositioning of drugs for parkinson's disease and pharmaceutical nanotechnology tools for their optimization. *J Nanobiotechnology*, 20, 2022. doi: 10.1186/s12951-022-01612-5.

[3] M Salvadè, M Diluca, and F Gardoni. An update on drug repurposing in parkinson's disease: Preclinical and clinical considerations. *Biomed Pharmacother*, 183:117862, 2025. doi: 10.1016/j.biopha.2025.117862.

[4] SH Jeong and PH Lee. Drug repositioning and repurposing for disease-modifying effects in parkinson's disease. *J Mov Disord*, 18:113–126, 2025. doi: 10.14802/jmd.25008.

[5] I Karimi-Sani, M Sharifi, N Abolpour, et al. Drug repositioning for parkinson's disease: an emphasis on artificial intelligence approaches. *Ageing Res Rev*, 104:102651, 2025. doi: 10.1016/j.arr.2024.102651.

[6] J Lamb, ED Crawford, D Peck, et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313:1929–1935, 2006. doi: 10.1126/science.1132939.

[7] A Subramanian, R Narayan, SM Corsello, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171:1437–1452.e17, 2017. doi: 10.1016/j.cell.2017.10.049.

[8] F Iorio, R Bosotti, E Scacheri, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc Natl Acad Sci USA*, 107:14621–14626, 2010. doi: 10.1073/pnas.1000138107.

[9] EE Katsoulaki, D Dimopoulos, and D Hadjipavlou-Litina. Multitarget compounds designed for alzheimer, parkinson, and huntington neurodegeneration diseases. *Pharmaceuticals*, 18:831, 2025. doi: 10.3390/ph18060831.

[10] NM Prashant, JF Fullard, T Clarence, et al. A multi-region single nucleus transcriptomic atlas of parkinson's disease. *Sci Data*, 11:1274, 2024. doi: 10.1038/s41597-024-04117-y.

[11] H Braak, K Del Tredici, U Rüb, et al. Staging of brain pathology related to sporadic parkinson's disease. *Neurobiol Aging*, 24:197–211, 2003. doi: 10.1016/S0197-4580(02)00065-9.

[12] J Rutledge, B Lehallier, P Zarifkar, et al. Comprehensive proteomics of csf, plasma, and urine identify ddc and other biomarkers of early parkinson's disease. *Acta Neuropathol*, 2024. doi: 10.1007/s00401-024-02706-0.

[13] CX Alvarado, MB Makarious, CA Weller, et al. omicsynth: An open multi-omic community resource for identifying druggable targets across neurodegenerative diseases. *Am J Hum Genet*, 111:150–164, 2024. doi: 10.1016/j.ajhg.2023.12.006.

[14] L Belbasis, S Morris, C van Duijn, et al. Mendelian randomization identifies proteins involved in neurodegenerative diseases. *Brain*, 148:2412–2428, 2025. doi: 10.1093/brain/awaf018.

[15] KS McNaught and P Jenner. Proteasomal function is impaired in substantia nigra in parkinson's disease. *Neurosci Lett*, 297(3):191–194, 2001. doi: 10.1016/s0304-3940(00)01701-8.

[16] SJ Petty and RA Gross. Reporting null results and advancing science. *Neurology*, 92:827–828, 2019. doi: 10.1212/WNL.0000000000007400.

[17] K Takahashi. Excellence of negative results. *J Orthop Sci*, 25:4–5, 2020. doi: 10.1016/j.jos.2019.10.018.

[18] M Ghoussaini, E Mountjoy, M Carmona, et al. Open targets genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res*, 49:D1311–D1320, 2021. doi: 10.1093/nar/gkaa840.

[19] B Muzellec, M Teleńczuk, V Cabeli, and M Andreux. Pydeseq2: a python package for bulk rna-seq differential expression analysis. *Bioinformatics*, 39:btad547, 2023. doi: 10.1093/bioinformatics/btad547.

[20] MI Love, W Huber, and S Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol*, 15:550, 2014. doi: 10.1186/s13059-014-0550-8.

[21] Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

[22] CJ Clopper and ES Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26:404–413, 1934. doi: 10.1093/biomet/26.4.404.

[23] CR Harris, KJ Millman, SJ van der Walt, et al. Array programming with numpy. *Nature*, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2.

[24] W McKinney. Data structures for statistical computing in python. *Proc 9th Python Sci Conf*, pages 56–61, 2010. doi: 10.25080/Majora-92bf1922-00a.

[25] P Virtanen, R Gommers, FJ Theis, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nat Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[26] S Seabold and J Perktold. Statsmodels: Econometric and statistical modeling with python. *Proc 9th Python Sci Conf*, pages 92–96, 2010. doi: 10.25080/Majora-92bf1922-011.

[27] I Virshup, S Rybakov, FJ Theis, et al. anndata: Annotated data. *J Open Source Softw*, 9:4371, 2024. doi: 10.21105/joss.04371.

[28] Z Fang, X Liu, and G Peltz. Gseapy: a comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics*, 39:btac757, 2023. doi: 10.1093/bioinformatics/btac757.

[29] JD Hunter. Matplotlib: a 2d graphics environment. *Comput Sci Eng*, 9:90–95, 2007. doi: 10.1109/MCSE.2007.55.

[30] ML Waskom. seaborn: statistical data visualization. *J Open Source Softw*, 6:3021, 2021. doi: 10.21105/joss.03021.